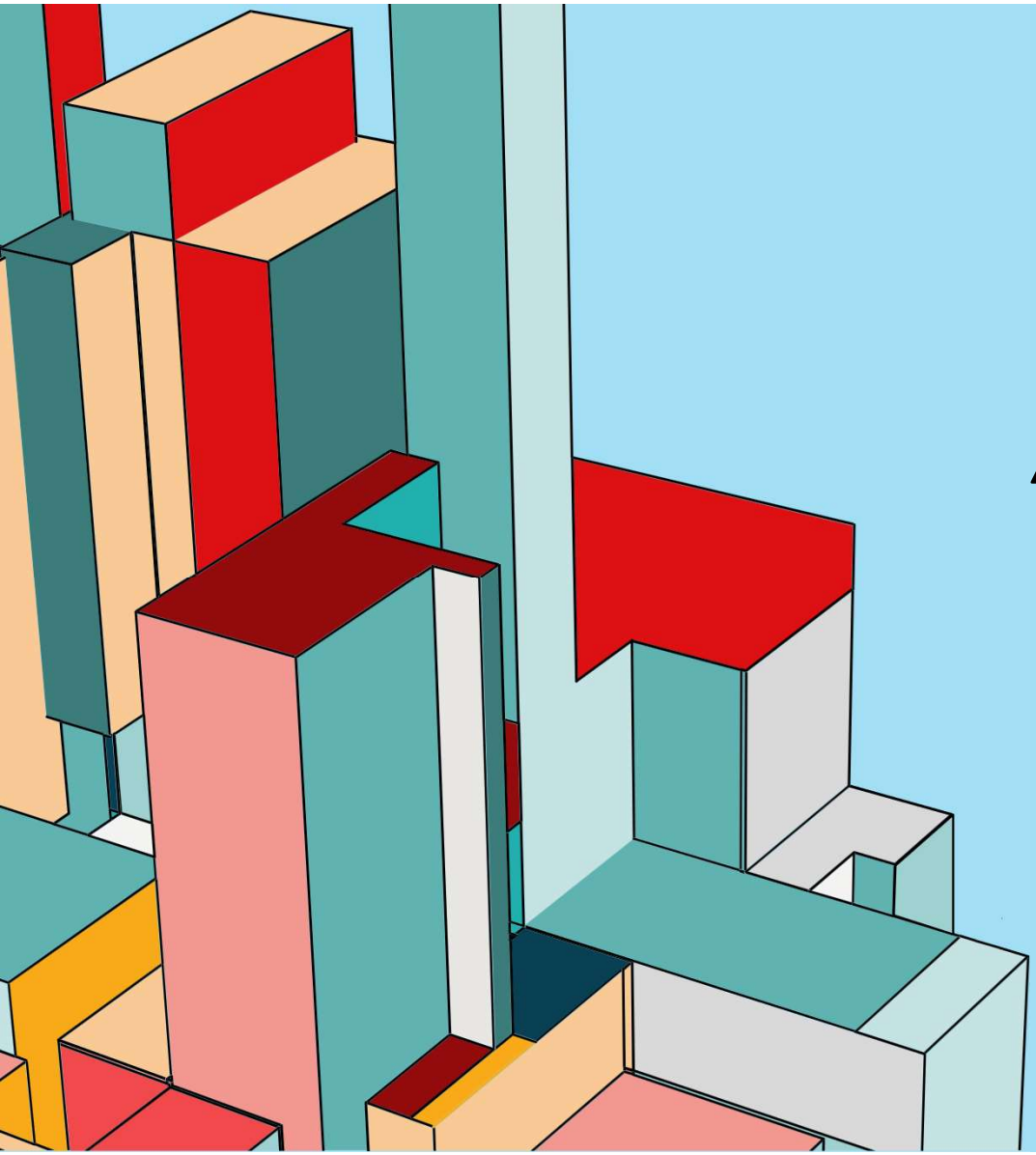


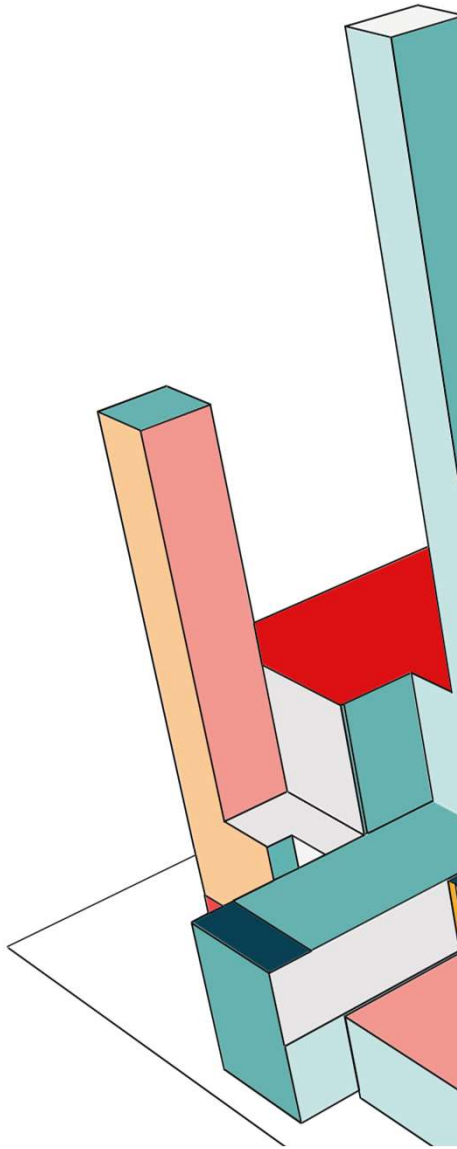
DAY 1:

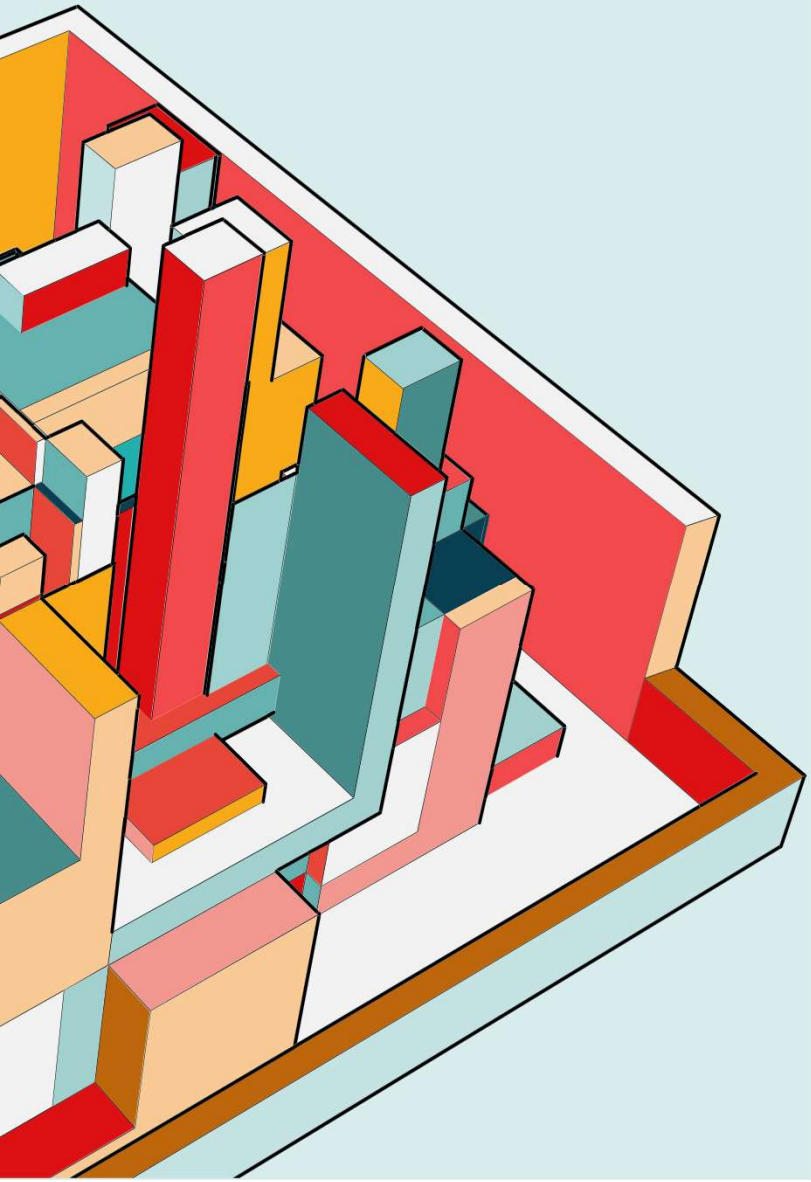
DATA ANALYTICS AND PREDICTION MODEL WITH PYTHON

AGENDA

Day 1 : Morning Time

- Introduction to data analytics, BI and Data Science
- Introduction to data analytics
- Transforming Data
- Data Analysis and Presentation





INTRODUCTION TO DATA ANALYTICS, BI AND DATA SCIENCE

DATA ANALYTICS, BUSINESS INTELLIGENCE (BI) AND DATA SCIENCE

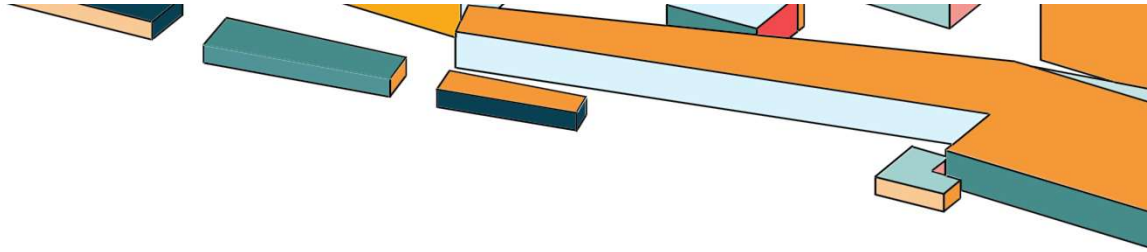
หัวข้อ	Data Analytics	Business Intelligence – BI	Data Science
ความหมาย	กระบวนการวิเคราะห์ข้อมูลทั้งในอดีตและปัจจุบัน เพื่อหาข้อสรุป แนวโน้ม และสาเหตุของเหตุการณ์ที่เกิดขึ้น ช่วยให้องค์กรเข้าใจ *"เกิดอะไรขึ้น"* และ *"ทำไมถึงเกิดขึ้น"*	ชุดของเครื่องมือและระบบที่รวบรวม จัดเก็บ และนำเสนอข้อมูลธุรกิจ เพื่อช่วยผู้บริหารและผู้ใช้งานในการติดตามผลการดำเนินงานและตัดสินใจได้รวดเร็ว	ศาสตร์ที่ผสมผสานระหว่างสถิติ คณิตศาสตร์ การเขียนโปรแกรม และความรู้ด้านธุรกิจ เพื่อนำข้อมูลมาใช้สร้างแบบจำลองเชิงพยากรณ์ (Predictive) และเชิงแนะนำ (Prescriptive)
ลักษณะเด่น	• มุ่งเน้นการใช้สถิติและการแสดงผลข้อมูล (Data Visualization) • ทำงานกับข้อมูลย้อนหลังและข้อมูลปัจจุบัน • เป็นพื้นฐานในการตัดสินใจเชิงธุรกิจ	• ใช้ Dashboard, KPI, และรายงานต่าง ๆ • เน้นการนำเสนอข้อมูลที่ใช้งานง่าย เข้าใจได้ทันที • ส่วนใหญ่เป็นการวิเคราะห์เชิงพรรณนา (Descriptive) และวิเคราะห์สาเหตุ (Diagnostic)	• ใช้ Machine Learning (ML) และ Artificial Intelligence (AI) • ตอบคำถามว่า *"จะเกิดอะไรขึ้น?"* และ *"ควรทำอย่างไร?"* • ต้องการทักษะการเขียนโปรแกรม (Python, R) และการจัดการ Big Data
ตัวอย่าง	• วิเคราะห์ยอดขายย้อนหลังเพื่อหาสาเหตุการตกต่ำ • วิเคราะห์พฤติกรรมการยกเลิกบริการของลูกค้า (Customer Churn)	• Dashboard แสดงยอดขายแบบ Real-time • รายงานสรุปผลกำไรขาดทุนรายเดือนของบริษัท	• แบบจำลองทำนายยอดขายล่วงหน้า • ระบบแนะนำสินค้า (Recommendation System) เช่น Amazon, Netflix • การตรวจจับธุรกรรมทุจริต (Fraud Detection)

DATA ANALYTICS, BUSINESS INTELLIGENCE (BI) AND DATA SCIENCE

ระดับ	จุดเน้น	คำถามหลัก	ประเภทการวิเคราะห์*	ผู้ใช้งานหลัก	เครื่องมือ
Data Analytics **ใช้ข้อมูลเพื่อหาสาเหตุและอธิบายอดีต**	วิเคราะห์ข้อมูลในอดีต/ปัจจุบัน มีผลลัพธ์ Insight ลึก, แบบจำลอง, การคาดการณ์	*เกิดอะไรขึ้น? ทำไม?*	Descriptive, Diagnostic,	Data Scientist, Data Analyst	Excel, SQL, Visualization
Business Intelligence (BI) **ใช้เครื่องมือสร้าง Dashboard/ รายงานเพื่อช่วยตัดสินใจ**	การติดตามผลธุรกิจแบบ Real-time มีผลลัพธ์ Dashboard, KPI, รายงาน	*ตอนนี้เกิดอะไรขึ้น?*	Descriptive, Diagnostic	ผู้บริหาร, Business Users	Power BI, Tableau, SAP Analytics Cloud
Data Science **ใช้ ML/AI เพื่อทำนายและเสนอแนวทางอนาคต**	การทำนายและการแนะนำแนวทาง มีผลลัพธ์เป็น การคาดการณ์, ระบบอัตโนมัติ, การเพิ่มประสิทธิภาพ, การสนับสนุนการตัดสินใจ	*จะเกิดอะไรขึ้น? ควรทำอะไร?*	Descriptive, Diagnostic, Predictive, Prescriptive	Data Scientist, Data Engineer, ML Engineer, Business Analyst	Python, R, ML/AI

ประเภทการวิเคราะห์

1. **Descriptive Analytics** – อธิบายว่า *"เกิดอะไรขึ้น?"*
2. **Diagnostic Analytics** – อธิบายว่า *"ทำไมถึงเกิดขึ้น?"*
3. **Predictive Analytics** – ทำนายว่า *"จะเกิดอะไรขึ้นต่อไป?*
4. **Prescriptive Analytics** – แนะนำว่า *"ควรทำอย่างไร?"*



USE CASES

หมวด	Data Analytics	Business Intelligence (BI)	Data Science
	มุ่งเน้นการทำความเข้าใจสิ่งที่เกิดขึ้นและสาเหตุ	**มุ่งเน้นการติดตามผลธุรกิจแบบ Real-time และช่วยตัดสินใจรวดเร็ว**	**มุ่งเน้นการพยากรณ์และการแนะนำแนวทาง (Predictive & Prescriptive)**
Finance	วิเคราะห์ต้นทุน, กำไร	Dashboard งบการเงิน	Fraud Detection, Credit Scoring
Marketing	วิเคราะห์ ROI ของแคมเปญ	Dashboard แสดงยอดขาย/การตลาด	Recommendation Engine, Customer Segmentation
HR	วิเคราะห์การลาออก	Dashboard แสดงจำนวนพนักงาน	Predictive Attrition Model
Supply Chain	วิเคราะห์สาเหตุ delay	Dashboard Stock/Delivery	Demand Forecasting, Route Optimization
Manufacturing	วิเคราะห์ defect rate	Dashboard สรุปผลผลิต	Predictive Maintenance

COMPARE BETWEEN DATA ANALYTICS &BI

1. Data Analytics Example

เน้น: ค้นหาสาเหตุ รูปแบบ แนวโน้ม และการพยากรณ์

สถานการณ์: บริษัทอีคอมเมิร์ซอยากรู้ว่ายอดขายไตรมาสที่แล้วทำไมถึงลดลง

ขั้นตอน Data Analytics

- เก็บข้อมูลดิบ: ทราฟฟิกเว็บไซต์, ข้อมูลลูกค้า, แคมเปญโฆษณา, รายการขาย
- ใช้เครื่องมือวิเคราะห์ (Python, R, SQL)
- ทำการพยากรณ์ (เช่น Regression) เพื่อดูว่า การใช้จ่ายด้านการตลาด, การตั้งราคา, หรือปัญหาเว็บ มีผลกับยอดขายอย่างไร
- แบ่งกลุ่มลูกค้า (Customer Segmentation) เช่น Cluster เพื่อหาลูกค้ากลุ่มสำคัญ

ผลลัพธ์

- “ลูกค้าวัย 25–34 ปี ที่มาจากโฆษณา Instagram มีอัตราการซื้อสูงกว่า 40%”
- “การที่เว็บไซต์ล่มทำให้ยอดขายลดลง 10%”

2. Business Intelligence (BI) Example

เน้น: การติดตาม รายงาน แดชบอร์ด เพื่อใช้ตัดสินใจ

สถานการณ์: บริษัทอีคอมเมิร์ซอยากให้ผู้จัดการสามารถติดตามยอดขายประจำวันได้

ขั้นตอน BI

- เชื่อม Data Warehouse กับเครื่องมือ BI (Power BI, Tableau, SAP Analytics Cloud)
- สร้าง Dashboard แสดง KPI: ยอดขายตามสินค้า, ยอดขายเทียบเป้า, ลูกค้า Top 10, รายได้ตามภูมิภาค, สต็อกสินค้า
- จัดทำรายงานอัตโนมัติส่งผู้จัดการทุกวันจันทร์

ผลลัพธ์

- Dashboard แสดงว่า: “ยอดขายสัปดาห์ที่แล้ว = 1.2M บาท (ต่ำกว่าเป้า 5%) สินค้าขายดีสุด: หูฟังไร้สาย”
- สามารถเจาะดูรายละเอียดยอดขายตามภูมิภาค สินค้า หรือช่องทางการตลาด
- แจ้งเตือน: “สินค้ารหัส X เหลือน้อยกว่ากำหนดในสต็อก”

แนวโน้มสำคัญใน DATA ANALYTICS

- ✓ **การวิเคราะห์แบบเรียลไทม์ (Real-time Analytics)**

  การใช้ข้อมูลและตอบสนองในทันที เช่น Snowpipe โหลดข้อมูลเข้า Snowflake Data Warehouse แบบอัตโนมัติและเกือบ Real-time โดยไม่ต้องรัน Batch Load แบบ Manual เมื่อมีไฟล์ใหม่ เช่น CSV, JSON

- ✓ **Data Mesh & Decentralization**

  การกระจายความรับผิดชอบด้านข้อมูลสู่แต่ละหน่วยงาน ช่วยให้การเข้าถึงและวิเคราะห์ข้อมูลเร็วขึ้นและครอบคลุมกว่าเดิม

  ตัวอย่างการใช้งานจริงก่อนใช้ Data Mesh (Centralized)ทีม Marketing ขอ Data จาก Data Lake ส่วนกลาง ทำให้ต้องรอ Data Engineer รวมข้อมูลจากหลายระบบทำให้ใช้เวลานานหลังใช้ Data Mesh (Decentralized)ทีม Marketing มี Data Owner + Pipeline ของตนเอง ทำให้ สร้าง Data Product "Campaign Performance" และทีม Finance สามารถ reuse Data Product นี้ได้โดยตรงผ่าน data catalog โดยไม่ต้องทำ integration เอง

- ✓ **NLP และ Natural Language Interfaces**

  เทคโนโลยีที่ทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์ เช่น Google Translate, CHATGBT, ระบบตรวจสอบไวยากรณ์

- ✓ **ความสดใหม่ & คุณภาพของข้อมูล (Veracity & Data Observability)**

  การตรวจสอบและควบคุมคุณภาพข้อมูลเพื่อให้มั่นใจว่าข้อมูลที่ใช้วิเคราะห์ถูกต้องและทันสมัย

- ✓ **GenAI & RAG (Retrieval-Augmented Generation)**

  สามารถ "สร้าง" ข้อมูลใหม่จากโมเดล Machine Learning ที่เรียนรู้ pattern ของข้อมูลเดิม เช่น ChatGPT ⟶ สร้างข้อความ, โค้ด, บทความ co-pilot สร้าง code

แนวโน้มสำคัญใน BUSINESS INTELLIGENCE (BI)

- ✓ **Augmented Analytics**

  การนำ AI และ Machine Learning มาอัตโนมัติในงานจัดเตรียมข้อมูล การวิเคราะห์ และการสร้างความเข้าใจ ได้แก่ การใช้ Natural Language Query (NLQ) เช่น SAP Analytics Cloud: ผู้ใช้พิมพ์ “ยอดขายปีนี้เทียบกับปีที่แล้ว จะได้กราฟออกมา หรือ Power BI Q&A: ผู้ใช้พิมพ์ “Top 5 customers by revenue” ระบบสร้างรายงานอัตโนมัติ

- ✓ **Semantic Layer & MCP (Metric Context Protocol)**

  คือ ชั้นกลาง (abstraction layer) ระหว่าง Data Source (Database, Data Lake, Warehouse) และ Business User/BI Tool ทำหน้าที่แปลง raw data ให้อยู่ในรูปแบบที่ “เข้าใจตรงกัน” และ “สื่อสารด้วยภาษาธุรกิจ” ช่วยให้ผู้ใช้ไม่ต้องเขียน SQL เอง และ MCP เป็น protocol ที่ทำให้ semantic layers จากหลายระบบ คุยกันได้

- ✓ **Self-Service BI & Democratization**

  ผู้ใช้งานธุรกิจสามารถสร้างรายงานได้เอง โดยไม่ต้องพึ่งทีม IT, เพิ่มทักษะและการเข้าถึงข้อมูลอย่างทั่วถึง เช่น การใช้ drag &drop /low code no code

- ✓ **Data Governance, Security & Quality**

  การรักษาความน่าเชื่อถือของข้อมูล การดูแลคุณภาพความสมบูรณ์ของข้อมูล และการจัดการความปลอดภัย เพื่อเสริมให้ข้อมูลจาก BI มีความน่าเชื่อถือ

- ✓ **BI-as-a-Service & Collaborative BI**

  บริษัทขนาดกลางที่ใช้ BI-as-a-Service (Cloud) + Embedded BI โดย BI ไม่ได้แยกเป็น platform เดียว แต่ฝังอยู่ใน ระบบงานหลัก (Operational Systems) เช่น ERP, CRM, HRM ⟶ ผู้ใช้เห็น dashboard/analytics อยู่ใน workflow เดิม ⟶ ไม่ต้องสลับหน้าจอ ช่วยให้ข้อมูลเป็นส่วนหนึ่งของการตัดสินใจประจำวัน ไม่ใช่แค่รายงานที่ดูย้อนหลัง

