ข้อมูลและชนิดของข้อมูล (Data and Types of data)

“**ข้อมูล (Data)**” คือ ข้อเท็จจริง หรือ สิ่งที่ถือว่าเป็นความจริงสำหรับใช้เป็นหลักอนุมานหาความจริง หรือ การคำนวณ โดย ณ ปัจจุบัน ข้อมูลถูกแสดงในหลายรูปแบบ อาทิเช่น ตัวเลข (numbers), ข้อความ (text), วัน-เวลา (date-times), รูปภาพ (images), เสียง (audio) และ วิดีโอ (video) เป็นต้น แต่ละรูปแบบของข้อมูลจะมีการดำเนินการที่ไม่เหมือนกัน อาทิเช่น ข้อมูลตัวเลขจะสามารถดำเนินการคำนวณต่าง ๆ ได้ เช่น +, −, ×, ÷ และอื่น ๆ ส่วนข้อความจะสามารถทำการเชื่อมต่อกันระหว่างข้อความ (concatenate) ได้

ข้อมูลแบบมีโครงสร้างและไม่มีโครงสร้าง (Structured and Unstructured data)

ข้อมูลแบบมีโครงสร้าง (Structured data) จะเป็นข้อมูลที่ที่เป็นข้อสังเกตหรือคุณลักษณะต่าง ๆ ซึ่งโดยส่วนใหญ่จะถูกจัดเก็บอยู่ในรูปแบบตารางข้อมูลที่จะประกอบด้วยแถวและคอลัมน์ของข้อมูล ตัวอย่างเช่น ผลการทดลองที่ถูกจัดเก็บโดยนักวิทยาศาสตร์ที่มักจะถูกจัดเก็บอย่างเป็นระเบียบ เป็นต้น ข้อมูลแบบมีโครงสร้างจะเป็นข้อมูลที่จัดการได้ค่อนข้างง่ายและสามารถเข้าใจได้โดยง่ายเนื่องจากมีรูปแบบที่ตายตัว ซึ่งสามารถประยุกต์ใช้กับโมเดลทางด้านสถิติ (Statistical models) และโมเดลทางด้านการรู้จำเครื่อง (Machine learning models) ได้โดยง่าย

ข้อมูลแบบไม่มีโครงสร้าง (Unstructured data) จะเป็นข้อมูลที่ถูกจัดเก็บแบบอิสระไม่มีการจัดเรียงหรือมาตรฐานในการจัดเก็บ อาทิเช่น ข้อมูลประเภทข้อความต่าง ๆ อีเมล์ (e-mail) ล็อกของเซิร์ฟเวอร์ (server logs) หรือ ข้อความที่ถูกโพสบนเฟซบุ๊ค (Facebook posts) เป็นต้น ข้อมูลแบบไม่มีโครงสร้างจะเป็นข้อมูลที่พบเห็นได้โดยง่ายและมีประมาณ 80 – 90% ของข้อมูลทั้งหมดบนโลกที่ซึ่งจะต้องทำการประมวลผลข้อมูลเบื้องต้น (Preprocessing) เพื่อให้อยู่ในรูปแบบที่พร้อมใช้งาน

ข้อมูลเชิงปริมาณและข้อมูลเชิงคุณลักษณะ (Quantitative and qualitative data)

ข้อมูลเชิงปริมาณ (Quantitative data) จะสามารถเรียกอีกอย่างหนึ่งว่า **Numerical data** เป็นข้อมูลในเชิงวัดผล เกี่ยวกับขนาด เกี่ยวกับปริมาณ หรือ การนับที่เป็นจำนวน โดยแบ่งย่อยได้อีก 2 ประเภทคือ