แนวโน้มสำคัญใน DATA SCIENCE

✓ **จัดการกับ "Dark Data"**

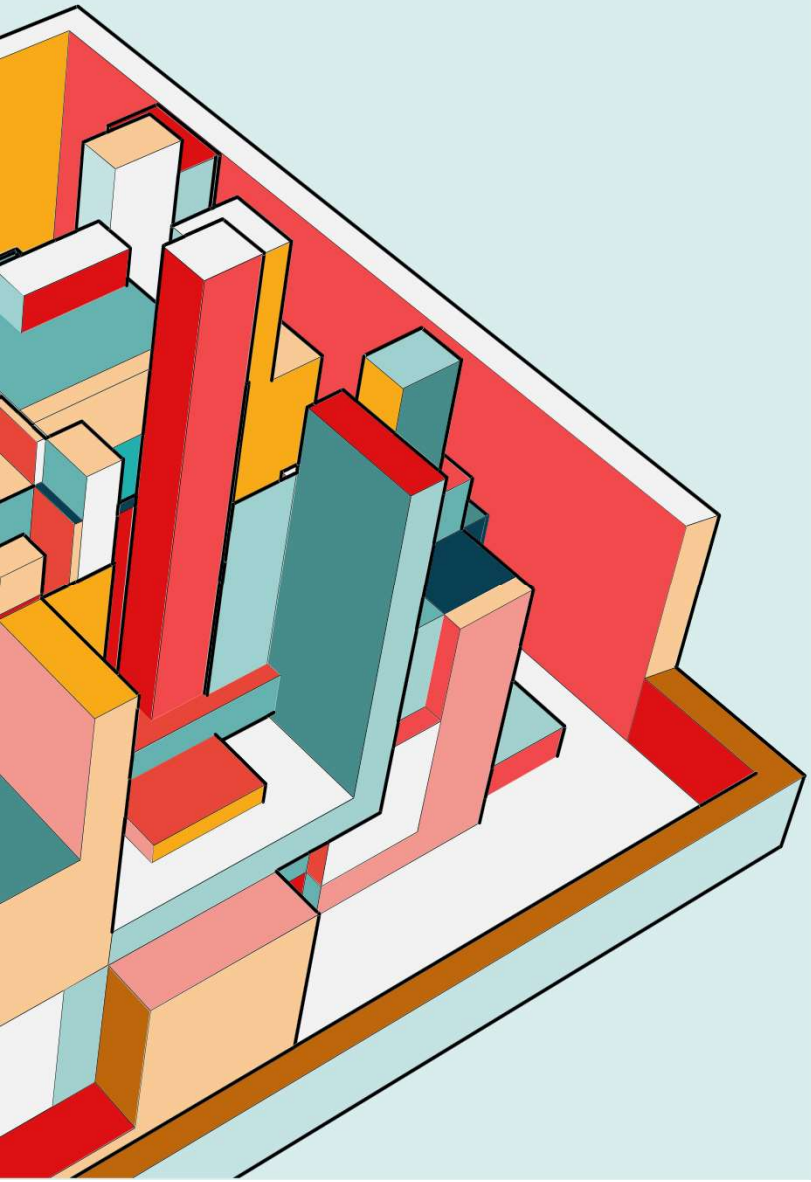
Dark Data = "ข้อมูลที่เก็บไว้แต่ไม่ได้ใช้" เช่น อีเมล เอกสาร ⟶ ถ้าไม่จัดการจะกลายเป็น ต้นทุน + ความเสี่ยง
การจัดการ = ค้นหา ⟶ ประเมิน ⟶ จัดการ (เก็บ/ลบ/ย้าย) ⟶ สร้าง governance ⟶ หาประโยชน์ถ้าเป็นไปได้
สร้างมูลค่าด้วย AI, Knowledge Graph และระบบอัจฉริยะ

✓ **ลงทุนในโครงสร้างข้อมูลที่เชื่อถือได้ก่อนใช้ AI**

การพัฒนาคุณภาพข้อมูลและระบบที่มั่นคงเป็นพื้นฐานก่อนเริ่มใช้ AI โดยเฉพาะองค์กรใหญ่ที่เก่งในด้าน Governance ได้ผลดีกว่ามาก

✓ **GenAI จากแค่ช่วยเรื่องความเร็ว ⟶ เข้าถึงการตัดสินใจธุรกิจ**

AI ไม่ใช่แค่ช่วยทำงานเร็ว แต่เริ่มถูกใช้วิเคราะห์ตลาด, ทำ M&A insights และช่วยควบคุมความเสี่ยงอย่างมีประสิทธิภาพ



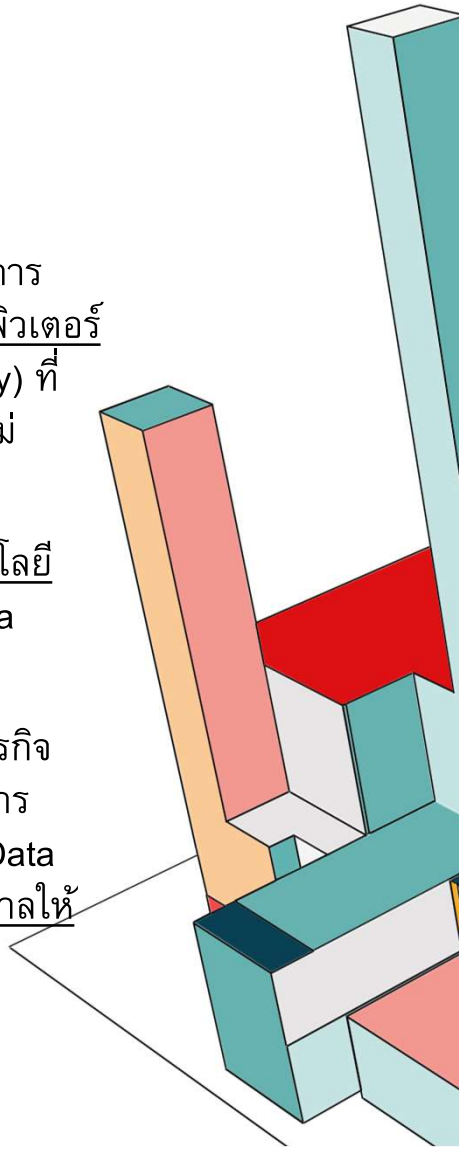
INTRODUCTION TO DATA ANALYTICS

BACKGROUND

ข้อมูล (Data) มีการเพิ่มพูนอย่างก้าวกระโดดในอัตราทวีคูณ (Exponential Growth) ปรากฏการณ์การแพร่กระจายของข้อมูล (Data Proliferation) นี้เกิดจากการพัฒนาของพลังการประมวลผลของคอมพิวเตอร์ (Computer Processing Power) ที่เพิ่มสูงขึ้น ความสามารถในการจัดเก็บข้อมูล (Storage Capacity) ที่ขยายตัวอย่างต่อเนื่อง ตลอดจนแบนด์วิดท์ (Bandwidth) ที่มีศักยภาพมากขึ้น ซึ่งจนถึงปัจจุบันยังไม่ปรากฏสัญญาณของการชะลอตัวลงแต่อย่างใด

ภายใต้บริบทของโลกอินเทอร์เน็ตในทุกสรรพสิ่ง (Internet of Things: IoT) เมื่อผนวกเข้ากับเทคโนโลยีแบนด์วิดท์ยุคที่ห้า (5G Bandwidth) ปริมาณข้อมูลที่ถูกสร้างขึ้นและส่งต่อมาจากแหล่งกำเนิด (Data Sources) ที่หลากหลายจะยิ่งทวีจำนวนมากขึ้นอย่างมีนัยสำคัญ

คำถามเชิงวิชาการที่สำคัญจึงเกิดขึ้นว่า ข้อมูลจำนวนมหาศาลเหล่านี้มีคุณค่าในเชิงใด และองค์กรธุรกิจสามารถใช้ข้อมูลดังกล่าวเพื่อสร้างองค์ความรู้เชิงลึก (Insights) และเสริมสร้างความได้เปรียบทางการแข่งขัน (Competitive Advantage) ได้หรือไม่ บทนี้จึงมุ่งศึกษาแนวคิดว่าด้วยการวิเคราะห์ข้อมูล (Data Analytics) และเครื่องมือที่เกี่ยวข้อง (Toolsets) ซึ่งมีความจำเป็นต่อการเปลี่ยนข้อมูลปริมาณมหาศาลให้กลายเป็นสารสนเทศ (Information) ที่มีคุณค่าและสามารถนำไปใช้ประโยชน์ได้จริง



BACKGROUND

การใช้ข้อมูล (Data) อย่างเหมาะสมได้กลายเป็นประเด็นที่มีความสำคัญอย่างยิ่งสำหรับวิชาชีพบัญชี นักบัญชีในทุกสาขาการปฏิบัติงานต่างก็นำข้อมูลมาใช้ในรูปแบบที่น่าสนใจและหลากหลาย ตัวอย่างเช่น

สายงาน (Practice Area)	การใช้ข้อมูล (Data Usage)	ประโยชน์ที่ได้รับ (Benefits)
ผู้สอบบัญชี (Auditors – Internal & External)	- ใช้การวิเคราะห์ข้อมูล (Data Analytics) และระบบอัตโนมัติ (Automation) เพื่อตรวจสอบธุรกรรมทั้งประชากร (Full Population) แทนการสุ่มตัวอย่าง (Sample) - จัดหาหลักฐานที่ชัดเจนมากขึ้นในการพิสูจน์การปฏิบัติตามมาตรการทางการบัญชี (Accounting Rules)	- เพิ่มคุณภาพและความน่าเชื่อถือของการตรวจสอบ - ลดความเสี่ยงจากการพึ่งพาตัวอย่างขนาดเล็ก - สนับสนุนการปฏิบัติตามกฎระเบียบอย่างมีประสิทธิภาพ
นักบัญชีฝ่ายองค์กร (Corporate Accountants)	- ใช้ข้อมูลเพื่อคำนวณต้นทุนสินค้าและบริการ (Costing Products & Services) ได้อย่างแม่นยำ - ประเมินความเสี่ยง (Risk Assessment) ได้ดีขึ้น - ระบุโอกาสในการรักษาและเพิ่มมูลค่า (Preserve & Enhance Value)	- การตัดสินใจด้านต้นทุนและการกำหนดราคาแม่นยำขึ้น - ลดความเสี่ยงทางการเงิน - เพิ่มศักยภาพในการสร้างและรักษามูลค่าองค์กร
ผู้เชี่ยวชาญด้านภาษี (Tax Professionals)	- ใช้ Data Analytics เพื่อประมาณการผลกระทบทางภาษี (Tax Consequences) แบบเรียลไทม์ (Real-time Estimates) - ตอบสนองต่อการตรวจสอบจากหน่วยงานกำกับดูแล (Regulators) ได้มีประสิทธิภาพมากขึ้น	- เพิ่มความสามารถในการวางแผนภาษีเชิงกลยุทธ์ - ลดความเสี่ยงจากข้อพิพาททางภาษี - มีบทบาทเชิงรุกต่อการตัดสินใจทางธุรกิจระดับสูง
ที่ปรึกษาการลงทุน (Investment Advisors)	- ใช้ข้อมูลเพื่อระบุโอกาสการลงทุนที่ให้ผลตอบแทนที่ดีกว่า (Favorable Investment Opportunities) และนำมาแนะนำแก่ลูกค้า	- เพิ่มความแม่นยำในการคัดเลือกการลงทุน - สร้างความเชื่อมั่นกับลูกค้า - เพิ่มความสามารถในการแข่งขันในตลาดการเงิน

BACKGROUND

จากตัวอย่างที่กล่าวมา แสดงให้เห็นว่าผู้ใช้งานจำเป็นต้องมีความเข้าใจเกี่ยวกับข้อมูล (Data) และการเปลี่ยนแปลงที่เกิดขึ้นกับโลกธุรกิจ เพื่อทำความเข้าใจขอบเขตของการปฏิวัติข้อมูล (Data Revolution) สิ่งสำคัญคือการพิจารณาแนวคิด “4 V’s of Big Data” ได้แก่ Volume, Velocity, Variety และ Veracity

4 V’s ของ Big Data

- **Volume (ปริมาณ):** ข้อมูลจำนวนมหาศาลที่องค์กรเก็บและสร้างขึ้น
- **Velocity (ความเร็ว):** ความเร็วในการสร้างและส่งต่อข้อมูล
- **Variety (ความหลากหลาย):** รูปแบบข้อมูลที่แตกต่างกัน (ตัวเลข, ข้อความ, ภาพ ฯลฯ)
- **Veracity (ความน่าเชื่อถือ):** คุณภาพและความถูกต้องของข้อมูล

ประโยชน์ของ Big Data ต่อสายการบิน

- ปรับปรุง **ความสัมพันธ์กับลูกค้า**
- วางแผน **ตารางซ่อมบำรุง**
- กำหนด **เส้นทางการบินที่เหมาะสมกว่า**
- จัดตาราง **พนักงาน** ได้มีประสิทธิภาพ
- แก้ปัญหาธุรกิจอื่น ๆ

ตัวอย่างจากอุตสาหกรรมสายการบิน

- ความสำเร็จของสายการบินขึ้นอยู่กับการ **คาดการณ์เวลาเดินทางถึง/ออกเดินทางที่แม่นยำ**
- เดิมทีใช้การประเมินของนักบิน ⟶ ขาดความแม่นยำ (โดยเฉพาะช่วงลงจอดที่ต้องใช้สมาธิสูง)
- **PASSUR Aerospace**
  - ❖ ใช้ข้อมูลหลายแหล่ง: ตารางบิน, สภาพอากาศ, และเรดาร์แบบพาสซีฟใกล้สนามบิน
  - ❖ เรดาร์ส่งข้อมูลทุก 4.6 วินาที ⟶ ความเร็วสูง (Velocity)
  - ❖ เก็บข้อมูลย้อนหลังมหาศาล ⟶ ปริมาณมาก (Volume)
  - ❖ ผลลัพธ์: ลดช่องว่างระหว่างเวลา “จริง” และเวลา “ที่คาดการณ์” เกือบหมด (Veracity)

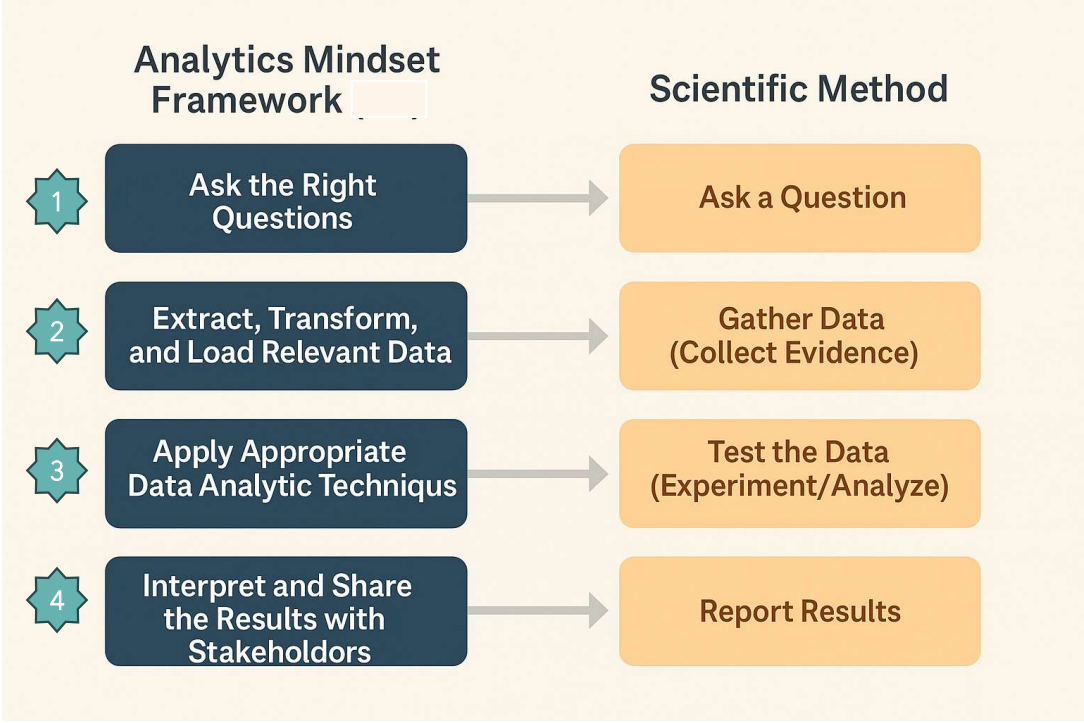
ข้อคิดสำคัญ

- Big Data อย่างเดียวไม่พอ ⟶ ต้องมี **การวิเคราะห์ (Analytics)**
- ความสำเร็จขึ้นอยู่กับ **Mindset ที่มองข้อมูลเชิงองค์รวม** ไม่ใช่แค่ใช้เครื่องมือ

ANALYTIC MINDSET

กรอบความคิด (Mindset) หมายถึง ทัศนคติทางจิตใจ วิธีคิด หรือกรอบความคิดที่บุคคลใช้ในการรับรู้และตัดสินใจ กรอบความคิดเป็นการรวบรวมของความเชื่อและความคิดที่ทรงพลัง ซึ่งมีอิทธิพลต่อการคิด ความรู้สึก และการกระทำของบุคคล

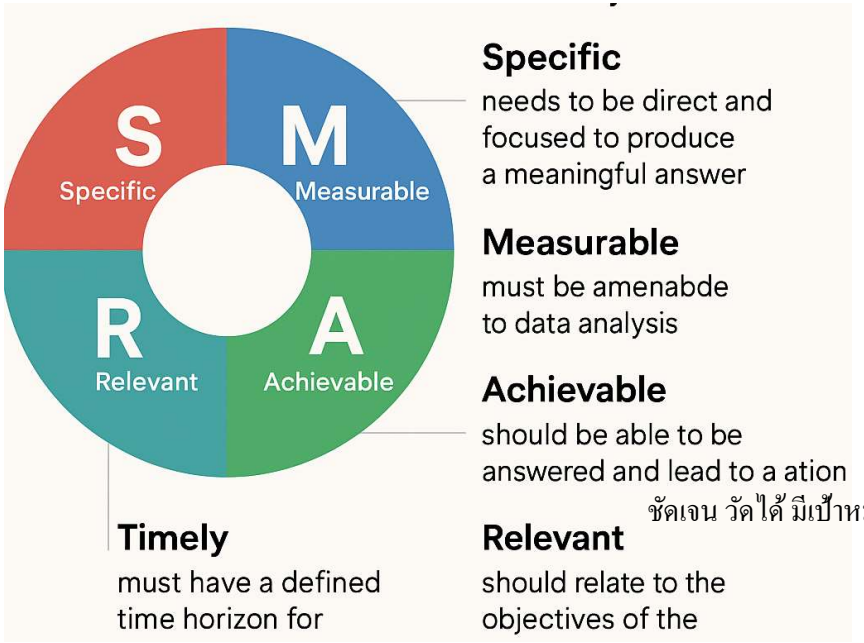
ตารางนี้ช่วยให้เห็นชัดว่า Analytics Mindset ในทางบัญชี มีรากฐานคล้ายกับ Scientific Method ที่ใช้ในงานวิจัยวิทยาศาสตร์



1. ASK THE RIGHT QUESTION

คำนิยามของข้อมูล (Data) ว่าเป็นข้อเท็จจริงที่ถูกรวบรวม (Collected) บันทึก (Recorded) จัดเก็บ (Stored) และประมวลผล (Processed) โดยระบบ ดังนั้น ข้อมูลเพียงลำพังยังมีคุณค่าไม่มากนัก แต่เมื่อข้อมูลถูกแปลงเป็นสารสนเทศ (Information) จึงจะสร้างคุณค่าได้

การเริ่มต้นกระบวนการเปลี่ยนข้อมูลให้เป็นสารสนเทศ จำเป็นต้องมีคำถาม (Question) หรือผลลัพธ์ที่ต้องการ (Desired Outcome) การตั้งคำถามที่ถูกต้อง (Asking the Right Question) ถือเป็นก้าวแรกของ กรอบความคิดเชิงการวิเคราะห์ (Analytics Mindset) เพื่อกำหนดว่า “คำถามที่ดี” (Good Question) หรือ “คำถามที่ถูกต้อง” (Right Question) ในบริบทของการวิเคราะห์ข้อมูล (Data Analytics) เป็นอย่างไร คำถามควรถูกออกแบบตามหลักการ SMART ได้แก่:



ตัวอย่าง

แบบทั่วไป (ไม่ชัด)	แบบฉลาด (Smart Question)
ยอดขายเราเป็นยังไง?	ยอดขาย Q2/2025 ของสินค้ากลุ่ม SUV ในไทย เพิ่ม/ลดกี่ % เมื่อเทียบกับ Q2/2024?
ลูกค้าลดลงไหม?	อัตราการ Churn ของลูกค้า SME ในกลุ่ม Telecom ช่วง H1/2025 เท่าไร และสาเหตุหลักคืออะไร?
ต้นทุนสูงไหม?	ต้นทุนต่อหน่วย (Unit Cost) ของ Product A ในโรงงานเชียงใหม่ Q1/2025 สูงกว่ามาตรฐานที่ตั้งไว้ (120 บาท/หน่วย) หรือไม่?
พนักงานแฮปปี้ไหม?	ค่า Employee Engagement Score ปี 2025 ในฝ่าย IT เทียบกับปี 2024 แตกต่างกันอย่างไร?

2. EXTRACT, TRANSFORM, AND LOAD RELEVANT DATA

กระบวนการ ETL หรือการ ดึง (Extract), แปลง (Transform) และ บรรทุก (Load) ข้อมูล ถือเป็นขั้นตอนที่ใช้เวลามากที่สุดของการสร้าง Analytics Mindset เนื่องจากกระบวนการนี้แตกต่างกันไปตามแต่ละโปรแกรม ฐานข้อมูล หรือระบบที่จัดเก็บและใช้งานข้อมูล ทำให้ต้องมีการส่งออก (Export) แปลง (Transform) และบรรทุก (Load) เข้าสู่ระบบใหม่ซ้ำหลายครั้ง

การดึงข้อมูล (Extracting Data)

ขั้นแรกของ ETL คือ การดึงข้อมูล (Extraction) ซึ่งแบ่งเป็น 3 ขั้นตอนหลัก:

1.เข้าใจความต้องการของข้อมูลและข้อมูลที่มีอยู่ (Understand Data Needs and Available Data)

- ต้องกำหนดความต้องการข้อมูลอย่างชัดเจน (สัมพันธ์กับ Analytics Mindset ข้อแรกคือ Asking the Right Question) หากกำหนดไม่ดี อาจดึงข้อมูลผิดหรือไม่ครบ ต้องทำ ETL ซ้ำ ⟶ เสียเวลาและทรัพยากร

2.ดำเนินการดึงข้อมูล (Perform the Data Extraction)

- ต้องเข้าใจแหล่งที่อยู่ของข้อมูล (Location) การเข้าถึง (Accessibility) และโครงสร้าง (Structure)
- องค์กรนิยมจัดเก็บข้อมูลในรูปแบบ:
  - ✓ Data Warehouse ⟶ โกดังข้อมูลเชิงโครงสร้าง (Structured Data) จากหลายแหล่ง
  - ✓ Data Mart ⟶ คลังข้อมูลย่อย แบ่งตามภูมิภาค/หน้าที่ (เช่น Sales Mart, Marketing Mart)
  - ✓ Data Lake ⟶ แหล่งรวมข้อมูลทุกประเภท (Structured, Semi-structured, Unstructured)

3.ตรวจสอบคุณภาพการดึงและบันทึก (Verify and Document Extraction)

- ต้องตรวจสอบคุณภาพของข้อมูลที่ดึงมา และบันทึกขั้นตอนเพื่อความโปร่งใส

ประเภทของข้อมูล

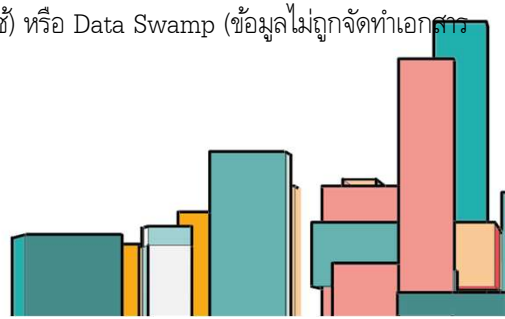
- Structured Data: ข้อมูลที่มีโครงสร้างแน่นอน เช่น General Ledger, Relational Database, Spreadsheet
- Semi-structured Data: ข้อมูลที่มีรูปแบบบางส่วน เช่น CSV, XML, JSON, Log Files
- Unstructured Data: ข้อมูลไร้โครงสร้าง เช่น รูปภาพ, วิดีโอ, เสียง, โซเชียลมีเดีย, เอกสาร

ข้อควรระวัง

Data Warehouse ⟶ ใหญ่และซับซ้อนมาก (เช่น Facebook ปี 2014 มี 300 Petabytes ใน 800,000 ตาราง)

Data Mart ⟶ เล็กกว่า เข้าถึงเร็วกว่า และควบคุมสิทธิ์ผู้ใช้ได้ดีกว่า

Data Lake ⟶ ยืดหยุ่นสูง แต่ถ้าไม่มีการจัดการ อาจกลายเป็น Dark Data (ข้อมูลที่มีแต่ไม่ได้ใช้) หรือ Data Swamp (ข้อมูลไม่ถูกจัดทำเอกสาร ทำให้ใช้ไม่ได้)



EXTRACT, TRANSFORM, AND LOAD RELEVANT DATA

ตารางเปรียบเทียบ Data Warehouse, Data Mart และ Data Lake

ประเภท (Type)	โครงสร้าง (Structure)	ประเภทข้อมูล (Data Types)	จุดแข็ง (Strengths)	ข้อควรระวัง (Cautions)
Data Warehouse	โครงสร้างชัดเจน (Highly Structured) เหมาะสำหรับ Reporting / BI	ข้อมูลเชิงโครงสร้าง (Structured Data) เช่น General Ledger, Relational Database, Spreadsheet	- รวมข้อมูลจากหลายแหล่งในองค์กร - รองรับการวิเคราะห์เชิงลึก - มาตรฐานสูงและน่าเชื่อถือ	- ขนาดใหญ่มาก ⟶ เข้าถึงช้า - ค่าใช้จ่ายสูงในการจัดเก็บและบำรุงรักษา
Data Mart	โครงสร้างชัดเจน แต่ขอบเขตเล็กกว่า (Smaller, Structured)	ข้อมูลเชิงโครงสร้าง (Structured Data) เฉพาะด้าน เช่น Sales Mart, Marketing Mart	- ขนาดเล็ก ⟶ เข้าถึงเร็ว - ควบคุมสิทธิ์การเข้าถึงง่าย - ตอบโจทย์เฉพาะหน่วยงาน	- มักไม่ครอบคลุมข้อมูลทั้งองค์กร - อาจเกิดความซ้ำซ้อนของข้อมูลระหว่าง Mart ต่าง ๆ
Data Lake	ยืดหยุ่นสูง (Flexible, Scalable) เพิ่มพลัง AI/ML, Advanced Analytics	รวมได้ทุกประเภท: - Structured (โครงสร้างชัดเจน) - Semi-structured (CSV, XML, JSON) - Unstructured (ภาพ, วิดีโอ, เสียง, Social Media)	- รองรับข้อมูลทุกประเภท - รวมข้อมูลภายในและภายนอกองค์กร - รองรับการประมวลผลแบบ Big Data & AI	- เสี่ยงกลายเป็น Dark Data (เก็บแต่ไม่ใช้) - เสี่ยงกลายเป็น Data Swamp (ข้อมูลไม่ถูกจัดทำเอกสาร ใช้ต่อไม่ได้)

EXTRACT, TRANSFORM, AND LOAD RELEVANT DATA

การออกแบบ Data Warehouse, Data Mart และ Data Lake

การออกแบบ Data Warehouse, Data Mart และ Data Lake ได้หลายวิธี เช่น:

1.Data Warehouse ⟶ Data Mart

- ออกแบบ Data Warehouse ให้เชื่อมต่อกับแหล่งข้อมูลธุรกรรม (Transaction Data) หรือข้อมูลเชิงโครงสร้าง (Structured Data) ทั้งหมด
- จากนั้นสร้าง Data Mart โดยอ้างอิงจากข้อมูลใน Data Warehouse

2.Data Mart ⟶ Data Warehouse

- ออกแบบ Data Mart ให้เชื่อมต่อโดยตรงกับแหล่งข้อมูลธุรกรรม
- นำข้อมูลจาก Data Mart ต่าง ๆ มารวม (Aggregate) เป็น Data Warehouse

3.Data Warehouse ⟶ Data Lake (ทั่วไปที่สุด)

- โดยทั่วไป Data Warehouse ถือเป็นฐานหลัก และมักถูกใช้ในการสร้าง Data Lake

4.Independent Access

- องค์กรอาจออกแบบให้ Data Warehouse, Data Mart และ Data Lake เข้าถึงแหล่งข้อมูลโดยตรงอย่างอิสระ

การเข้าใจโครงสร้างการออกแบบเหล่านี้เป็นสิ่งสำคัญ เพราะจะช่วยผู้ใช้งานระบุได้ว่าข้อมูลที่ต้องการเก็บไว้ที่ใด และสามารถเข้าถึงได้อย่างถูกต้อง

ความสำคัญของ Data Dictionary และ Metadata

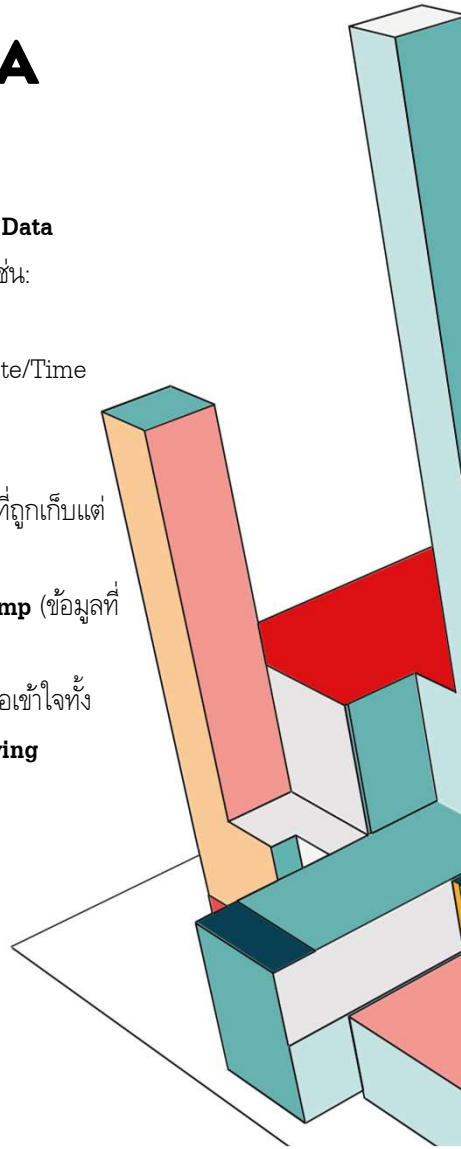
เพื่อทำความเข้าใจโครงสร้างข้อมูล วิธีที่ดีที่สุดคือการศึกษา **Data Dictionary** ซึ่งเก็บ **Metadata** (ข้อมูลที่ใช้อธิบายข้อมูล) เช่น:

- ✓ จำนวนตัวอักษรสูงสุดในฟิลด์ (Field Length)
- ✓ ประเภทข้อมูล (Data Type) เช่น Integer, Text, Date/Time
- ✓ รูปแบบข้อมูล (Data Format)

ประโยชน์ของ Metadata ที่ถูกต้องและทันสมัย

- ✓ ช่วยป้องกันไม่ให้ข้อมูลกลายเป็น **Dark Data** (ข้อมูลที่ถูกเก็บแต่ไม่ได้วิเคราะห์)
- ✓ ลดความเสี่ยงที่ Data Lake จะกลายเป็น **Data Swamp** (ข้อมูลที่ไร้การจัดทำเอกสาร ใช้งานต่อไม่ได้)

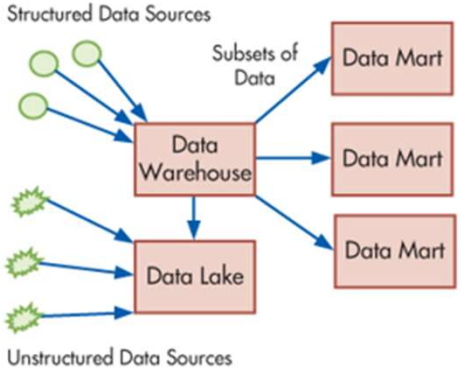
ดังนั้น การตรวจสอบ Data Dictionary อย่างรอบคอบ เพื่อเข้าใจทั้ง **ข้อมูล (Data)** และ **โครงสร้างพื้นฐานของข้อมูล (Underlying Objects)** จึงเป็นขั้นตอนสำคัญของกระบวนการ **ETL**



EXTRACT, TRANSFORM, AND LOAD RELEVANT DATA

แผนภาพแสดงโครงสร้างการเชื่อมโยง Data Warehouse – Data Mart – Data Lake

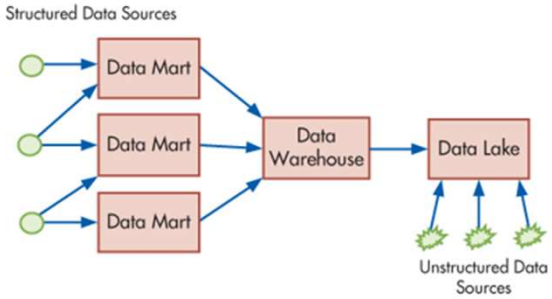
แบบที่ 1



จาก Enterprise Data Warehouse ⟶ Sales Data Mart

- DWH เก็บข้อมูลทั้งหมด: ลูกค้า, สินค้า, inventory, accounting, HR
- Sales Data Mart ⟶ ดึงเฉพาะข้อมูล **ยอดขาย, ลูกค้า, campaign**
- ใช้ตอบคำถาม: *"Top 10 ลูกค้าปีนี้?"*, *"ยอดขายตาม region เดือนนี้เทียบกับปีก่อน?"*

แบบที่ 2



1. Retail (ค้าปลีก)

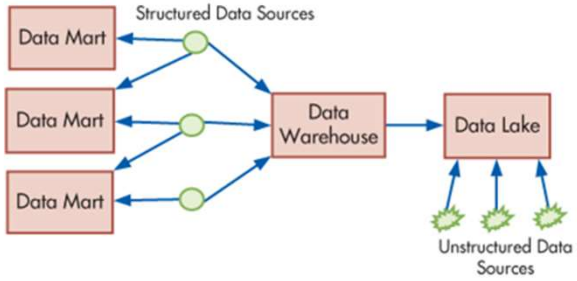
เดิม:

- Data Warehouse เก็บข้อมูลยอดขาย POS, Inventory, Finance
- ใช้ทำรายงาน KPI เช่น ยอดขายรายวัน, Margin, Stock level

เปลี่ยนเป็น Data Lake:

- รวมข้อมูลจาก POS + clickstream บน e-commerce + social media feedback + IoT sensor จากห้าง
- นำไปทำ **customer 360 view** และ **personalized recommendation (AI/ML)**

แบบที่ 3



2. Banking (ธนาคาร)

เดิม:

DWH เก็บ transaction, loan portfolio, balance sheet ⟶ ใช้ทำรายงานตามกฎหมาย (regulatory reporting)

เปลี่ยนเป็น Data Lake:

- เก็บข้อมูลธุรกรรมแบบ real-time, call center logs, mobile app clickstream
- ใช้ทำ **fraud detection, customer churn prediction, credit scoring model**

3. Manufacturing (โรงงาน/อุตสาหกรรม)

เดิม:

- DWH เก็บข้อมูลการผลิต, ต้นทุน, BOM ⟶ ใช้ทำ cost analysis

เปลี่ยนเป็น Data Lake:

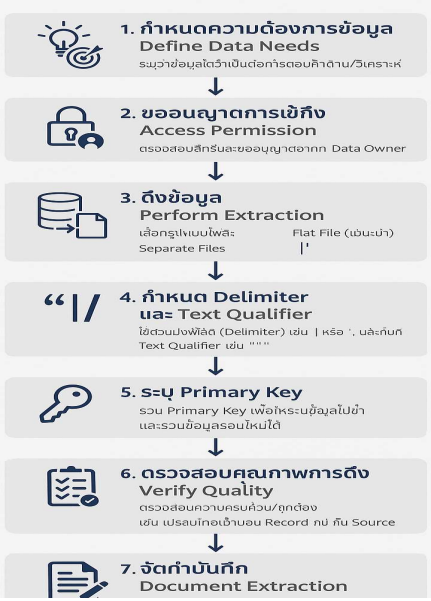
- เก็บ IoT sensor data จากเครื่องจักร, maintenance logs, weather data
- ใช้ทำ **predictive maintenance, yield optimization**

ตาราง: ตัวอย่าง METADATA ใน DATA DICTIONARY

Field Name (ชื่อฟิลด์)	Data Type (ชนิดข้อมูล)	Allowed Characters (อักขระที่อนุญาต)	Format (รูปแบบข้อมูล)	Example (ตัวอย่างค่า)
Customer_ID	Integer (จำนวนเต็ม)	ตัวเลข 0–9 เท่านั้น	8 หลัก (Fixed Length)	00012345
Customer_Name	Text (ข้อความ)	ตัวอักษร A–Z, a–z, เว้นวรรค	สูงสุด 100 ตัวอักษร	Somchai Prasert
Invoice_Date	Date/Time (วันที่/เวลา)	ตัวเลขและตัวแบ่งวันที่	YYYY-MM-DD	2025-09-02
Invoice_Amount	Decimal (ทศนิยม)	ตัวเลข 0–9 และ “.”	สูงสุด 2 ตำแหน่งทศนิยม	12500.75
Email_Address	Text (ข้อความ)	ตัวอักษร A–Z, a–z, ตัวเลข, @, .	สูงสุด 150 ตัวอักษร	somchai@email.com
Product_Code	Text/Alphanumeric (อักษรผสมตัวเลข)	ตัวอักษร A–Z และตัวเลข 0–9	10 ตัวอักษร (Fixed Length)	PRD2025001

EXTRACT,

TRANSFORM,
AND LOAD
RELEVANT
DATA



ตารางสรุปขั้นตอนการดึงข้อมูล (Extract Data)

ขั้นตอน (Step)	สิ่งที่ต้องทำ (What to Do)	เครื่องมือ/แนวทาง (Tools / Guidelines)
1. กำหนดความต้องการข้อมูล (Define Data Needs)	ระบุว่าข้อมูลใดจำเป็นต่อการตอบคำถาม/วิเคราะห์	เชื่อมโยงกับ Analytics Mindset – Ask the Right Question
2. ขออนุญาตการเข้าถึง (Access Permission)	ตรวจสอบสิทธิ์และขออนุญาตจาก Data Owner	Internal Control / Approval Workflow
3. ดึงข้อมูล (Perform Extraction)	เลือกรูปแบบไฟล์: - Separate Files - Flat File (แนะนำ)	Tools: SQL, Data Export Function, ETL Tool
4. กำหนด Delimiter และ Text Qualifier	ใช้ตัวแบ่งฟิลด์ (Delimiter) เช่น “	” หรือ “,” และกำกับด้วย Text Qualifier เช่น “ ”
5. ระบุ Primary Key	รวม Primary Key เพื่อให้ระบุข้อมูลไม่ซ้ำ และรวมข้อมูลรอบใหม่ได้	Relational Database Key
6. ตรวจสอบคุณภาพการดึง (Verify Quality)	ตรวจสอบความครบถ้วน/ถูกต้อง เช่น เปรียบเทียบจำนวน Record กับ Source	Batch Processing Controls / Record Count
7. จัดทำบันทึก (Document Extraction)	บันทึกขั้นตอนและผลการดึงข้อมูลเพื่อ Audit Trail	Data Extraction Log / ETL Documentation

EXTRACT,

TRANSFORM, AND LOAD RELEVANT DATA

Examples of Delimiters and Text Qualifiers

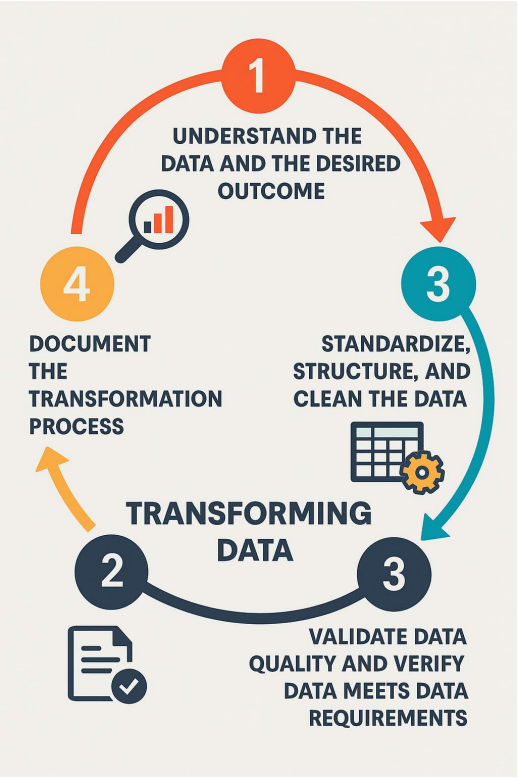
Pipe Delimited, Text Qualifiers						
Flat File, Header Row	Fname|Lname|PerformanceScore|"PerformanceReview"					
Flat File, Data Row	Renee|Armstrong|99|"Smiles a lot, Enthusiastic, could improve technique"					
Separated Data, Header Row	FName	LName	PerformanceScore	PerformanceReview		
Separated Data, Data Row	Renee	Armstrong	99	Smiles a lot, Enthusiastic, could improve technique		
Comma Delimited, No Text Qualifiers						
Flat File, Header Row	Fname,Lname,PerformanceScore,PerformanceReview					
Flat File, Data Row	Renee,Armstrong,99,Smiles a lot, Enthusiastic, could improve technique					
Separated Data, Header Row	FName	LName	PerformanceScore	PerformanceReview	**ERROR**	**ERROR**
Separated Data, Data Row	Renee	Armstrong	99	Smiles a lot	Enthusiastic	could improve technique
Comma Delimited, Text Qualifiers						
Flat File, Header Row	Fname,Lname,PerformanceScore,PerformanceReview					
Flat File, Data Row	Renee,Armstrong,99,"Smiles a lot, Enthusiastic, could improve technique"					
Separated Data, Header Row	FName	LName	PerformanceScore	PerformanceReview		
Separated Data, Data Row	Renee	Armstrong	99	Smiles a lot, Enthusiastic, could improve technique		

ตาราง BEST PRACTICES หลังการดึงข้อมูล (POST-EXTRACTION BEST PRACTICES)

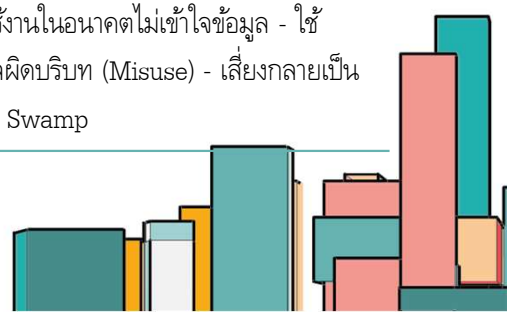
ขั้นตอน (Step)	สิ่งที่ต้องทำ (What to Do)	ประโยชน์ (Benefits)
1. Reperform Data Extraction	เลือก **Sample Records** แล้วทำการดึงข้อมูลซ้ำ จากนั้นเปรียบเทียบกับ **Full Data Extract**	- ยืนยันความถูกต้องของกระบวนการดึงข้อมูล - สร้างหลักฐานการตรวจสอบ (Audit Evidence)
2. Create New Data Dictionary	สร้าง Data Dictionary ใหม่จากผลการดึงข้อมูล - คัดลอกจาก Source Data Dictionary - อัปเดตฟิลด์ที่มีการเปลี่ยนแปลง	- ผู้ใช้เข้าใจข้อมูลได้ตรงกัน - รองรับการใช้งานต่อเนื่อง
3. Add Metadata	ระบุแหล่งที่มาของข้อมูล (Source) ใน Metadata และเพิ่มคำอธิบายสำหรับข้อมูลที่นำเข้าจากภายนอก	- ป้องกันการเกิด **Data Swamp** - เพิ่มความโปร่งใสในการใช้งานข้อมูล
4. Define New Fields Clearly	หากมีข้อมูลที่ไม่เคยถูกกำหนดไว้ใน Data Dictionary เดิม ต้องกำหนด Field ใหม่อย่างถูกต้อง	- ป้องกันความเข้าใจผิดในอนาคต - ช่วยให้นักวิเคราะห์ใช้ข้อมูลได้อย่างมีประสิทธิภาพ

EXTRACT,
TRANSFORM,
AND LOAD
RELEVANT
DATA

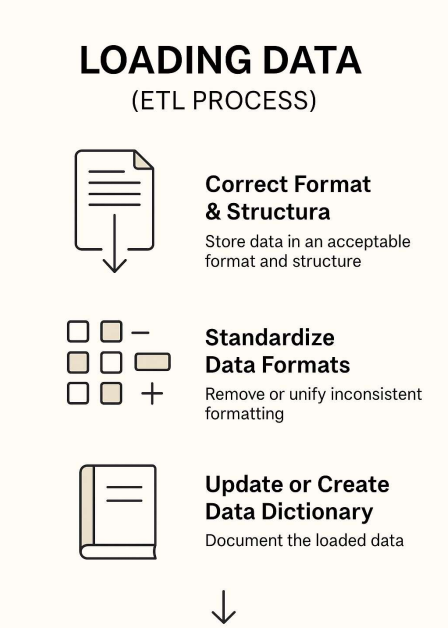
การแปลงข้อมูล (Data Transformation Process) คือการทำให้ข้อมูลมีมาตรฐาน (Standardizing) มีโครงสร้างที่ชัดเจน (Structuring) และการทำความสะอาดข้อมูล (Cleaning) เพื่อให้ข้อมูลอยู่ในรูปแบบที่พร้อมสำหรับการวิเคราะห์ กระบวนการนี้ใช้เวลาและมีความซับซ้อนสูง ภาพรวมของกระบวนการ 4 ขั้นตอน คือ:



ขั้นตอน (Step)	สิ่งที่ต้องทำ (What to Do)	ความเสี่ยงถ้าไม่ทำ (Risks if Skipped)
1. เข้าใจข้อมูลและผลลัพธ์ที่ต้องการ (Understand the Data & Desired Outcome)	- ศึกษาข้อมูลที่ Extract มา - สร้าง/ตรวจสอบ Data Dictionary - กำหนดข้อกำหนด เช่น Format, Delimiter, Filtering, Level of Detail	- เลือกข้อมูลผิดหรือไม่ครบถ้วน - ข้อมูลไม่ตอบโจทย์วิเคราะห์ - ต้องทำ ETL ซ้ำ เสียเวลาและทรัพยากร
2. ทำให้ข้อมูลมีมาตรฐาน จัดโครงสร้าง และทำความสะอาด (Standardize, Structure & Clean)	- รวมข้อมูลจากหลายระบบให้อยู่ใน Format เดียวกัน - ลบ/แก้ไขข้อมูลซ้ำ ข้อมูลผิดพลาด - สร้างความสม่ำเสมอ (Consistency)	- ข้อมูลไม่สามารถรวมกันได้ - วิเคราะห์ผิดพลาดเพราะโครงสร้างไม่ตรงกัน - ใช้เวลามากในการแก้ไขภายหลัง
3. ตรวจสอบคุณภาพและความสอดคล้อง (Validate Data Quality & Requirements)	- ตรวจสอบความถูกต้อง (Accuracy) และความครบถ้วน (Completeness) - ตรวจสอบความสอดคล้องกับวัตถุประสงค์ (Data Requirements)	- ข้อมูลแม้ถูกต้อง แต่ไม่ตรงกับความต้องการวิเคราะห์ ⟶ ใช้ไม่ได้ - เกิดข้อผิดพลาดในการตัดสินใจ
4. จัดทำเอกสาร (Document the Process)	- อัปเดต Data Dictionary - บันทึกการเปลี่ยนแปลงและเหตุผล - จัดทำเอกสารสำหรับการใช้งานต่อไป	- ผู้ใช้งานในอนาคตไม่เข้าใจข้อมูล - ใช้ข้อมูลผิดบริบท (Misuse) - เสี่ยงกลายเป็น Data Swamp



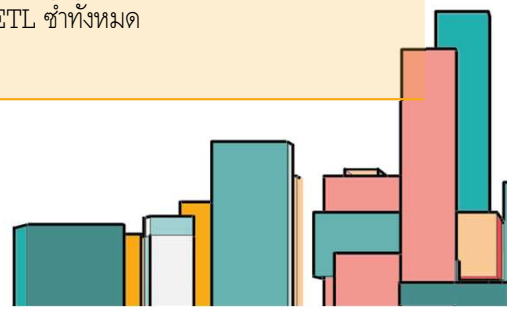
EXTRACT, TRANSFORM, AND **LOAD RELEVANT DATA**



การโหลดข้อมูล (Loading Data)

เมื่อข้อมูลถูกจัดโครงสร้าง (Structured) และทำความสะอาด (Cleaned) แล้ว ก็พร้อมที่จะถูกนำเข้า (Imported) ไปยังเครื่องมือหรือโปรแกรมที่จะใช้ในการวิเคราะห์ หากข้อมูลได้รับการแปลง (Transformed) อย่างถูกต้องแล้ว ขั้นตอนนี้มักจะดำเนินการได้อย่างรวดเร็วและไม่ซับซ้อน อย่างไรก็ตาม มีข้อควรพิจารณาที่สำคัญดังนี้:

ขั้นตอน (Step)	สิ่งที่ต้องทำ (What to Do)	ความเสี่ยงถ้าไม่ทำ (Risks if Skipped)
1. จัดเก็บข้อมูลในรูปแบบ/โครงสร้างที่ถูกต้อง (Correct Format & Structure)	- บันทึกไฟล์ให้อยู่ในรูปแบบที่ระบบรองรับ เช่น Text File + Delimiters, XBRL, JSON, Relational Tables - โหลดข้อมูลตามลำดับที่สอดคล้องกับ Referential Integrity	- โปรแกรมไม่สามารถนำเข้าข้อมูลได้ - ละเมิดกฎ Referential Integrity ⟶ เกิดข้อมูลผิดพลาด
2. ทำให้รูปแบบข้อมูลเป็นมาตรฐาน (Standardize Data Formats)	- ลบคอมม่า (,) ที่คั่นหลักพัน - ใช้เครื่องหมายลบ (–) แทนวงเล็บสำหรับค่าติดลบ - ใช้ Date Format มาตรฐานเดียวกัน	- โปรแกรมตีความข้อมูลผิดพลาด - เกิดค่าผิด เช่น ติดลบไม่ถูกต้อง หรือวันที่สลับเดือน/วัน
3. อัปเดต/สร้าง Data Dictionary (Update or Create Data Dictionary)	- จัดทำ Data Dictionary ใหม่ที่อธิบายทุกฟิลด์ - บันทึก Metadata เช่น แหล่งที่มาของข้อมูล และการเปลี่ยนแปลงระหว่างกระบวนการ	- ผู้ใช้งานภายหลังไม่เข้าใจข้อมูล - ใช้ข้อมูลผิดบริบท - อาจต้องทำ ETL ซ้ำทั้งหมด



3. APPLY APPROPRIATE DATA ANALYTIC TECHNIQUES

ตารางที่ 1: 4 ประเภทของ Data Analytics

ประเภท (Type)	คำถามหลัก (Key Question)	จุดประสงค์ (Purpose)	ตัวอย่าง (Example)
Descriptive Analytics	What happened? / What is happening?	ทำความเข้าใจอดีตหรือสถานการณ์ปัจจุบัน	ROI, Gross Margin
Diagnostic Analytics	Why did this happen?	ค้นหาความสัมพันธ์เชิงเหตุผล (Causal Relationships)	IT Budget ↑ ⟶ Efficiency ↑
Predictive Analytics	What might happen?	พยากรณ์สิ่งที่จะเกิดขึ้นในอนาคต	ราคาหุ้น, อัตราแลกเปลี่ยน
Prescriptive Analytics	What should be done?	เสนอแนวทางการตัดสินใจหรือการลงมือทำ	Loan Approval Algorithm

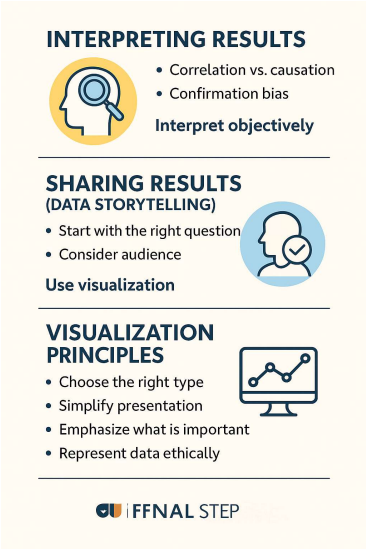
3. APPLY APPROPRIATE DATA ANALYTIC TECHNIQUES

ตารางที่ 2: ระดับทักษะ

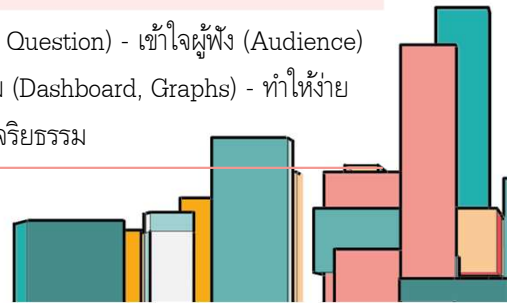
ระดับ (Level)	ความหมาย (Meaning)	ตัวอย่างการปฏิบัติ (Example in Practice)
Awareness (การรับรู้)	รู้จักว่าเครื่องมือ/เทคนิคคืออะไรและทำอะไรได้ แต่ทำเองไม่ได้	รู้ว่า Python ใช้ทำ Data Analysis ได้ แต่ยังเขียนโค้ดเองไม่ได้
Working Knowledge (ความรู้เชิงปฏิบัติ)	เคยทำงานคล้ายกันมาก่อน สามารถทำได้ แต่ต้องทบทวนหรือหาคู่มือช่วย	เขียน SQL Query ง่าย ๆ ได้ แต่ต้องเปิดคู่มือเมื่อเจอโจทย์ซับซ้อน
Mastery (ความเชี่ยวชาญ)	ทำได้ทันที เข้าใจลึก สามารถรับผิดชอบโครงการเต็มรูปแบบ	ได้รับงาน Data Analytics แล้วสามารถใช้ Python, SQL, Machine Learning ทำงานได้ครบถ้วน

4. INTERPRET AND SHARE THE RESULTS WITH STAKEHOLDERS

การตีความผลลัพธ์ (Interpreting Results) และ การสื่อสารผลลัพธ์ (Sharing Results / Data Storytelling) ตารางนี้ช่วยให้เห็นว่า Interpreting Results = Internal Understanding ในขณะที่ Sharing Results = External Communication



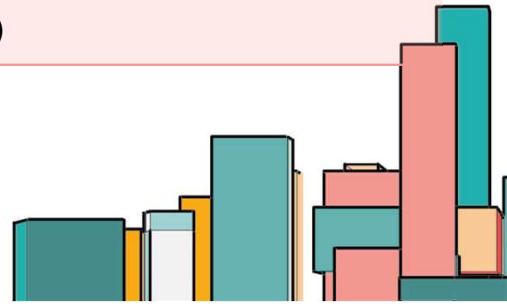
ด้าน (Aspect)	การตีความผลลัพธ์ (Interpreting Results)	การสื่อสารผลลัพธ์ (Sharing Results / Data Storytelling)
จุดมุ่งหมาย (Objective)	เข้าใจความหมายของผลวิเคราะห์อย่างถูกต้อง	ถ่ายทอดผลวิเคราะห์ให้ Stakeholders เข้าใจและตัดสินใจได้
ความท้าทาย (Challenges)	- ความสับสนระหว่าง Correlation vs Causation และ Confirmation Bias (ตีความตามความเชื่อ/ความปรารถนา)	- ความแตกต่างของผู้ฟัง (ประสบการณ์, ความรู้ด้าน Data) - เลือกรูปแบบการเล่าเรื่อง/Visualization ที่เหมาะสม
ตัวอย่าง (Example)	- ยอดขายอุปกรณ์หิมะสูงขึ้นเพราะโฆษณา? หรือเพราะเข้าหน้าหนาว? - ผู้จัดการให้คะแนนพนักงานเกินจริงเพราะเป็นเพื่อน	- CIO ต้องการเห็นแนวโน้มค่าใช้จ่ายด้านเทคโนโลยี ⟶ ใช้ Line Chart บน Dashboard แทนการรายงานเป็นตัวเลข
แนวทางที่ดี (Best Practices)	- ฝึกการตีความอย่างเป็นกลาง (Objective) - ใช้ความรู้และการฝึกฝนเพื่อเข้าใจผลลัพธ์อย่างแท้จริง	- เริ่มจากคำถามต้นทาง (Right Question) - เข้าใจผู้ฟัง (Audience) - ใช้ Visualization ที่เหมาะสม (Dashboard, Graphs) - ทำให้ง่าย เน้นประเด็น และสื่อสารอย่างมีจริยธรรม



4. INTERPRET AND SHARE THE RESULTS WITH STAKEHOLDERS

ตารางสรุป Interpreting vs Sharing vs Visualization Principles

หมวด (Category)	จุดมุ่งหมาย (Objective)	ความท้าทาย/ความเสี่ยง (Challenges/Risks)	แนวทางที่ดี (Best Practices)
Interpreting Results (การตีความผลลัพธ์)	เข้าใจความหมายที่แท้จริงของผลวิเคราะห์	- สับสนระหว่าง **Correlation vs Causation** - **Confirmation Bias**: ตีความเข้าข้างความเชื่อเดิม	- ตีความอย่าง **เป็นกลาง (Objective)** - ฝึกฝนเพื่อเพิ่มความแม่นยำ
Sharing Results / Data Storytelling (การสื่อสารผลลัพธ์)	ถ่ายทอดผลวิเคราะห์ให้ Stakeholders เข้าใจง่ายและใช้ตัดสินใจได้	- ระดับความรู้และประสบการณ์ของผู้ฟังต่างกัน - เลือกรูปแบบการเล่าเรื่องไม่ตรงกลุ่มเป้าหมาย	- เริ่มจาก **คำถามต้นทาง (Right Question)** - วิเคราะห์ **ผู้ฟัง (Audience)** - ใช้ **Visualization** ที่เหมาะสม
Visualization Principles (หลักการ Visualization)	ใช้กราฟ/แผนภาพช่วยสื่อสารข้อมูลให้ชัดเจนและทรงพลัง	- นำเสนอซับซ้อนเกินไป - ขาดจริยธรรมในการแสดงข้อมูล	- เลือกชนิด Visualization ให้เหมาะ - ทำให้เรียบง่าย (Simplify) - เน้นสิ่งสำคัญ (Emphasize) - แสดงข้อมูลอย่าง **มีจริยธรรม (Ethical)**

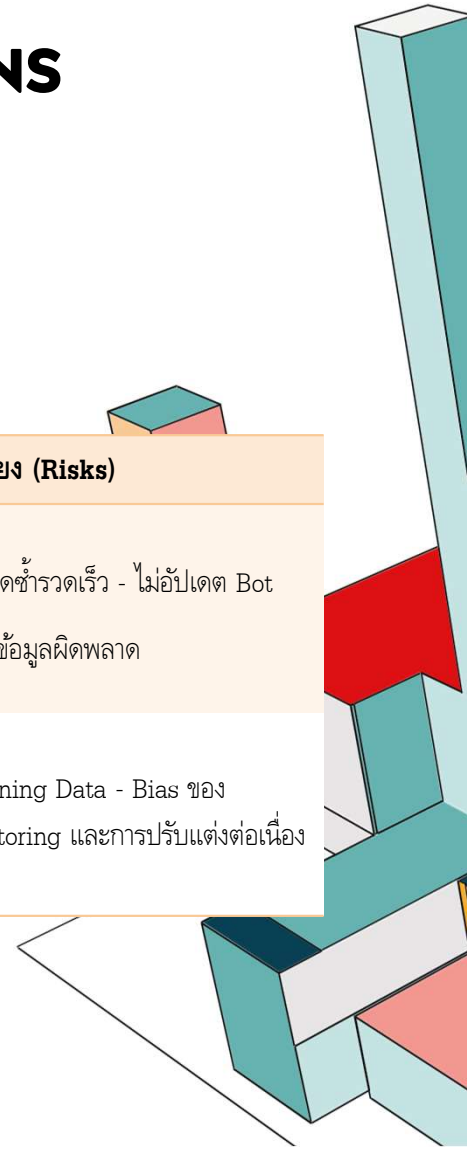


ADDITIONAL DATA ANALYTICS CONSIDERATIONS

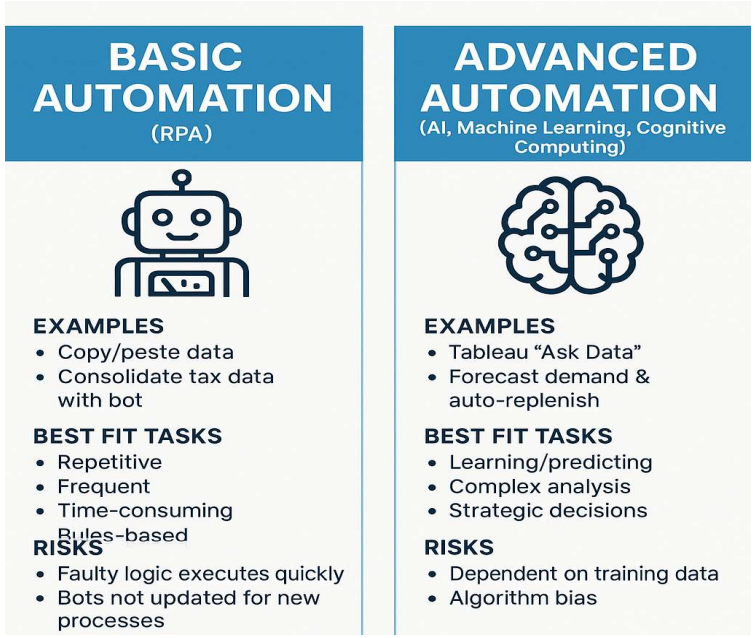
- **การทำงานอัตโนมัติ (Automation)**

- **ความหมาย:**
  Automation คือการใช้เครื่องจักรหรือซอฟต์แวร์ให้ทำงานโดยอัตโนมัติแทนที่มนุษย์ เช่น เดิมต้องคัดลอกและวางข้อมูลด้วยมือ
  ⟶ เขียนโปรแกรมให้ทำแทนได้

ประเภท (Type)	ตัวอย่าง (Examples)	งานที่เหมาะสม (Best Fit Tasks)	ความเสี่ยง (Risks)
Basic Automation (RPA – Robotic Process Automation)	- Copy/Paste Data - Consolidate Tax Data ด้วย Bot - Balance Journal Entries	- งานซ้ำซาก (Repetitive) - ทำบ่อย (Frequent) - ใช้เวลานาน (Time-consuming) - มีกฎเกณฑ์ชัดเจน (Rules-based)	- หาก Logic ผิด ⟶ ทำผิดซ้ำรวดเร็ว - ไม่อัปเดต Bot เมื่อ Process เปลี่ยน ⟶ ข้อมูลผิดพลาด
Advanced Automation (AI, Machine Learning, Cognitive Computing)	- Tableau “Ask Data” - Aera Technology (Self-driving enterprise) - Forecast Demand & Auto-replenish Inventory	- งานที่ต้อง **เรียนรู้/คาดการณ์** - วิเคราะห์ข้อมูลซับซ้อน - ตัดสินใจเชิงกลยุทธ์	- ความแม่นยำขึ้นอยู่กับ Training Data - Bias ของ Algorithm - ต้องการ Monitoring และการปรับแต่งต่อเนื่อง



ADDITIONAL DATA ANALYTICS CONSIDERATIONS



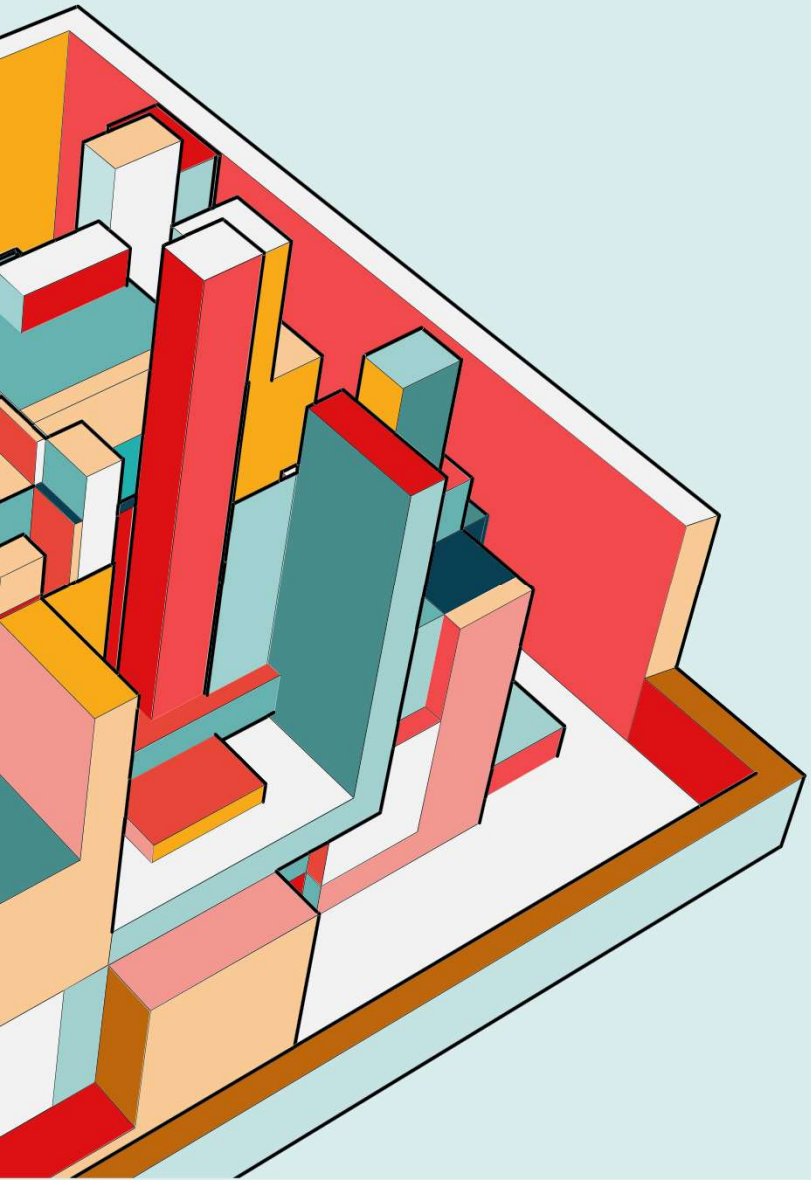
กระบวนการ Automation ส่งผลต่อมนุษย์ คือ (Human Element of Automation)

- พนักงานมักกังวลว่าการนำ Automation มาใช้จะทำให้ **สูญเสียงาน**
- จริงอยู่ Automation อาจถูกใช้เพื่อลดจำนวนพนักงาน แต่ก็สามารถใช้เพื่อลด **งานซ้ำซากและน่าเบื่อ** เพื่อให้พนักงานมีเวลาไปทำงานที่ **สร้างคุณค่า (Value-Added Tasks)** มากกว่า
- องค์กรควร **พิจารณาผลกระทบต่อบุคลากร** และเตรียมการบริหารจัดการความกังวลเหล่านี้ ก่อนเริ่มโครงการ Automation

ADDITIONAL DATA ANALYTICS CONSIDERATIONS

Data Analytics ไม่ใช่เครื่องมือที่เหมาะสมเสมอไป (Data Analytics Is Not Always the Right Tool)

- ข้อมูลที่น่าเชื่อถืออาจ **ไม่มี** สำหรับคำถามบางประเภท ⟶ ทำให้ Data Analytics ไม่สามารถให้คำตอบที่ถูกต้องได้
- **Human Judgment & Intuition** อาจมีความสำคัญ เช่น การประเมินปัจจัยด้านอารมณ์ ความรู้สึก หรือบริบทที่ไม่สามารถวัดผลเชิงตัวเลขได้
- ตัวอย่าง: หาก Data Analysis แสดงว่าการโกง (Fraud) จะทำให้ได้ผลตอบแทนสูงและมีโอกาสถูกจับน้อย ⟶ CEO ที่มีจริยธรรมจะรู้ว่าการกระทำดังกล่าว **ผิดแม้ผลวิเคราะห์บอกว่า "คุ้ม"**



TRANSFORMING DATA

INTRODUCTION

ปัญหาคุณภาพข้อมูล (Data Quality Problem)

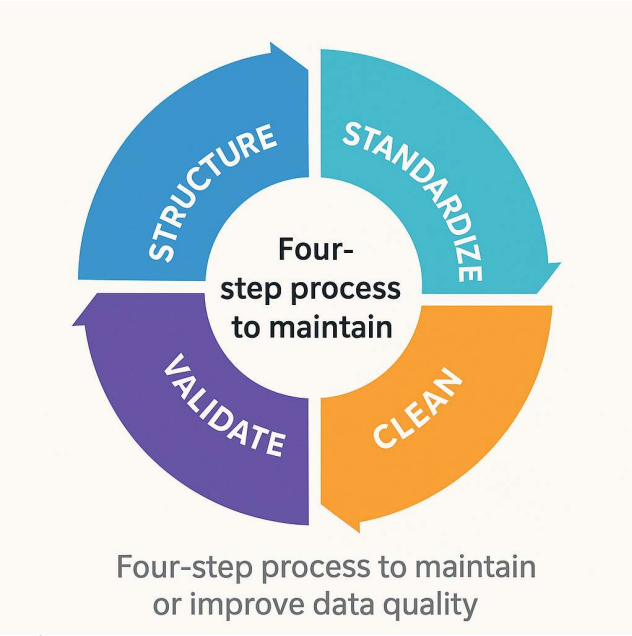
- IBM ประมาณการว่า **ข้อมูลคุณภาพต่ำ** สร้างความสูญเสียต่อเศรษฐกิจสหรัฐอเมริกา **3.1 ล้านล้านดอลลาร์/ปี**
- **1 ใน 3** ของผู้ตัดสินใจทางธุรกิจ **ไม่เชื่อถือข้อมูล** ที่ใช้ตัดสินใจ
- ผลสำรวจโดย Experian: ผู้บริหารเชื่อว่า **26% ของข้อมูลทั้งหมดไม่ถูกต้อง** และปัญหากำลังทวีความรุนแรงขึ้น