- **Discrete data** คือข้อมูลที่สามารถ**นับได้** ระบุจุดสิ้นสุดได้ในระยะเวลาใดเวลาหนึ่ง เช่น จำนวนลูกค้าที่สั่งซื้อสินค้าประจำเดือนจำนวน 500 คน
- **Continuous data** หมายถึงข้อมูลที่**ไม่สามารถนับได้ แต่สามารถวัด/ประเมินได้** เช่น ความยาว, น้ำหนัก, ส่วนสูง เป็นต้น

ข้อมูลเชิงคุณลักษณะ (Qualitative data) จะสามารถเรียกอีกอย่างหนึ่งว่า Categorical data เป็นข้อมูลในเชิงคุณลักษณะที่ไม่มีผลทางคณิตศาสตร์ เช่น เพศ, สีที่ชอบ, ประเภทสินค้า ซึ่งข้อมูลประเภทนี้เราอาจจะแทนชุดข้อมูลนี้ด้วยตัวเลขได้ เช่น 1 = Yes, 0 = No แต่ไม่มีความหมายในเชิงคณิตศาสตร์ (mathematical meaning) โดยแบ่งย่อยได้อีก 2 ประเภทคือ

- **Nominal**—ข้อมูลที่ถูกจัดเก็บในรูปแบบ category ที่ซึ่งมักถูกเรียกว่าเป็น categorical ตัวอย่างเช่น หิน สามารถแบ่งประเภทได้เป็น หินอัคนี (igneous rock), หินชั้น/หินตะกอน (sedimentary rock) และ หินแปร (metamorphic rock) เป็นต้น
- **Ordinal**—ข้อมูลที่ถูกจัดเก็บเป็น rank order of scores (1st, 2nd, 3rd, etc.) ตัวอย่างเช่น คะแนนรีวิวสินค้า 1 ดาว — 5 ดาว เมื่อเราทำการสำรวจความพึงพอใจในสินค้า ลูกค้าแต่ละคนทำการรีวิว 1 ดาว คือชอบน้อยสุดไปจนถึง 5 ดาว คือชอบมากสุด ทำให้ตัวเลข 1–5 นั้นมีความหมายในเชิงคณิตศาสตร์

นอกเหนือจากข้างต้น เราสามารถแยกประเภทข้อมูลออกเป็นในเชิงคณิตศาสตร์, เชิงสถิติ และเชิงวิทยาการคอมพิวเตอร์ ที่ซึ่งเราจะสามารถแยกประเภทของข้อมูลได้ดังนี้

- **ประเภทของข้อมูลเชิงคณิตศาสตร์ (Types of data in Mathematics)**
  - **Integers**—ตัวเลขจำนวนเต็มอาจมีค่าบวกหรือลบก็ได้
  - Rational Numbers—ตัวเลขที่แสดงถึงผลหารระหว่าง 2 integers : p/q โดยที่ค่าของ q จะต้องมีค่าไม่เท่ากับ 0
  - **Real Numbers**—ตัวเลขที่รวมถึง rational numbers และ irrational numbers (ตัวอย่างเช่น √2 = 1.41421356, **π** = 3.14159265... และ e = 2.71828...)
  - **Imaginary Numbers**—ตัวเลขที่ค่า square ของตัวเลขมีค่าน้อยกว่าหรือเท่ากับ 0 ตัวอย่างเช่น √-25 จะเป็น imaginary number เนื่องจากค่า square ของ √-25 มีค่าเท่ากับ -25