สาเหตุของข้อมูลคุณภาพต่ำ (Dirty Data Causes)

- **ข้อผิดพลาดเชิงเทคนิค:** ข้อมูลไม่สมบูรณ์ (Incomplete), ล้าสมัย (Outdated), ซ้ำซ้อน (Duplicate), หรือผิดพลาดเล็กน้อย (Typos, Spelling Mistakes)
- **ปัจจัยภายนอก:** การเปลี่ยนแปลงในโลกจริง (เช่น ลูกค้าเปลี่ยนเบอร์โทรศัพท์)
- **การรวมข้อมูล:** จากหลายแหล่งที่ใช้รูปแบบต่างกัน
- **ข้อผิดพลาดในการจัดการ:** Human Errors ในการบันทึกหรือประมวลผลข้อมูล

แนวทางแก้ไข (Solutions)

- **อุดมคติ:** เก็บและจัดเก็บข้อมูลที่สะอาดตั้งแต่แรก (Capture & Store Clean Data)
- **การควบคุมภายใน:** (ตามบทที่ 10–13) เพื่อสนับสนุนการเก็บข้อมูลที่ถูกต้อง
- **ความเป็นจริง:** แม้มี Internal Controls ข้อมูลก็ยังอาจกลายเป็น Dirty Data ⟶ จึงต้องใช้กระบวนการ **Data Transformation**



กระบวนการฟื้นฟูคุณภาพข้อมูล (Four-Step Data Transformation Process)

1.Structure the Data – จัดโครงสร้างข้อมูลให้เป็นระบบ

2.Standardize the Data – ทำให้ข้อมูลอยู่ในรูปแบบมาตรฐานเดียวกัน

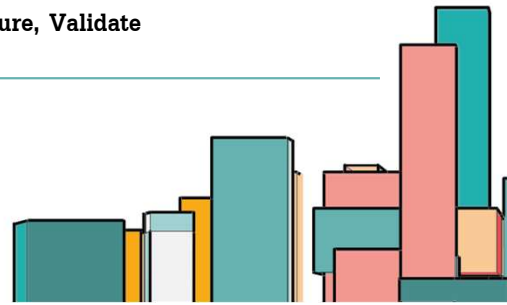
3.Clean the Data – แก้ไขหรือลบข้อมูลที่ผิดพลาด ซ้ำซ้อน หรือไม่สมบูรณ์

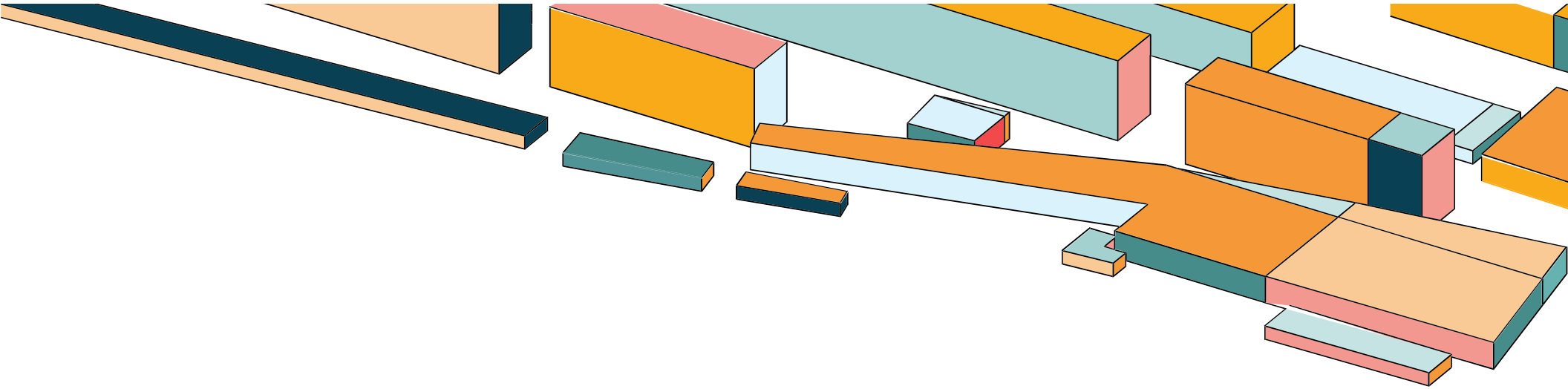
4.Validate the Data – ตรวจสอบคุณภาพ ความถูกต้อง และความสมบูรณ์ของข้อมูล

ATTRIBUTES OF HIGH-QUALITY DATA

คุณลักษณะของข้อมูลคุณภาพสูง (High-Quality Data Attributes)
ในทุก ๆ การทำงานกับข้อมูล สิ่งสำคัญคือ **ตรวจสอบให้แน่ใจว่าข้อมูลมีคุณลักษณะของข้อมูลคุณภาพสูง ดังนี้**

คุณลักษณะ (Attribute)	ความหมาย (Meaning)	ตัวอย่าง (Example)	ขั้นตอน Transformation ที่เกี่ยวข้อง
Accuracy (ความถูกต้อง)	ข้อมูลสะท้อนความจริงได้ถูกต้อง	เบอร์โทรลูกค้าปัจจุบันตรงกับความจริง	**Clean, Validate**
Completeness (ความสมบูรณ์)	ไม่มีข้อมูลที่ขาดหาย	รายการคำสั่งซื้อบันทึกทั้งชื่อ, ที่อยู่, วันที่	**Structure, Clean**
Consistency (ความสอดคล้องกัน)	ข้อมูลเหมือนกันในทุกระบบ/ทุกที่	“Customer ID” เหมือนกันทั้ง ERP และ CRM	**Standardize**
Timeliness (ความทันเวลา)	ข้อมูลเป็นปัจจุบันและพร้อมใช้งาน	ยอดขายรายวันอัปเดตภายในวันถัดไป	**Validate**
Uniqueness (ความไม่ซ้ำซ้อน)	ไม่มีข้อมูลซ้ำที่ไม่จำเป็น	ลูกค้า 1 คน = 1 Record เท่านั้น	**Clean**
Relevance (ความเกี่ยวข้อง)	ข้อมูลสอดคล้องกับวัตถุประสงค์ที่ใช้	ใช้ข้อมูล Vendor ไม่ใช่ Customer สำหรับวิเคราะห์ Supplier Performance	**Structure, Validate**





DATA STRUCTURING

DATA STRUCTURING

ความหมาย (Definition)Data Structure = วิธีการจัดเก็บข้อมูล รวมถึง ความสัมพันธ์ระหว่างฟิลด์ข้อมูลต่าง ๆData Structuring = กระบวนการปรับเปลี่ยนการจัดระเบียบและความสัมพันธ์ของข้อมูล เพื่อเตรียมพร้อมสำหรับการวิเคราะห์

ความสำคัญ (Importance)

- ✓ ข้อมูลที่ถูก Extracted ออกมา มักจะยัง **ไม่เหมาะสมสำหรับการวิเคราะห์ทันที**
- ✓ ต้องมีการจัดโครงสร้าง (Structuring) เพื่อให้ข้อมูลสามารถนำไปใช้วิเคราะห์ได้อย่างมีประสิทธิภาพ

เทคนิคที่ใช้ในการ Structuring (Techniques)

1. **Aggregating Data:** รวมข้อมูลในระดับรายละเอียดที่แตกต่างกัน (Summarization)
2. **Joining Data:** เชื่อมโยงข้อมูลจากหลายแหล่ง/หลายตารางเข้าด้วยกัน (Integration)
3. **Pivoting Data:** หมุนตารางข้อมูลเพื่อเปลี่ยนมุมมองหรือนำเสนอข้อมูลในรูปแบบที่เหมาะกับการวิเคราะห์

Key Insight:

การจัดโครงสร้างข้อมูล (Data Structuring) เป็นขั้นตอนสำคัญระหว่าง **การ Extract ข้อมูล** และ **การวิเคราะห์ข้อมูล** เพราะเป็นตัวเชื่อมที่ทำให้ข้อมูลพร้อมใช้และเชื่อมโยงกันได้อย่างถูกต้อง

1. AGGREGATE DATA

ความหมายของ Aggregate Data

Aggregate Data (ข้อมูลที่สรุปรวม) = การนำเสนอข้อมูลในรูปแบบสรุป (Summarized Form) มี **รายละเอียดน้อยกว่า (Fewer Details)** เมื่อเทียบกับ **Disaggregated Data (ข้อมูลเชิงละเอียด/ดิบ)**

ตัวอย่าง: การวิเคราะห์ยอดซื้อจาก Vendor ⟶ สามารถสรุปรวมเป็น **Total Units Purchased** และ **Total Expenditures** ต่อ Vendor แต่หากสรุปรวมแล้ว จะไม่สามารถรู้ได้ว่า **ซื้อสินค้าใดบ้าง** ต้องย้อนกลับไปยัง **Disaggregated Data** เท่านั้น

Aggegated at the Vendor Level			*Aggregated for All Vendors*	
VendorName	**UnitsPurch**	**TotalCosts**	**UnitsPurch**	**TotalCosts**
B&D	543	$ 18,511.00	3,472	$ 152,111.54
Black and Decker	702	$ 15,812.63		
Calphalon	261	$ 19,509.75		
Honeywell	400	$ 5,516.90		
KitchenAid	381	$ 43,282.53		
Oster	697	$ 28,020.11		
Panasonic	488	$ 21,458.62		

Examples of Different Levels of Aggregating Data

1. AGGREGATE DATA

ตารางเปรียบเทียบ Aggregate Data vs Disaggregate Data

ประเภทข้อมูล (Type)	คำจำกัดความ (Definition)	ตัวอย่าง (Example)	การใช้งาน (Use)	ข้อจำกัด (Limitations)	แนวปฏิบัติที่ดี (Best Practice)
Disaggregated Data (ข้อมูลเชิงละเอียด)	ข้อมูลที่บันทึกในระดับ **เหตุการณ์เดี่ยว/ธุรกรรม** โดยไม่มีการสรุปรวม	รายการขายแต่ละครั้ง, ใบสั่งซื้อ, รายการสินค้า	ใช้สำหรับ **การวิเคราะห์เชิงลึก**, Drill-down, Audit, Traceability	ปริมาณข้อมูลมาก ⟶ การประมวลผลอาจช้า	เก็บข้อมูลในรูปแบบนี้เป็นหลัก
Aggregated Data (ข้อมูลสรุปรวม)	ข้อมูลที่ถูก **สรุปรวม/รวมยอด** ตามมิติที่สนใจ	ยอดขายรวมต่อ Sales Manager, ยอดขายรวมทั้งหมดในงบกำไรขาดทุน	ใช้สำหรับ **รายงานสรุป, การตัดสินใจเชิงกลยุทธ์, Performance Evaluation**	รายละเอียดหายไป ไม่สามารถย้อนวิเคราะห์ถึงธุรกรรมย่อยได้	สร้าง Aggregation ผ่าน **Query/Report** ตามความต้องการ แทนที่จะเก็บแบบสรุปรวมถาวร

ความสำคัญในกระบวนการ Transforming Data

- ✓ ต้องเข้าใจ **ระดับการสรุปรวม (Level of Aggregation)** ของแต่ละแหล่งข้อมูล
- ✓ การรวมข้อมูลที่สรุปในระดับต่างกัน อาจก่อให้เกิด **ปัญหาในการวิเคราะห์**

Key Takeaway:

- ✓ **Disaggregated = เก็บ** (เก็บข้อมูลดิบเพื่อความยืดหยุ่น)
- ✓ **Aggregated = ใช้** (ใช้สร้างรายงาน/การวิเคราะห์ แต่ไม่ใช่รูปแบบเก็บถาวร)

2. DATA JOINING

การ Join ข้อมูลในฐานข้อมูล (Database Joins in ETL Process)

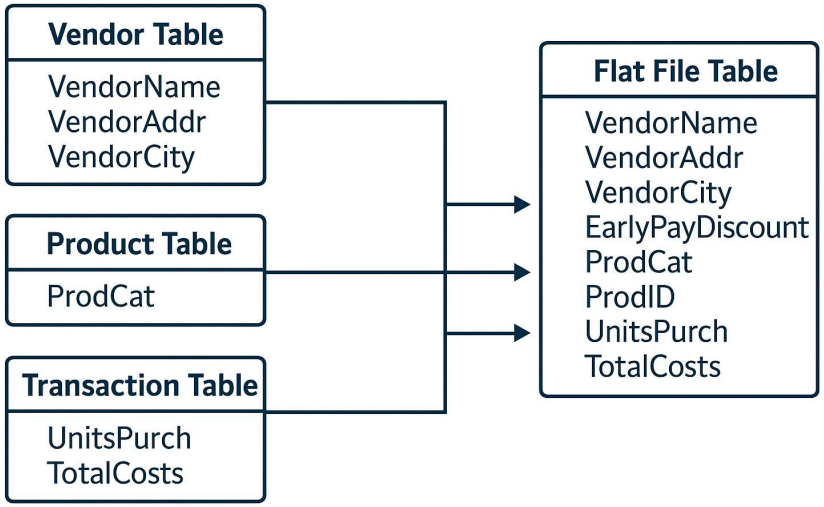
1. ความสำคัญของ Join

- ✓ การ Query ฐานข้อมูลมักต้อง **รวมข้อมูลจากหลายตาราง (Join)** เพื่อสร้างตารางเดียวที่พร้อมสำหรับการวิเคราะห์
- ✓ เป็นขั้นตอนสำคัญใน **ETL Process** โดยเฉพาะเมื่อข้อมูลสุดท้ายถูก Export ออกมาเป็น **Flat File**

2. ตัวอย่าง (S&S Dataset Example)

ตารางที่ได้จากการ Join มีข้อมูลมาจากหลายแหล่ง เช่น:

- ✓ **Vendor Table** ⟶ ฟิลด์ที่เกี่ยวกับผู้ขาย เช่น *Vendor Name, Discount*
- ✓ **Product Table** ⟶ ฟิลด์ที่เกี่ยวกับสินค้า เช่น *ProdCat, ProdID*
- ✓ **Transaction Table** ⟶ ฟิลด์ที่เกี่ยวกับธุรกรรมการซื้อ เช่น *UnitsPurch, TotalCosts*

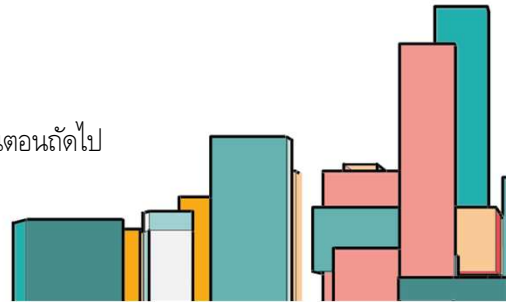


3. ผลลัพธ์

- ✓ หลังจาก Join ⟶ ได้ **Flat File Table** ที่รวมฟิลด์จากทั้ง **Vendor + Product + Transaction** เข้าด้วยกัน ตารางเดียวนี้ทำให้นักวิเคราะห์สามารถนำไปใช้ในการ **วิเคราะห์เชิงลึก (Detailed Analysis)** ได้ทันที

📌 **Key Insight:**

การ Join คือสะพานที่เชื่อม **ข้อมูลเชิงรายละเอียดจากหลายตาราง** มารวมเป็น **ตารางเดียว (Single Flat File)** เพื่อรองรับการวิเคราะห์และรายงานในขั้นตอนถัดไป



3. DATA PIVOTING

ความหมายของ Data Pivoting

Data Pivoting = การ “หมุน” ข้อมูลจาก **แถว (Rows)** ให้กลายเป็น **คอลัมน์ (Columns)**

ซอฟต์แวร์บางตัวออกแบบมาให้ทำงานกับ **Pivoted Data** มากกว่า Unpivoted Data การเข้าใจโครงสร้างของโปรแกรมปลายทางที่ใช้วิเคราะห์ข้อมูล จึงเป็นสิ่งสำคัญในการตัดสินใจว่าจะ Pivot ข้อมูลหรือไม่

ความเชื่อมโยงกับ Aggregation

- ✓ การ Pivot มักจะ **เกี่ยวข้องกับ Aggregation** (เช่น การ Sum ข้อมูลตาม Vendor และ Product Category)
- ✓ ข้อจำกัด: เมื่อ Aggregation เกิดขึ้น ⟶ รายละเอียดระดับเดิม (เช่น รายการสินค้าเดี่ยว) จะ **สูญหายและไม่สามารถย้อนกลับมาได้** เว้นแต่จะกลับไปใช้ **Disaggregated Data ดั้งเดิม**

แนวปฏิบัติที่ดี (Best Practice)

- ✓ เก็บข้อมูลในรูปแบบ **ละเอียดที่สุด (Disaggregated)**
- ✓ ใช้ Pivot & Aggregation **เฉพาะในขั้นตอนการวิเคราะห์/รายงาน**
- ✓ หลีกเลี่ยงการเก็บถาวรในรูปแบบ Pivoted เพราะจะสูญเสียความยืดหยุ่นในการวิเคราะห์

Figure 1

	ProdCat		
VendorName	**1**	**2**	**3**
B&D	$ 15,982.00	$ 2,529.00	
Black and Decker	$ 2,220.06	$ 568.00	$ 13,024.57
Calphalon	$ 19,509.75		
Honeywell		$ 5,516.90	
KitchenAid	$ 43,282.53		
Oster	$ 28,020.11		
Panasonic	$ 15,765.12	$ 5,693.50	

Figure 2

VendorName	ProdCat	TotalCosts
B&D	1	$15,982.00
B&D	2	$ 2,529.00
Black and Decker	1	$ 2,220.06
Black and Decker	2	$ 568.00
Black and Decker	3	$13,024.57
Calphalon	1	$19,509.75
Honeywell	2	$ 5,516.90
KitchenAid	1	$43,282.53
Oster	1	$28,020.11
Panasonic	1	$15,765.12
Panasonic	2	$ 5,693.50

ตัวอย่าง Pivot

(Pivot ครั้งแรก (Figure 1):

แต่ละแถว = **Vendor Name**

แต่ละคอลัมน์ = **Product Categories**

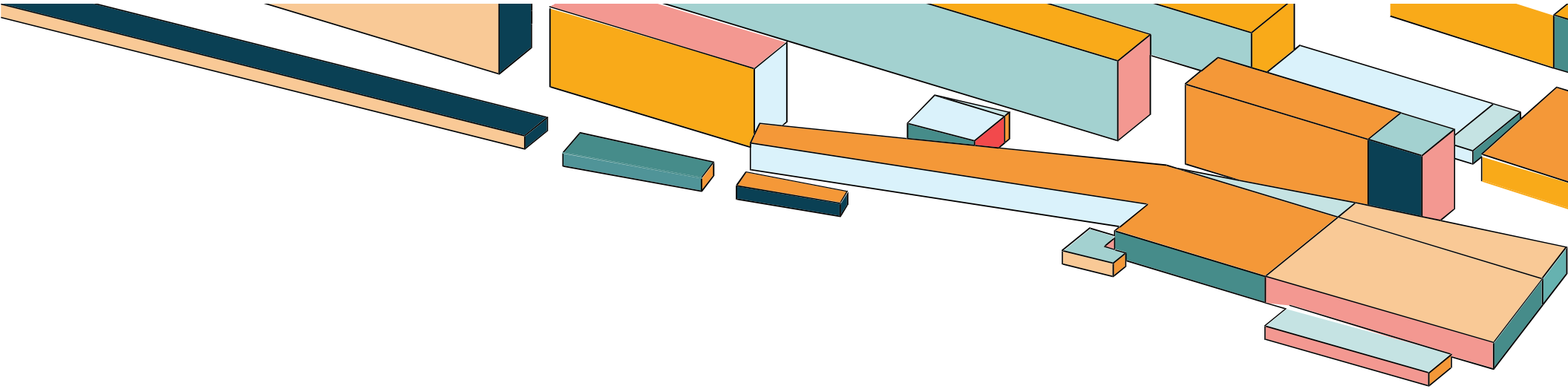
ค่าภายในตาราง = **Total Product Costs (Sum)**

Pivot กลับ (Figure 2):

ข้อมูลใกล้เคียงกับต้นฉบับมากขึ้น

แต่ยังคง **Aggregated** ในระดับสูงกว่า

ข้อมูลรายละเอียดของสินค้า (Individual Products) สูญหายไปตั้งแต่ Pivot แรก ⟶ ไม่สามารถกู้คืนได้



DATA STANDARIZATION

DATA STANDARDIZATION

ความหมายของ Data Standardization

Data Standardization = กระบวนการทำให้ **โครงสร้าง (Structure)** และ **ความหมาย (Meaning)** ของแต่ละข้อมูล (Data Element) เป็นมาตรฐานเดียวกัน เพื่อให้ข้อมูลสามารถ **นำไปวิเคราะห์และใช้ตัดสินใจ** ได้อย่างถูกต้อง

ความสำคัญ

- ✓ มีความสำคัญโดยเฉพาะเมื่อ **รวมข้อมูลจากหลายแหล่ง (Merging Data Sources)**
- ✓ ทำให้ข้อมูลจากระบบที่ต่างกันสามารถสื่อสารและเปรียบเทียบกันได้

วิธีการทำให้ข้อมูลเป็นมาตรฐาน (Achieving Standardized Data)

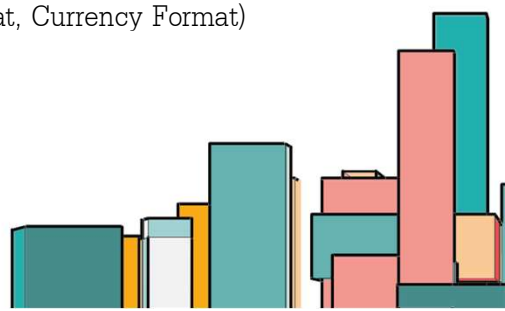
- ✓ แปลงข้อมูลให้เป็น **Format หรือ Data Type เดียวกัน**
- ✓ ใช้ **Coding Scheme** เดียวกัน (เช่น Country Code, Currency Code)
- ✓ ทำให้ข้อมูลอยู่ใน **Field ที่ถูกต้อง** และจัดเรียง Field ให้เป็นระบบ

แนวคิดเชิงปฏิบัติ (Practical Perspective)

- ✓ เมื่อทำงานกับ **Database File** หรือ **Flat File Format** ⟶ คิดว่าเป็นการทำให้ **Columns** ของข้อมูลถูกต้องและเป็นมาตรฐาน

ประเด็นที่ควรพิจารณา (Considerations in Data Standardization)

1. **Data Parsing** ⟶ การแยกข้อมูลออกจากกัน (เช่น ชื่อเต็ม ⟶ First Name, Last Name)
2. **Data Concatenation** ⟶ การรวมข้อมูลเข้าด้วยกัน (เช่น Address Line 1 + Address Line 2)
3. **Cryptic Data Values** ⟶ การใช้ค่าที่เข้าใจยาก เช่น Code ที่ไม่ชัดเจน ต้องแปลงให้เป็นความหมายที่เข้าใจได้
4. **Misfielded Data Values** ⟶ ข้อมูลอยู่ใน Field ที่ผิด (เช่น หมายเลขโทรศัพท์ถูกใส่ในช่อง Address)
5. **Data Formatting & Consistency** ⟶ การทำให้รูปแบบข้อมูลสอดคล้องกัน (เช่น Date Format, Currency Format)



1. DATA PARSING

ความหมาย (Definition)

- **Data Parsing** = การแยกข้อมูลจาก **ฟิลด์เดียว** ออกมาเป็น **หลายฟิลด์ย่อย**
- จุดประสงค์: เพื่อให้สามารถนำ **องค์ประกอบย่อย ๆ ของข้อมูล** ไปใช้วิเคราะห์ได้แยกกันอย่างมีประสิทธิภาพ

การใช้งาน (Uses)

- **Data Parsing = เปลี่ยนข้อมูลที่ซับซ้อนให้ง่ายต่อการวิเคราะห์**
- เมื่อข้อมูลถูกแยกเป็นองค์ประกอบย่อย ⟶ ทำให้สามารถใช้ข้อมูลนั้นได้ **ตรงตามวัตถุประสงค์เชิงวิเคราะห์** มากขึ้น

Data Parsing Example

	Original Column		*New Columns*		
RowNumber	**EarlyPayDiscount**		**DiscountRate**	**DiscountDays**	**BalanceDueDays**
1	2/10 N30		2	10	30
2	2/10 N30		2	10	30
3					
4	2/10 N30		2	10	30
5	2/10 N30		2	10	30
6	2/10 N15		2	10	15
7	2/10 N30		2	10	30
8	2/10 N15		2	10	15
9					
10	2/10 N30		2	10	30
11	2/10 N30		2	10	30
12	1/10 N30		1	10	30
13					
14	2/10 N30		2	10	30
15	1/10 N30		1	10	30
16	2/10 N30		2	10	30
17	2/10 N15		2	10	15

2. DATA CONCATENATION

ความหมาย (Definition)

Data Concatenation = การนำข้อมูลจาก สองฟิลด์หรือมากกว่า มารวมกันเป็น ฟิลด์เดียว

- **มักใช้ในการ รวมข้อมูลเพื่อให้อ่านง่ายขึ้น หรือ สร้างตัวระบุเฉพาะ (Unique Identifier)Key Insight**

การใช้งาน (Uses)

- ทำให้ **ข้อมูลแสดงผลได้สะดวกขึ้น** (เช่น Full Name)
- ใช้สร้าง **Unique Identifier** สำหรับตารางที่ Primary Key ไม่สมบูรณ์
- ใช้ในการรวม Field หลายตัวเพื่อสร้าง **Key สำหรับ Join** ระหว่างตาราง

Key Takeaway:

- **Parsing = แยก Field ⟶ หลาย Field**
- **Concatenation = รวม Field ⟶ Field เดียว**
- ทั้งสองเทคนิคมักถูกใช้ร่วมกันใน **Data Transformation** เพื่อทำให้ข้อมูล **สะอาด (Clean), มีโครงสร้าง (Structured), และพร้อมใช้งานเชิงวิเคราะห์**

Data Concatenation Example

	Original Columns		*New Column*
RowNumber	**FName**	**LName**	**FullName**
1	DeShawn	Williams	DeShawn Williams
2	Milton	Armstrong	Milton Armstrong
3	Chiyo	Tanaka	Chiyo Tanaka
4	M.	Armstrong	M. Armstrong
5	Milton	Armstrong	Milton Armstrong
6	Maryam	Ahmad	Maryam Ahmad
7	Milton G.	Armstrong	Milton G. Armstrong
8	Maryam	Ahmad	Maryam Ahmad
9	Chiyo	Tanaka	Chiyo Tanaka
10	DeShawn	Williams	DeShawn Williams
11	DeShawn	Williams	DeShawn Williams
12	Larsena	Hansen	Larsena Hansen
13	Jacobsen	Sofia	Jacobsen Sofia
14	Milton	Armstrong	Milton Armstrong
15	Larsena	Hansen	Larsena Hansen
16	Milton	Armstrong	Milton Armstrong
17	Maryam	Ahmad	Maryam Ahmad

3. CRYPTIC DATA VALUES

ความหมาย (Definition)

- **Cryptic Data Values** = ค่าข้อมูลที่ **ไม่มีความหมายชัดเจน** หากไม่เข้าใจ **Coding Scheme** ที่ใช้กำหนดค่า
- ผู้ใช้จะต้องอ้างอิง **Data Dictionary** หรือ **ตาราง Mapping** จึงจะเข้าใจค่าที่บันทึกไว้

ตาราง: Cryptic Data Values & Dummy Variables

ประเภท (Type)	ตัวอย่าง (Example)	ปัญหา (Problem)	แนวทางแก้ไข (Solution)	Best Practice
Cryptic Data Values	- Employee Position: 1 = Partner, 2 = Senior Consultant, 3 = Analyst - ProdCat: 1 = Kitchen Appliances, 2 = Fans, 3 = Hand Tools	ผู้ใช้ไม่เข้าใจค่าตัวเลข/รหัส หากไม่อ้างอิง Data Dictionary	- Replace ด้วยคำเต็ม - หรือ Add Column ใหม่เพื่อเก็บคำอธิบาย	ใช้ **Join Data** เพื่อเก็บ ID เดิม และเพิ่มความหมายประกอบ
Dummy Variables (Dichotomous)	- Preferred Vendor: 1 = Yes, 0 = No	ค่าตัวเลข (0/1) เข้าใจได้ยาก ถ้า Field Name ไม่สื่อความหมาย	ใช้ 0 และ 1 ตามมาตรฐาน (0 = Absence, 1 = Presence)	ตั้งชื่อ Field ให้ชัดเจน เช่น **PreferredVendor** แทนที่จะใช้ VendorStatus

Key Takeaway:

✓ ปัญหา Cryptic Data Values = ทำให้ผู้ใช้ **ไม่สามารถตีความข้อมูลได้ทันที**

✓ แนวทางแก้ที่ดีที่สุดคือ **ใช้ Data Dictionary + Join Data เพื่อเติมความหมาย** และตั้งชื่อฟิลด์ให้เข้าใจง่าย

4. MISFIELDED DATA VALUES

ความหมาย (Definition)

- ✓ **Misfielded Data Values** = ค่าข้อมูลที่มี **รูปแบบถูกต้อง (Correctly Formatted)** แต่ถูกบันทึกใน **Field ที่ไม่ถูกต้อง**
- ✓ ปัญหาคือ **ข้อมูลอยู่ผิดที่** ไม่ใช่ว่าข้อมูลมีรูปแบบผิด

ลักษณะปัญหา (Issue)	ตัวอย่าง (Example)	การแก้ไข (Correction)	ระดับปัญหา (Level)
ค่าข้อมูลอยู่ผิด Field	Field: City = “Germany” (ควรอยู่ใน Country)	ย้ายค่า “Germany” ไป Field Country	Individual Row
ทั้ง Column อยู่ผิด Field	Field: City เก็บ State Field: State เก็บ City	สลับข้อมูลให้ตรง Field ที่ถูกต้อง	Entire Column
ค่าข้อมูลถูกต้อง แต่ Field ผิด	ตัวเลขโทรศัพท์ถูกบันทึกใน Field Address	ย้ายค่ากลับไป Field Phone	Individual Row
รูปแบบข้อมูลไม่ผิด แต่ Field ผิด	Field: PostalCode = “Bangkok”	ย้าย “Bangkok” ไป Field City	Individual Row

Key Takeaway:

- ✓ **Misfielded Data Values = Data ถูก แต่ที่ผิด**
- ✓ อาจเกิดได้ทั้ง **ระดับแถว (Row)** และ **ระดับคอลัมน์ (Column)**
- ✓ ต้องอาศัย **Data Dictionary / Schema** เพื่อค้นพบและแก้ไขให้ถูกต้อง

5. DATA FORMATTING AND DATA CONSISTENCY

ความหมาย (Definition)

- **Data Formatting** = รูปแบบการแสดงข้อมูล เช่น วันที่แสดงเป็น “03/04/82” หรือ “April 3, 1982”
- **Data Consistency** = ความสม่ำเสมอในการเก็บและใช้ข้อมูล ⟶ ทุกค่าต้องถูกเก็บในรูปแบบเดียวกัน

ตัวอย่าง ข้อมูลที่มี format ต่างกัน

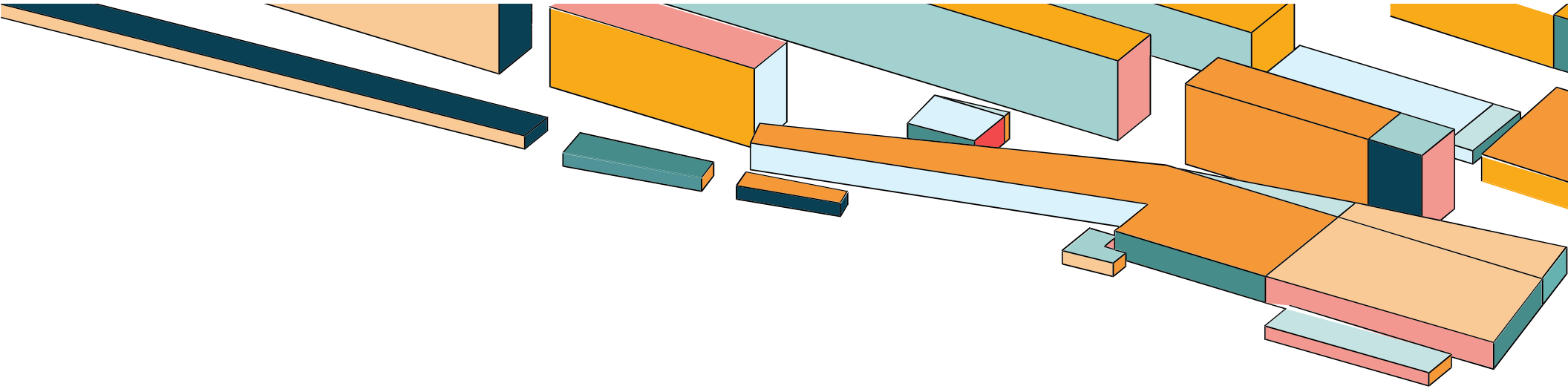
PhoneNumber
907961-4917
(844) 986-0858
3605659487
(551) 7779393
(790)447-1783
(551) 7779393

ตารางเปรียบเทียบ Data Formatting vs Data Consistency

หัวข้อ	Data Formatting	Data Consistency
ความหมาย (Definition)	วิธีการ **แสดงผลข้อมูล** (Display Format)	วิธีการ **เก็บข้อมูลจริง** (Storage Format)
ตัวอย่าง (Example)	วันที่เดียวกันแสดงเป็น - April 3, 1982 - 03/04/82 - 04/03/82	วันที่เก็บเป็น - Excel Serial Date = 29170.0416 - Unix Epoch Time = 311130000
ปัญหา (Problem)	- ทำให้ **ผู้ใช้สับสน** - อาจทำให้ข้อมูล **ผิดเพี้ยนตอน Import/Export**	- ข้อมูล **ไม่สอดคล้องกันระหว่างระบบ** - การวิเคราะห์ผิดเพราะเก็บ **หลายรูปแบบผสมกัน**
กรณีศึกษา (Case Study)	MI5: Spreadsheet Error ⟶ เลขโทรศัพท์ 134 หมายเลขท้ายถูกเปลี่ยนเป็น “000”	ระบบเก็บวันที่บางส่วนเป็น Serial Date และบางส่วนเป็น Epoch Time ใน Field เดียวกัน
Best Practice	- เลือก **Format เดียว** ใช้ทั้ง Field/File - กำหนด Format ไว้ใน **Data Dictionary**	- เก็บข้อมูลด้วย **Format เดียวกัน** (เช่น UTC สำหรับเวลา) - หลีกเลี่ยงการสลับไปมาระหว่างวิธีเก็บหลายแบบ

Key Takeaway: ทั้งสองอย่างต้องควบคู่กัน ⟶ เพื่อความถูกต้องและน่าเชื่อถือของข้อมูล

- ✓ **Formatting = ความสม่ำเสมอในการ “แสดงผล”**
- ✓ **Consistency = ความสม่ำเสมอในการ “เก็บข้อมูล”**



DATA CLEANING

DIRTY DATA AND DATA CLEANING

ปัญหาหลัก

- ข้อมูลสกปรก (Dirty Data) = ข้อมูลที่ **ไม่สอดคล้องกัน / ไม่ถูกต้อง / ไม่สมบูรณ์**
- พบมากที่สุดจากการสำรวจผู้ใช้งาน Kaggle 16,000 คน

ความเสี่ยงและต้นทุน

- ทำให้การตัดสินใจผิดพลาด
- สูญเสียชื่อเสียงและความน่าเชื่อถือ
- ตัวอย่าง: Fidelity Magellan Fund
    - รายงานปันผล $4.32 ต่อหุ้น ⟶ **ผิดพลาด**
    - นักบัญชีภาษีลืมใส่เครื่องหมายลบ (ขาดทุน 1.3 พันล้านดอลลาร์) ระบบตีความเป็นกำไร

      ⟶ **เกิดความคลาดเคลื่อน 2.6 พันล้านดอลลาร์**

ข้อผิดพลาดที่พบบ่อย

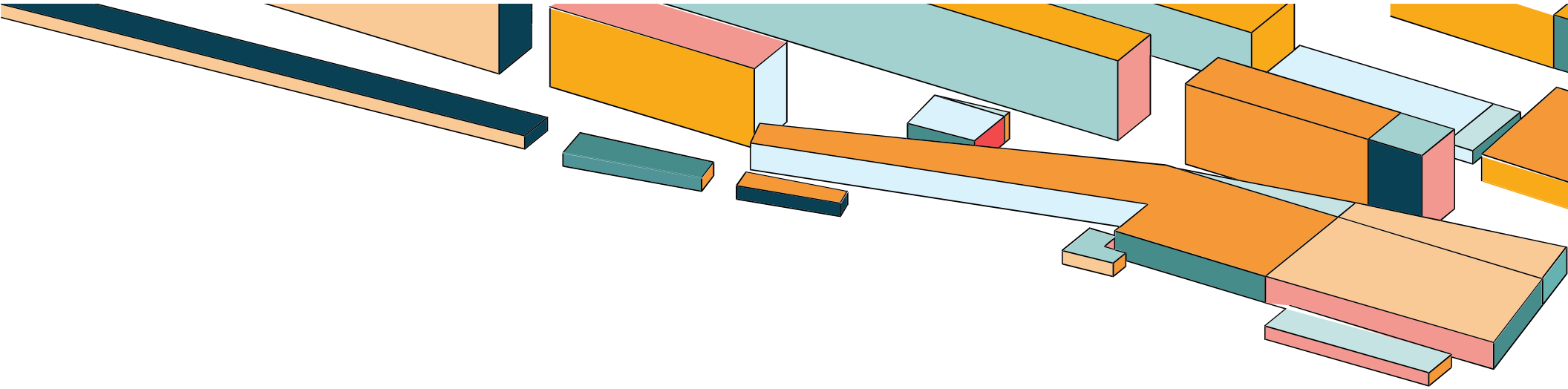
- ✓ รูปแบบข้อมูลไม่ตรงกัน (วันที่, หน่วย, ตัวพิมพ์)
- ✓ ข้อมูลหาย/ไม่สมบูรณ์
- ✓ ข้อมูลซ้ำซ้อน
- ✓ ค่าผิดปกติ (Negative Age, Impossible Dates)
- ✓ รหัสหมวดหมู่ไม่ถูกต้อง (NY ≠ N.Y. ≠ New York)
- ✓ ความผิดพลาดจากการป้อนข้อมูล (decimal, transposed digits, missing sign)

DATA QUALITY ISSUE

ประเภทปัญหา	คำนิยาม	ตัวอย่างจาก S&S	แนวทางแก้ไข
Data De-Duplication (การลบข้อมูลซ้ำ)	การลบระเบียนที่มีค่าข้อมูลเหมือนกันทุกช่อง	แถวที่ 17 เป็นข้อมูลซ้ำกับแถวที่ 6	ตรวจสอบว่าไม่ใช่คำสั่งซื้อจริงซ้ำ แล้วลบข้อมูลซ้ำออก
Data Filtering (การกรองข้อมูล)	การลบระเบียน/คอลัมน์ที่ไม่เกี่ยวข้อง หรือจัดการค่าที่หาย (null)	ไฟล์มี **หมวดสินค้า 3 (เครื่องมือช่าง)** ที่ไม่เกี่ยวข้องกับการวิเคราะห์	กรองเฉพาะหมวดที่เกี่ยวข้อง จัดการค่าที่หายด้วยการเก็บไว้, ลบ หรือทำ Data Imputation
Contradiction Errors (ข้อมูลขัดแย้ง)	ข้อมูลสิ่งเดียวกันแต่ไม่สอดคล้องกัน	หมายเลขโทรศัพท์ของ **Milton Armstrong** ไม่ตรงกัน (แถว 16 เทียบกับแถวอื่น)	ตรวจสอบจาก log ระบบ หรือยืนยันกับ vendor แล้วแก้ให้สอดคล้องกัน
Threshold Violations (เกินค่าที่กำหนด)	ค่าอยู่นอกช่วงที่กำหนดหรือไม่ตรงตามรูปแบบฟิลด์	ฟิลด์ PhoneNumber ควรมี 10 หลัก แต่แถวที่ 13 มี 13 หลัก	แก้ไขข้อมูลให้ตรงตาม threshold ที่กำหนด (10 หลัก)
Attribute Dependency Errors (ความสัมพันธ์ผิดพลาด)	ค่าของคุณลักษณะรองไม่ตรงกับคุณลักษณะหลัก	เมือง **Concord, NC** ใส่รหัสไปรษณีย์ของ **Concord, CA (94519)**	แก้ไขรหัสไปรษณีย์เป็น 28027 ให้ถูกต้องสำหรับ Concord, NC
Data Entry Errors (การป้อนข้อมูลผิด)	ความผิดพลาดจากการกรอกข้อมูล เช่น พิมพ์ผิด, ตัวเลขสลับ, ข้อมูลหาย	เมือง **San Diego** พิมพ์ผิดเป็น “SanDgieo” (แถว 6, 8, 17)	แก้ไขการสะกดให้ถูกต้องเป็น “San Diego”

Example : Duplicate

	Original Columns		*New Column*
RowNumber	**FName**	**LName**	**FullName**
1	DeShawn	Williams	DeShawn Williams
2	Milton	Armstrong	Milton Armstrong
3	Chiyo	Tanaka	Chiyo Tanaka
4	M.	Armstrong	M. Armstrong
5	Milton	Armstrong	Milton Armstrong
6	Maryam	Ahmad	Maryam Ahmad
7	Milton G.	Armstrong	Milton G. Armstrong
8	Maryam	Ahmad	Maryam Ahmad
9	Chiyo	Tanaka	Chiyo Tanaka
10	DeShawn	Williams	DeShawn Williams
11	DeShawn	Williams	DeShawn Williams
12	Larsena	Hansen	Larsena Hansen
13	Jacobsen	Sofia	Jacobsen Sofia
14	Milton	Armstrong	Milton Armstrong
15	Larsena	Hansen	Larsena Hansen
16	Milton	Armstrong	Milton Armstrong
17	Maryam	Ahmad	Maryam Ahmad



DATA VALIDATION

DATA VALIDATION

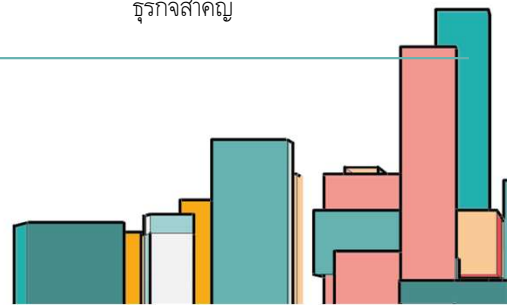
การตรวจสอบความถูกต้องของข้อมูล (Data Validation)

ความหมาย

- กระบวนการวิเคราะห์ว่าข้อมูลมีคุณสมบัติเป็น **ข้อมูลที่มีคุณภาพสูง** หรือไม่
- ใช้ทั้ง **แบบเป็นทางการ** (หลังการแปลงข้อมูล) และ **แบบไม่เป็นทางการ** (ตรวจสอบระหว่างการแปลงข้อมูล)
- เป็นขั้นตอนสำคัญก่อนการทำ Data Cleaning และมักทำแบบวนซ้ำ (Iterative Process) เพื่อทั้งค้นหาปัญหาที่ต้องแก้ไข และยืนยันว่าการแปลงข้อมูลสำเร็จและถูกต้อง

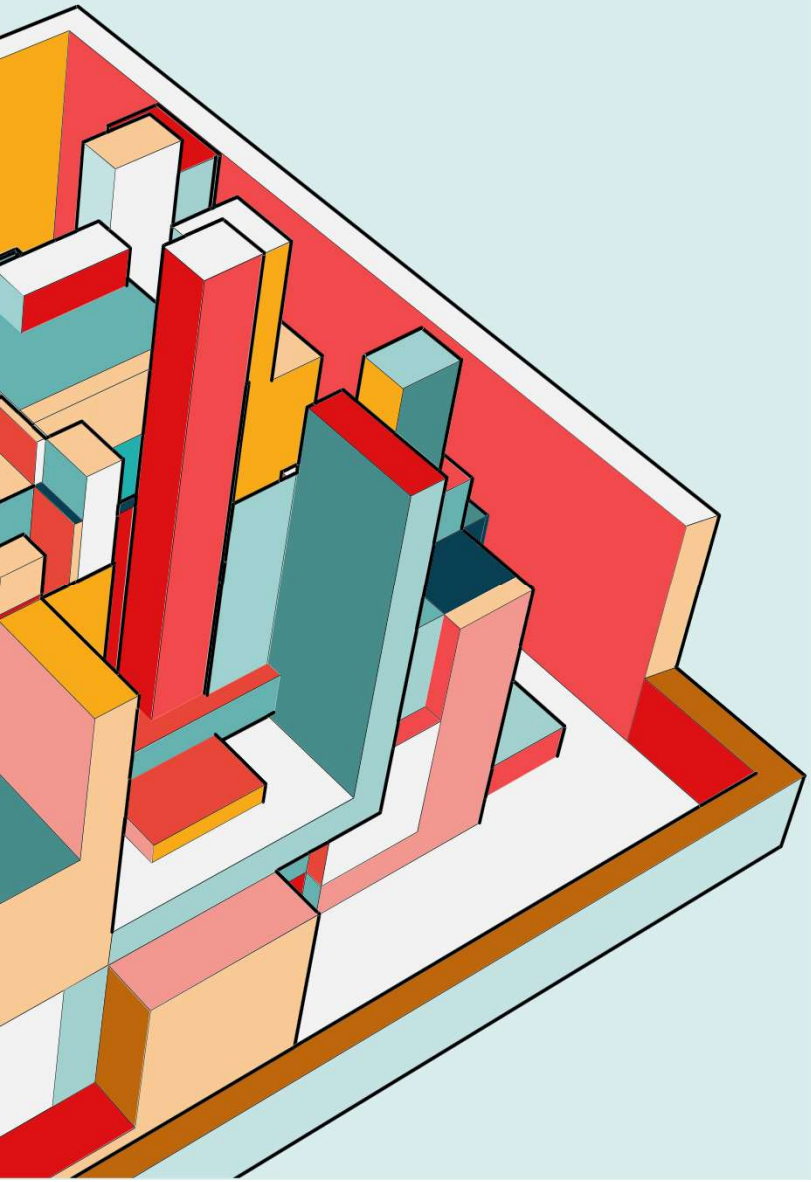
ตาราง **สรุปการตรวจสอบความถูกต้องของข้อมูล (Data Validation Techniques)**

Technique	วิธีการ	ตัวอย่าง	จุดแข็ง
Visual Inspection (การตรวจสอบด้วยสายตา)	ใช้สายตามองหาความผิดปกติ เช่น การเรียงข้อมูล, การหาค่าซ้ำ (unique values)	ตรวจพบ “Milton Armstrong” มีหลายรูปแบบการสะกด หรือเบอร์โทรศัพท์ผิดรูปแบบ	เร็ว ง่าย ใช้ได้ทั้งข้อมูลเล็กและใหญ่
Basic Statistical Tests (การทดสอบทางสถิติพื้นฐาน)	คำนวณสถิติเบื้องต้น (min, max, mean, median, sum) และเปรียบเทียบก่อน–หลังการแปลง	ราคาสินค้าติดลบ, ยอดรวมหลังแปลงไม่ตรงกับต้นทาง	ช่วยตรวจหาค่าผิดปกติและการสูญหายของข้อมูล
Audit a Sample (ตรวจสอบตัวอย่างข้อมูล)	เลือกสุ่มระเบียนมาวิเคราะห์และตรวจสอบกับข้อมูลต้นทาง	Dataset 100,000 ระเบียน ⟶ ตรวจ 1,000 ระเบียน ⟶ พบ error rate 7%	ใช้ประเมินคุณภาพโดยรวมและคำนวณ Error Rate ได้
Advanced Testing (การทดสอบขั้นสูง)	ใช้ Business Rules หรือความรู้เฉพาะด้านเพื่อตรวจสอบ	ตรวจสอบว่า Debit = Credit, Sub-ledger = GL, จำนวน × ราคา = ยอดรวม	ให้ความเชื่อมั่นสูง เหมาะกับข้อมูลทางธุรกิจสำคัญ



EXAMPLE OF TRANSFORM DATA

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	RowNumber	VendorName	FName	LName	PhoneNumber	Address	City	State	ZipCode	DiscountRate	DiscountDays	BalanceDueDays	ProdCat	CatDesc	ProdID	ProdDesc	UnitsPurch	TotalCosts	AvgCostPerUnit
2	1	Oster	DeShawn	Williams	(907) 961-4917	1535 N Hickory St	Owosso	MI	48867	2	10	30	1	Kitchen Appliances	3	Toaster	186	$ 2,068.32	$ 11.12
3	2	Black and Decker	Milton	Armstrong	(844) 986-0858	1000 Stanley Dr	Concord	NC	28027	2	10	30	1	Kitchen Appliances	3	Toaster	227	$ 2,220.06	$ 9.78
4	3	KitchenAid	Chiyo	Tanaka	(360) 565-9487	1701 Kitchen Aid Way	Greenville	OH	45331				1	Kitchen Appliances	2	Microwave	210	$40,849.20	$ 194.52
5	4	Black and Decker	Milton	Armstrong	(844) 986-0858	1000 Stanley Dr	Concord	NC	28027	2	10	30	3	Hand tools	1	Hand Drill	273	$10,851.75	$ 39.75
6	5	Black and Decker	Milton	Armstrong	(844) 986-0858	1000 Stanley Dr	Concord	NC	28027	2	10	30	2	Fans	2	Desk Fan	80	$ 568.00	$ 7.10
7	6	Panasonic	Maryam	Ahmad	(551) 777-9393	7625 Panasonic Way	San Diego	CA	92154	2	10	15	1	Kitchen Appliances	2	Microwave	51	$ 7,882.56	$ 154.56
8	7	Black and Decker	Milton	Armstrong	(844) 986-0858	1000 Stanley Dr	Concord	NC	28027	2	10	30	2	Fans	1	Box Fan	281	$ 2,529.00	$ 9.00
9	8	Panasonic	Maryam	Ahmad	(551) 777-9393	7625 Panasonic Way	San Diego	CA	92154	2	10	15	2	Fans	2	Desk Fan	386	$ 5,693.50	$ 14.75
10	9	KitchenAid	Chiyo	Tanaka	(360) 565-9487	1701 Kitchen Aid Way	Greenville	OH	45331				1	Kitchen Appliances	3	Toaster	171	$ 2,433.33	$ 14.23
11	10	Oster	DeShawn	Williams	(907) 961-4917	1535 N Hickory St	Owosso	MI	48867	2	10	30	1	Kitchen Appliances	1	Blender	263	$ 5,062.75	$ 19.25
12	11	Oster	DeShawn	Williams	(907) 961-4917	1535 N Hickory St	Owosso	MI	48867	2	10	30	1	Kitchen Appliances	2	Microwave	248	$20,889.04	$ 84.23
13	12	Honeywell	Larsena	Hansen	(790) 447-1783	10640 Freeport Dr	Louisville	KY	40258	1	10	30	2	Fans	2	Desk Fan	54	$ 621.00	$ 11.50
14	13	Calphalon	Sofia	Bruner	(933) 937-5654	20750 Midstar Dr	Bowling Green	OH	43402				1	Kitchen Appliances	1	Blender	261	$19,509.75	$ 74.75
15	14	Black and Decker	Milton	Armstrong	(844) 986-0858	1000 Stanley Dr	Concord	NC	28027	2	10	30	1	Kitchen Appliances	2	Microwave	262	$15,982.00	$ 61.00
16	15	Honeywell	Larsena	Hansen	(790) 447-1783	10640 Freeport Dr	Louisville	KY	40258	1	10	30	2	Fans	1	Box Fan	346	$ 4,895.90	$ 14.15
17	16	Black and Decker	Milton	Armstrong	(844) 986-0858	1000 Stanley Dr	Concord	NC	28027	2	10	30	3	Hand tools	2	Jigsaw	122	$ 2,172.82	$ 17.81



DATA ANALYSIS AND PRESENTATION

TEST CASE

Case Study: TD Bank Group – Building an Analytics Mindset

Case

- TD Bank Group (ธนาคารชั้นนำของแคนาดาและอเมริกาเหนือ)
- ใช้เวลา 5 ปีสร้าง **Data Lake**
- ลงทุนใน **เครื่องมือและการฝึกอบรม** ให้พนักงานวิเคราะห์ข้อมูลเองได้

Actions

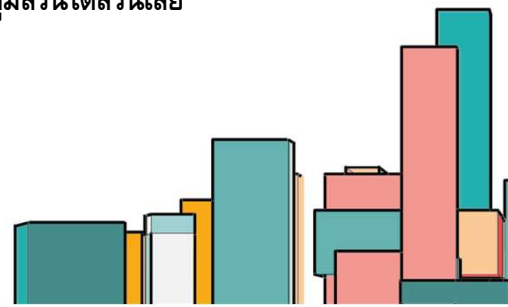
- สร้างระบบให้พนักงานเข้าถึงข้อมูลโดยตรง (ไม่ต้องพึ่ง Data Science Team)
- จัดหาเครื่องมือ Self-service Analytics & Visualization
- อบรมให้พนักงานทุกระดับสามารถทำ Data Analysis ได้

Results

- +90% ประสิทธิภาพโครงการวิเคราะห์
- -60% ต้นทุนการจัดการข้อมูล
- -30% ข้อร้องเรียนจากลูกค้าซ้ำซ้อน

Lesson Learned

- ✓ ETL อย่างเดียว **ไม่พอ** ⟶ ต้อง **วิเคราะห์** และ **สื่อสารผลลัพธ์**
- ✓ Analytics Mindset ต้องครอบคลุม:
  - ✓ การเลือกใช้ **เทคนิควิเคราะห์ข้อมูลที่เหมาะสม**
  - ✓ การ **ตีความและสื่อสารผลลัพธ์กับผู้มีส่วนได้ส่วนเสีย**

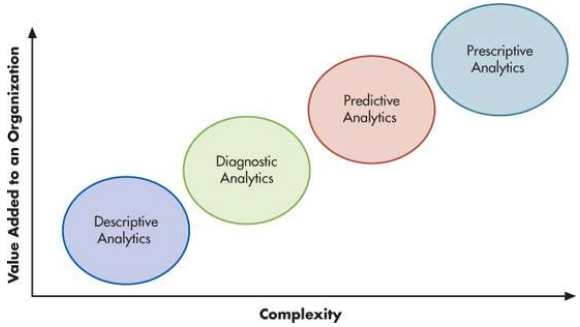


DATA ANALYSIS

ประเภทของการวิเคราะห์ข้อมูล (Categories of Data Analytics)

มีการแบ่งการวิเคราะห์ข้อมูลออกเป็น 4 ประเภท ได้แก่ **Descriptive, Diagnostic, Predictive และ Prescriptive** โดยแต่ละประเภทมีระดับความซับซ้อนและคุณค่าที่เพิ่มให้แก่องค์กรแตกต่างกัน การเป็นผู้เชี่ยวชาญในแต่ละด้านจำเป็นต้องอาศัยการฝึกฝนและเรียนรู้เพิ่มเติม

ประเภท	คำถามหลัก	วิธีการ/เทคนิค	ตัวอย่างการใช้งาน
Descriptive Analytics (การวิเคราะห์เชิงพรรณนา)	เกิดอะไรขึ้น? (*What happened?*)	- สถิติพื้นฐาน (mean, median, variance) - Visualization (charts, dashboards)	- อัตรากำไรขั้นต้น - Inventory turnover - Budget vs. Actual
Diagnostic Analytics (การวิเคราะห์เชิงวินิจฉัย)	ทำไมสิ่งนี้ถึงเกิดขึ้น? (*Why did it happen?*)	- Root cause analysis - "5 Why's" - Hypothesis testing	- วิเคราะห์สาเหตุที่กำไรขั้นต้นลดลง - ตรวจสอบสัดส่วนสินค้า/กลยุทธ์การตลาด
Predictive Analytics (การวิเคราะห์เชิงทำนาย)	สิ่งใดน่าจะเกิดขึ้น? (*What is likely to happen?*)	- Regression analysis - Machine learning models - Forecasting	- พยากรณ์ยอดขาย - การคาดการณ์ความเสี่ยงด้านเครดิต - Predictive maintenance
Prescriptive Analytics (การวิเคราะห์เชิงกำหนด)	ควรทำอย่างไรต่อไป? (*What should we do?*)	- Optimization models - Simulation - Scenario analysis	- แนะนำกลยุทธ์การตั้งราคา - กำหนดเส้นทางการขนส่งที่ดีที่สุด - การจัดสรรทรัพยากร



DATA ANALYSIS

ปัญหาที่พบบ่อยในการทำ Data Analytics

1.GIGO (Garbage In, Garbage Out)

- หากข้อมูลไม่คุณภาพ ⟶ ผลการวิเคราะห์ไม่มีค่า

2.Overfitting

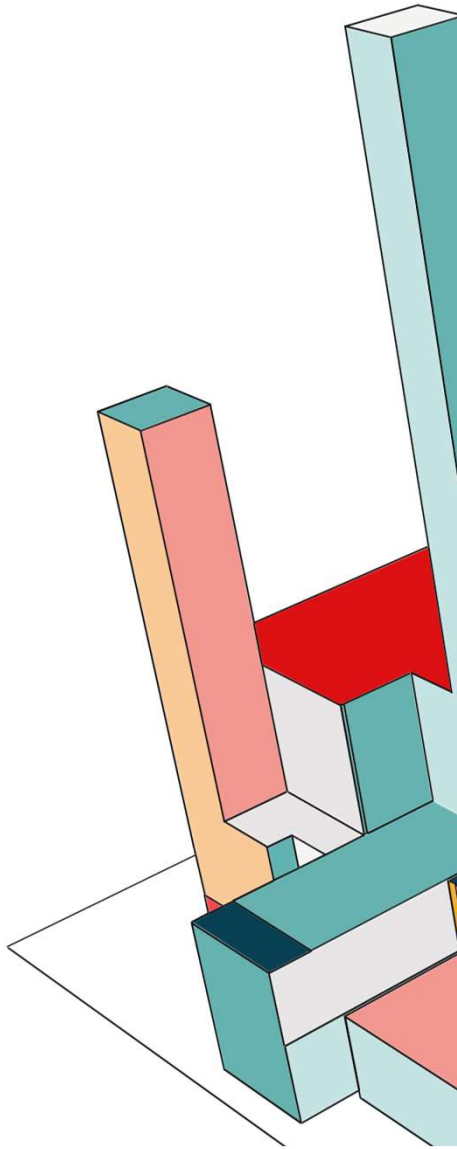
- โมเดลตรงกับข้อมูลเก่าเกินไป ⟶ ทำนายข้อมูลใหม่ไม่ได้

3.Extrapolation เกินช่วงข้อมูล

- ใช้โมเดลเกินขอบเขตข้อมูลเดิม ⟶ ผลลัพธ์ผิดเพี้ยน (เช่น โมเดลบ้าน 2,000–3,000 ตร.ฟุต ⟶ ไปใช้ทำนายบ้าน 7,000 ตร.ฟุต)

4.Variation (ความแปรปรวน)

- ไม่มีโมเดลใดทำนายได้แม่นยำ 100% ⟶ ควรรายงานเป็นช่วง (เช่น อุณหภูมิ 75–85 องศา ที่ความมั่นใจ 95%) ไม่ใช่ค่าจุดเดียว



1. DESCRIPTIVE ANALYTICS

1. Descriptive Analytics (การวิเคราะห์เชิงพรรณนา)

คำถามหลัก: “เกิดอะไรขึ้น?”

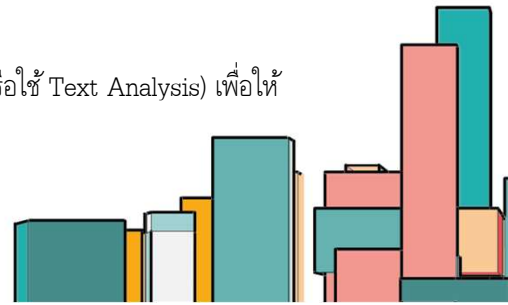
- ใช้เทคนิค Exploratory Data Analysis (EDA) เพื่อสำรวจข้อมูลโดยไม่อิงสมมติฐานหรือโมเดลที่ซับซ้อน
- ตัวอย่างการใช้งาน:
  - ✓ ผู้ตรวจสอบบัญชีภายนอกคำนวณ อัตรากำไร, leverage ratios เพื่อดูความเสี่ยงทางธุรกิจและระบุการทุจริต
  - ✓ นักบัญชีองค์กรคำนวณ ต้นทุนต่อหน่วย, inventory turnover, customer acquisition cost, variance budget-to-actual เพื่อติดตามประสิทธิภาพการดำเนินงาน

2. เทคนิคสำคัญใน Descriptive Analytics

- **Central Tendency (แนวโน้มเข้าสู่ศูนย์กลาง)**
  - ✓ ค่าเฉลี่ย (Mean) และค่ามัธยฐาน (Median)
  - ✓ ใช้เปรียบเทียบเพื่อดูว่ามี Outlier หรือไม่
- **Spread (การกระจาย)**
  - ✓ ช่วงของข้อมูล (Range)
  - ✓ ส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation)
  - ✓ ค่าควอไทล์ (Quartiles)
- **Distribution (การกระจายตัว)**
  - ✓ ตรวจสอบรูปแบบการกระจาย เช่น การกระจายปกติ (Normal Distribution)
  - ✓ ใช้เพื่อเลือกสถิติที่เหมาะสมในการวิเคราะห์
- **Correlation (ความสัมพันธ์)**
  - ✓ วัดด้วยค่าสหสัมพันธ์ (Correlation Coefficient) ระหว่าง −1 ถึง +1
  - ✓ แสดงการเคลื่อนไหวของตัวแปรที่สัมพันธ์กัน แต่ไม่เท่ากับเหตุและผล (Correlation $\neq$ Causation)

👉 **การแสดงผลด้วย Visualization (Viz) เช่น กราฟ แผนภาพ หรือแอนิเมชัน มักช่วยให้เข้าใจได้เร็วและชัดเจนขึ้น**

การวิเคราะห์ข้อมูลเชิงคุณภาพ ข้อมูลเช่น **โพสต์ในโซเชียลมีเดีย** สามารถแปลงเป็นข้อมูลเชิงปริมาณได้ (เช่น นับจำนวนข้อความเชิงบวก/เชิงลบ หรือใช้ Text Analysis) เพื่อให้สามารถวิเคราะห์ต่อได้เหมือนข้อมูลเชิงปริมาณ



2. DIAGNOSTIC ANALYTICS

2. Diagnostic Analytics (การวิเคราะห์เชิงวินิจฉัย)

คำถามหลัก: “ทำไมสิ่งนี้จึงเกิดขึ้น?”

สร้างต่อจาก Descriptive Analytics โดยเน้นหาสาเหตุของเหตุการณ์

แบ่งเป็น 2 แบบ:

แบบไม่เป็นทางการ (Informal Analysis): ใช้ตรรกะและสถิติพื้นฐาน เช่น

หากกำไรขั้นต้นลดลง ⟶ ตรวจสอบสัดส่วนสินค้า ⟶ พบว่าสินค้ามาร์จินต่ำขายเพิ่มขึ้นเพราะการตลาดโฆษณามาก

ใช้หลักการ “5 Why’s” ⟶ ถาม “ทำไม?” ซ้ำหลายครั้งจนกว่าจะเจอสาเหตุที่แท้จริง

แบบเป็นทางการ (Formal/Confirmatory Analysis): ใช้ **Hypothesis Testing** เพื่อตรวจสอบความสัมพันธ์หรือสมมติฐาน โดยกระบวนการคือ:

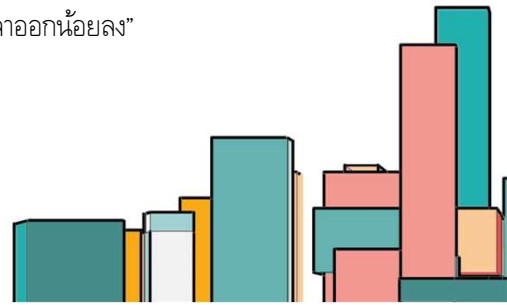
ระบุสมมติฐานศูนย์ (Null) และสมมติฐานทางเลือก (Alternative)

กำหนดระดับนัยสำคัญ (Significance Level)

เก็บตัวอย่างข้อมูลและคำนวณค่า P-value

เปรียบเทียบ P-value กับระดับนัยสำคัญ ⟶ ตัดสินใจว่าปฏิเสธหรือยอมรับสมมติฐาน

👉 สมมติฐานควรเขียนเป็น **ข้อความที่ทดสอบได้** ไม่ใช่คำถาม เช่น: “หากเราจ่ายค่าตอบแทนพนักงานเพิ่มขึ้น พนักงานจะมีแนวโน้มลาออกน้อยลง”



3. PREDICTIVE ANALYTICS

3. Predictive Analytics (การวิเคราะห์เชิงทำนาย)

คำถามหลัก: "สิ่งใดมีแนวโน้มจะเกิดขึ้นในอนาคต?"
ใช้ข้อมูลในอดีตเพื่อสร้าง **รูปแบบ (Patterns)** และคาดการณ์เหตุการณ์ในอนาคต
ตัวอย่างการใช้งาน:

Amazon ⟶ ใช้พฤติกรรมการค้นหา/การซื้อ เพื่อแนะนำสินค้า
Match.com ⟶ ใช้อัลกอริทึมทำนายความเข้ากันได้ของคู่รัก
Boston Medical Center ⟶ ใช้ "Hospital IQ" คาดการณ์ความต้องการผู้ป่วย ⟶ ปรับแผนบุคลากรและบริการสุขภาพ

ขั้นตอนการสร้าง Predictive Model

1. **เลือกตัวแปรเป้าหมาย (Target Variable)**
   - Categorical (เช่น "ลูกค้าจะซื้อ/ไม่ซื้อ")
   - Numeric (เช่น "ลูกค้าจะใช้จ่ายเท่าไหร่")
2. **เตรียมข้อมูลที่เหมาะสม (ETL)**
   - ข้อมูลยิ่งหลากหลาย ยิ่งมีประโยชน์ต่อการทำนาย
3. **สร้างและตรวจสอบโมเดล**
   - แบ่งข้อมูลเป็น **Training Set** และ **Test Set** ⟶ ป้องกันปัญหา *Overfitting*
   - วิธีที่ใช้:
     - ✓ Regression (Linear, Polynomial, Regression Trees) ⟶ สำหรับตัวเลข
     - ✓ Classification (Logistic Regression, Random Forests, Decision Trees, KNN, SVM) ⟶ สำหรับกลุ่ม/ประเภท

ประเด็นสำคัญ

- **Overfitting:** โมเดลแม่นกับข้อมูลเก่า แต่ทำนายข้อมูลใหม่ไม่ได้ ⟶ ต้องใช้ Test Set ตรวจสอบ
- **Model Update:** ความสัมพันธ์ของข้อมูลเปลี่ยนแปลง (เช่น E-commerce ⟶ Online Shopping) ⟶ โมเดลต้องอัปเดตอย่างต่อเนื่อง
- **Machine Learning:** ระบบเรียนรู้และปรับปรุงโมเดลโดยอัตโนมัติ เช่น การทำนายเครดิตลูกค้า

4. PRESCRIPTIVE ANALYTICS

4. Prescriptive Analytics (การวิเคราะห์เชิงกำหนด)

คำถามหลัก: “ควรทำอะไรต่อไป?”
ไม่ใช่แค่ทำนาย แต่ยังเสนอแนวทางปฏิบัติหรือการตัดสินใจอัตโนมัติ
ตัวอย่างการใช้งาน:

UPS ⟶ พัฒนา Prescriptive Model สำหรับเส้นทางส่งพัสดุ ⟶ ลดระยะทาง, ประหยัดเวลา, ลดการปล่อยคาร์บอน, ประหยัดต้นทุน $300–400 ล้าน/ปี
ใช้เทคนิค: AI, Machine Learning, Optimization, Simulation

DATA PRESENTATION

การแสดงผลข้อมูล (Data Visualization)

Choosing the Right Visualization

เหตุผลที่การแสดงผลข้อมูลสำคัญ

สำนวนที่ว่า *"ภาพหนึ่งภาพแทนคำนับพันคำ"* สอดคล้องกับงานวิจัยที่ยืนยันว่ามนุษย์ถูกออกแบบให้ **ประมวลผลข้อมูลเชิงภาพได้ดีกว่าข้อมูลที่เป็นตัวหนังสือ** โดยประโยชน์ของ Data Visualization ได้แก่:

- ✓ **ประมวลผลเร็วกว่า** การอ่านข้อมูลในตารางหรือตัวอักษร
- ✓ **ใช้งานง่ายกว่า** ผู้ใช้ไม่ต้องการคำอธิบายมากในการค้นหาข้อมูล
- ✓ **สอดคล้องกับรูปแบบการเรียนรู้หลักของคนส่วนใหญ่** (เป็น Visual Learners)

ตัวอย่างจากธุรกิจจริง

บริษัทขายตรงแห่งหนึ่งลงทุนสร้างโมเดลคาดการณ์ว่าลูกค้าในพื้นที่ใดมีแนวโน้มจะซื้อสินค้ามากที่สุด แต่เมื่อส่งผลการวิเคราะห์เป็นรายงานยาว ๆ ให้ทีมขาย ⟶ **ไม่ช่วยให้ขายได้จริง**

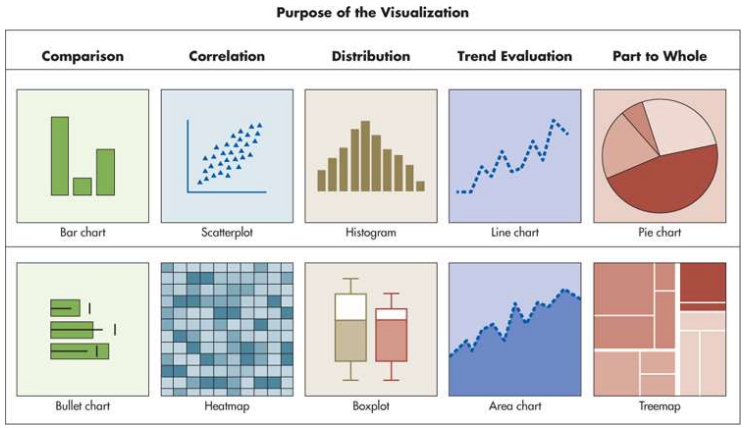
เมื่อบริษัทเปลี่ยนมานำเสนอข้อมูลเป็น **แผนที่ที่มี Layer สี** แสดงโอกาสหาลูกค้าใหม่ ⟶ ทีมขายเข้าใจทันทีว่าควรไปโฟกัสที่ไหนส่งผลให้มีแรงจูงใจมากขึ้นและสร้างผลลัพธ์ได้ทันที

ตาราง Cheat Sheet สรุป 5 วัตถุประสงค์หลักของ Data Visualization

Purpose	Visualization ที่เหมาะสม	ตัวอย่างการใช้งาน	ข้อดี / จุดเด่น
Comparison (การเปรียบเทียบ)	Bar Chart, Column Chart, Bullet Graph	เปรียบเทียบค่าตอบแทนระหว่างเพศ - ความพึงพอใจ vs ผลการทำงาน - Performance ปีนี้ vs ปีที่แล้ว	เห็นความแตกต่างระหว่างกลุ่มชัดเจน ใช้ง่าย
Correlation (ความสัมพันธ์)	Scatterplot (+ Regression Line)	วิเคราะห์ความสัมพันธ์ระหว่างราคาสินค้าและยอดขาย – ความสัมพันธ์ระหว่างค่าธรรมเนียมล่าช้าและการชำระตรงเวลา	แสดงให้เห็นความสัมพันธ์ของตัวแปรเชิงปริมาณ 2 ตัว
Distribution (การกระจายตัว)	Histogram, Boxplot	วิเคราะห์คะแนนสอบของนักเรียนทั้งห้อง – การกระจายของรายได้ลูกค้า	ระบุ Outliers และรูปแบบการกระจายของข้อมูลได้
Trend Evaluation (การดูแนวโน้มตามเวลา)	Line Chart, Area Chart	แนวโน้มยอดขายรายเดือน – จำนวนผู้ใช้งานแอปพลิเคชันต่อวัน	เห็นทิศทาง/การเปลี่ยนแปลงของข้อมูลตามกาลเวลา
Part-to-Whole (ส่วนต่อทั้งหมด)	Pie Chart, Stacked Bar, Treemap	สัดส่วนรายได้ตามหน่วยธุรกิจ – ส่วนแบ่งตลาดของคู่แข่ง	เข้าใจโครงสร้างส่วนประกอบภายในทั้งหมดได้ทันที

DATA PRESENTATION

ตาราง Visualization Purposes and Types



Data Visualization – Purposes & Examples

Purpose	Charts ที่เหมาะสม	Examples (ตัวอย่าง)	Usefulness (ข้อดี)
Comparison (การเปรียบเทียบ)	Bar Chart, Column Chart, Bullet Graph	- Gender pay gap - Job satisfaction vs performance - Performance year vs year	เข้าใจความแตกต่างระหว่างกลุ่ม/หมวดหมู่ชัดเจน
Correlation (ความสัมพันธ์)	Scatterplot, Heatmap	- Performance vs income - Job satisfaction vs performance - Training vs performance (levels)	เห็นความสัมพันธ์เชิงสถิติของตัวแปร 2 ตัว
Distribution (การกระจายตัว)	Histogram, Boxplot	- Salary distribution (bins of $10,000) - Performance rating distribution - Salary by department	ระบุ Outliers และรูปแบบการกระจายของข้อมูลได้
Trend Evaluation (แนวโน้มตามเวลา)	Line Chart, Area Chart	- Annual compensation change - Total compensation by department (per year)	แสดงการเปลี่ยนแปลงตามกาลเวลา เข้าใจทิศทาง
Part-to-Whole (ส่วนต่อทั้งหมด)	Pie Chart, Treemap	- % pay by department - Degree mix contribution	เข้าใจสัดส่วนองค์ประกอบของทั้งหมดได้ชัด
Other (อื่น ๆ)	Maps, Sankey Diagram, Combo Chart, Tables	- Data overlays on maps - Flow diagrams - Detailed reports in tables	ใช้เมื่อเป้าหมายซับซ้อน เช่น Spatial, Flow, หรือข้อมูลละเอียด

DATA PRESENTATION

Designing High-Quality Visualizations

3 หลักการออกแบบ Visualization ที่มีคุณภาพสูง

1. Simplification (การทำให้ง่ายขึ้น)

ความหมาย: ทำให้ Visualization เข้าใจง่าย ตรงประเด็น ลดความซับซ้อนที่ไม่จำเป็น

แนวทาง:

Title ⟶ กระชับ ชี้ insight โดยตรง
Axes ⟶ ตัด tick mark/เส้น/gridline ที่เกินความจำเป็น
Legend ⟶ แสดงเฉพาะ series สำคัญ หรือเขียน label บนกราฟเลย
Data Area ⟶ ใช้กราฟที่เหมาะสม, สีไม่เกิน 4–5 สี, หลีกเลี่ยง 3D/เอฟเฟกต์รบกวน

2. Emphasis (การเน้นประเด็นสำคัญ)

- **ความหมาย:** ทำให้ผู้ดูเห็น message หลักได้ทันที
- **แนวทาง:**
  - Title ⟶ ใช้ข้อความเชิง insight เช่น *"Q4 Sales Increased 20%"*
  - Axes ⟶ ใช้ scale ที่เหมาะสม ไม่บิดเบือน และช่วยเน้น pattern
  - Legend ⟶ ใช้สีหรือการจัดลำดับเพื่อเน้น series ที่สำคัญ
  - Data Area ⟶ ใช้การเน้นสี, ขนาด, หรือ annotation highlight จุดสำคัญ

3. Ethical Presentation (การนำเสนออย่างมีจริยธรรม)

ความหมาย: หลีกเลี่ยงการทำ Visualization ที่ทำให้เข้าใจผิดโดยตั้งใจหรือไม่ตั้งใจ

แนวทาง:

Title ⟶ ไม่ใช้ข้อความที่ชี้นำเกินจริง
Axes ⟶ ไม่ตัดแกน (axis truncation) จนทำให้การเปลี่ยนแปลงดูเว่อร์เกินจริง
Legend ⟶ ใช้สีที่เป็นกลาง ไม่ทำให้ข้อมูลถูกตีความผิด
Data Area ⟶ ใช้มาตราส่วนที่ถูกต้อง, ไม่ใช้การบิดเบือน perspective (เช่น Pie 3D ที่หลอกตา)

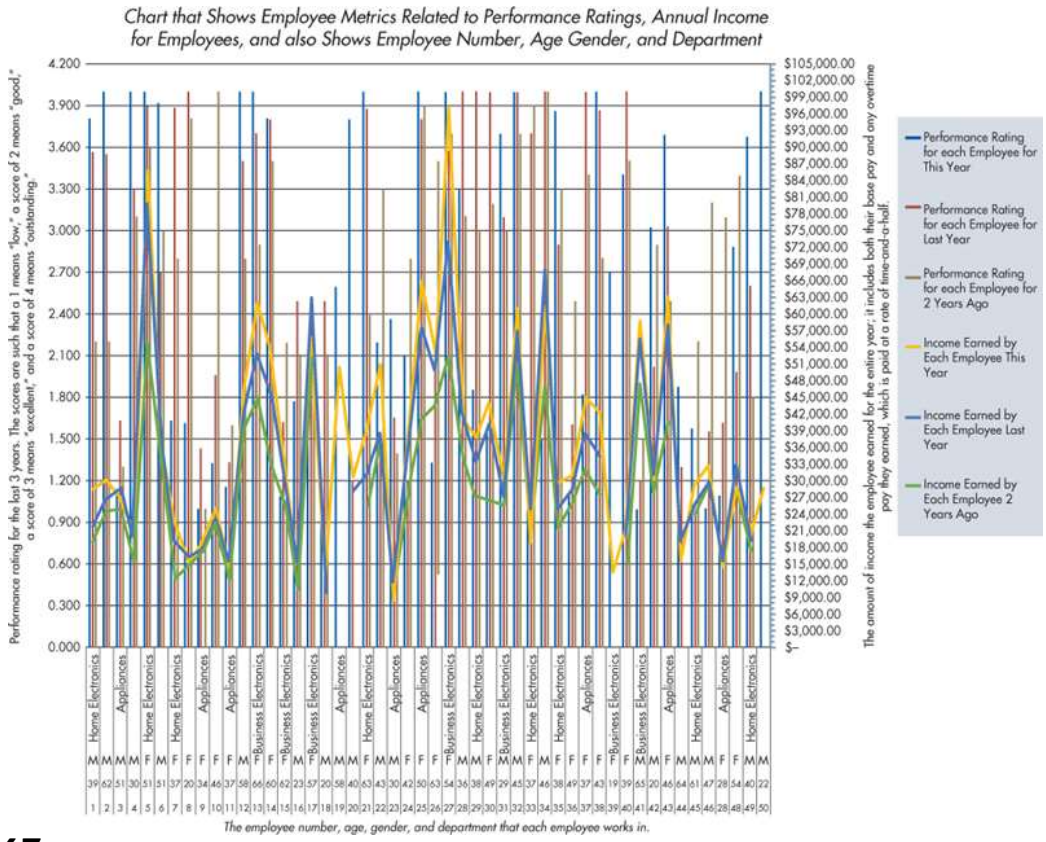
สรุป

Visualization ที่ดีต้องมี 3 สิ่งครบ:

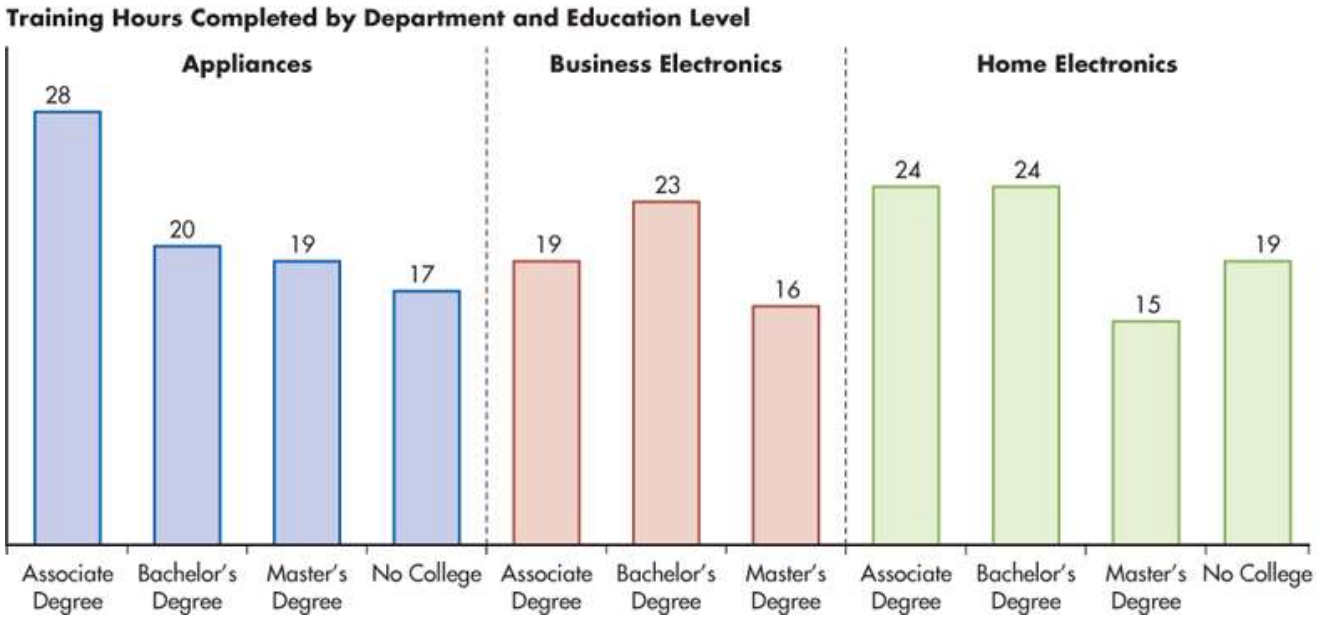
- **Simplification** ⟶ เข้าใจง่าย ไม่รก
- **Emphasis** ⟶ เห็นประเด็นสำคัญชัดเจน
- **Ethical Presentation** ⟶ ไม่บิดเบือนหรือทำให้ผู้ชมเข้าใจผิด

DATA PRESENTATION

ตัวอย่าง GRAPH ที่ดูได้ยาก

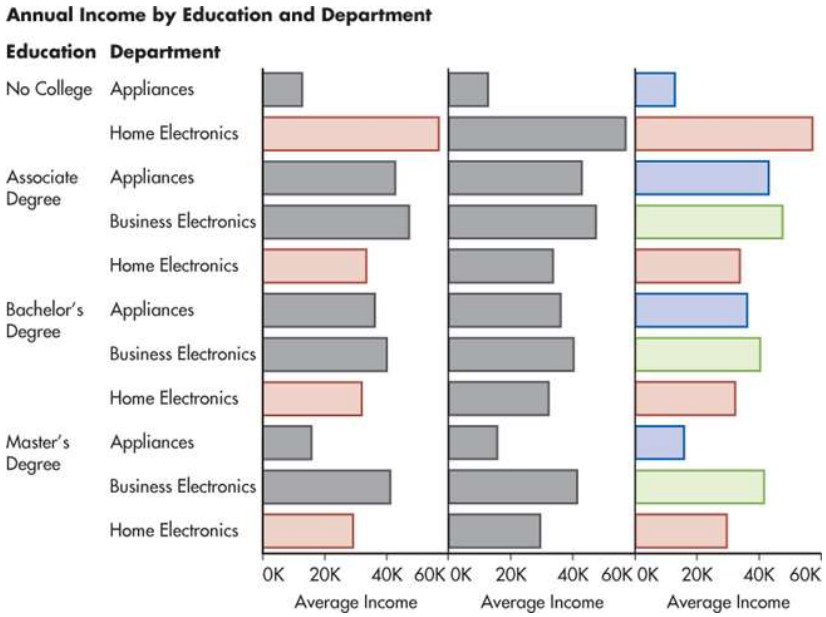


PRINCIPLES OF HIGH-QUALITY VISUALIZATION



Example of Simplifying a Visualization Using Distance

PRINCIPLES OF HIGH-QUALITY VISUALIZATION



1. **ซ้าย:** ใช้สีเน้น "แผนก Home Electronics" ⟶ ผู้ใช้หาได้ทันที เหมาะถ้าต้องการเน้นเฉพาะแผนกนี้
2. **กลาง:** ใช้สีเทาเหมือนกันหมด ⟶ ผู้ใช้ต้องตีความเองว่าอะไรสำคัญ ⟶ อาจพลาดสารที่ต้องการสื่อ
3. **ขวา:** ใช้สีหลายสีสำหรับแต่ละแผนก ⟶ เหมาะถ้าอยากเปรียบเทียบแผนก แต่ไม่ช่วยเน้นเฉพาะจุด

เทคนิคหลัก 3 แบบในการสร้าง Emphasis

1.Highlighting (การไฮไลต์/เน้นสี)

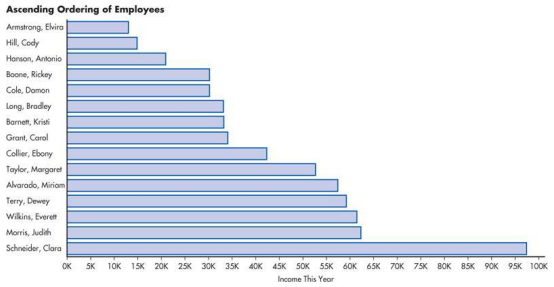
1. ใช้สี, ความต่าง (contrast), label, font, call-out, หรือแม้กระทั่งลูกศร เพื่อดึงความสนใจ
2. มักใช้ใน **Data Area** (เช่น บาร์, เส้น, พาย) มากที่สุด
3. **สี** เป็นเครื่องมือที่ทรงพลังที่สุดในการเน้นข้อมูล

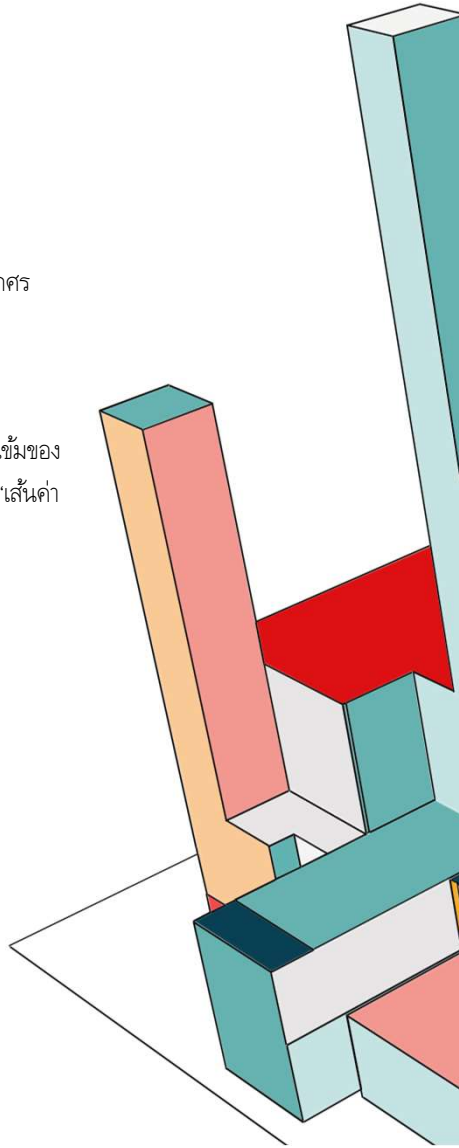
2.Weighting (การให้ความสำคัญด้วยขนาด/น้ำหนัก)ใช้ความหนาของเส้น, ความเข้มของสี, หรือขนาดของวัตถุ เพื่อบอกว่าส่วนไหนสำคัญกว่าเช่น กราฟเส้นแนวโน้มอาจทำ "เส้นค่าเป้าหมาย (Target Line)" หนากว่าเส้นอื่น

3.Ordering (การจัดลำดับ)

การจัดลำดับข้อมูลสามารถช่วยเน้นย้ำ insight ได้

เช่น การเรียงแท่งกราฟจากมากไปน้อย ⟶ ทำให้เห็นแผนกที่มียอดสูงสุดทันที





PRINCIPLES OF HIGH-QUALITY VISUALIZATION

Data Deception (การหลอกลวงด้วยข้อมูล) คือ "การนำเสนอข้อมูลในรูปแบบกราฟ โดยตั้งใจหรือไม่ตั้งใจก็ตาม ที่ทำให้ผู้ชมเข้าใจผิดหรือสร้างความเชื่อที่ไม่ตรงกับความจริงของข้อมูล"

หลักปฏิบัติเพื่อเลี่ยงการหลอกลวงข้อมูล

1.แสดงข้อมูลตามสัดส่วนจริง (Proportional Representation)

1. ตัวเลขที่นำเสนอควรสะท้อนค่าที่แท้จริง
2. ตัวอย่าง: **เริ่มแกน Y จากศูนย์ (0)** ⟶ ช่วยหลีกเลี่ยงการขยายความต่างให้ดูมากเกินจริง

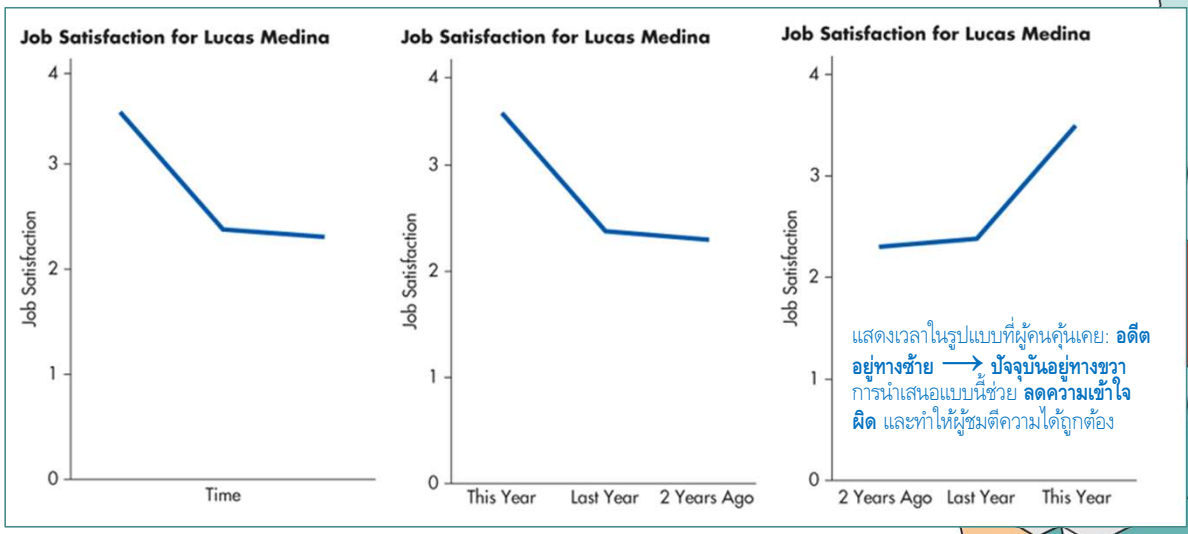
2.ใน Visualization ที่แสดงแนวโน้ม (Trends)

1. ควรให้ **เวลา (Time)** บนแกน X **แสดงจากซ้ายไปขวา**
2. สอดคล้องกับธรรมชาติของการอ่านข้อมูลและหลีกเลี่ยงการสร้างความสับสน

3.นำเสนอข้อมูลอย่างครบถ้วนตามบริบท (Present Complete Data)

1. ไม่ตัดข้อมูลที่สำคัญออกไปจนทำให้สารผิดเพี้ยน

เช่น การเลือกแสดงเฉพาะบางปี บางช่วง หรือบางกลุ่ม โดยไม่ชี้แจง ⟶ อาจทำให้ผู้ชมเข้าใจผิด



- **แสดงเหมือนกับว่า Lucas Median มีความพึงพอใจในงานลดลงเรื่อย ๆ ตามกาลเวลาซึ่งไม่ถูกต้อง เพราะแกนเวลา จัดเรียงกลับด้าน**

ปัญหาคือ **ผู้ใช้ส่วนใหญ่คาดหวังว่าเวลา ต้องเรียงจากซ้าย ⟶ ขวา (เก่า ⟶ ใหม่)** การกลับทิศเวลา **ทำให้ผู้ชมสับส**

THANK YOU

Manatchaya Sriphanlam