- **ประเภทของข้อมูลเชิงสถิติ (Types of data in Statistics)**

- **Nominal**—ข้อมูลที่อธิบายถึงชื่อหรือหมวดหมู่ ที่ซึ่งมักถูกเรียกว่าเป็น categorical ตัวอย่างเช่น เพศ เชื้อชาติ ศาสนา เป็นต้น
- **Ordinal**—ข้อมูลที่ถูกจัดเก็บเป็น rank order of scores (1st, 2nd, 3rd, etc.) ตัวอย่างเช่น คะแนนรีวิวสินค้า 1 ดาว — 5 ดาว เมื่อเราทำการสำรวจความพึงพอใจในสินค้า ลูกค้าแต่ละคนทำการรีวิว 1 ดาว คือชอบน้อยสุดไปจนถึง 5 ดาว คือชอบมากสุด ทำให้ตัวเลข 1–5 นั้นมีความหมายในเชิงคณิตศาสตร์
- **Interval**—ข้อมูลที่แสดงถึง**ความแตกต่างระหว่างข้อมูล** และยังสามารถระบุถึงความเท่าเทียมกันระหว่างความแตกต่าง 2 interval ใด ๆ ได้ ตัวอย่างเช่น ความแตกต่างระหว่างอุณหภูมิ 100 และ 90 องศาเซลเซียส จะเท่ากับความแตกต่างระหว่าง 90 และ 80 องศา หรืออีกตัวอย่างหนึ่ง คือ 100 ปีระหว่างศตวรรษที่ 20 และ 21 จะเท่ากับ 100 ปีระหว่างศตวรรษที่ 21 และ 22 เป็นต้น
- **Ratio**—เป็นข้อมูลแบบ interval data ที่มีการระบุถึงค่า 0 อย่างมีความหมาย ที่ซึ่งค่า 0 ใน ratio จะหมายถึงการไม่ปรากฏขึ้น ตัวอย่างเช่น ค่าความสูงเป็น 0 จะหมายถึงไม่มีความสูง หรือ ค่าน้ำหนัก 0 กรัม หมายถึง ไม่มีน้ำหนัก เป็นต้น นอกจากนั้นจะไม่มีความแตกต่างระหว่าง 2 ข้อมูลที่มีค่าเป็นลบ ตัวอย่างเช่น การวัดส่วนสูงด้วยเซนติเมตร นิ้ว หรือฟุต จะเป็นข้อมูลประเภท ratio เนื่องจาก ค่าความสูงไม่สามารถมีค่าติดลบได้ แต่ในทางกลับกัน ค่าของคุณอุณหภูมิจะสามารถมีค่าติดลบได้เช่น -10 องศา (แต่จะไม่มีความสูง -10 เซนติเมตร หรือ -10 นิ้ว เป็นต้น)

Provides:	Nominal	Ordinal	Interval	Ratio
The “order” of values is known		✔	✔	✔
“Counts,” aka “Frequency of Distribution”	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has “true zero”				✔

รูปที่ 1 คุณลักษณะของข้อมูลในเชิงสถิติ[1]

[1] http://www.mymarketresearchmethods.com/wp-content/uploads/2016/05/summary-of-data-types-and-scales.png

- ประเภทของข้อมูลเชิงวิทยาการคอมพิวเตอร์ (Types of data in Computer science)
  - Bit—เป็นข้อมูลหน่วยเล็กที่สุดที่สามารถแสดงได้ 2 ข้อมูล คือ 1 หรือ 0 ซึ่งข้อมูลใน 1 bit จะเรียกว่า binary data และเมื่อนำข้อมูลมามาเรียงต่อกันจะเรียกว่า byte ที่ซึ่งจะสามารถจัดเก็บข้อมูลอยู่ในช่วง 0 – 255 (00000000 – 11111111) ตัวอย่างเช่น byte หนึ่งๆมีข้อมูล 10110100 = 180 เป็นต้น
  - Boolean—เป็นข้อมูลเชิงตรรกะที่สามารถแสดงข้อมูลได้ 2 ค่าข้อมูล คือ “true” และ “false” ที่ซึ่งสามารถประยุกต์ใช้กับการเปรียบเทียบ อาทิเช่น x = y ? ถ้าคำตอบคือใช่ ค่าคำตอบของการเปรียบเทียบจะมีค่าเป็น “true” แต่ในทางกลับกัน ค่าคำตอบของการเปรียบเทียบจะมีค่าเป็น “false” ตามลำดับ
  - Alphanumeric—เป็นข้อมูลที่ใช้สำหรับจัดเก็บตัวอักษร/อักขระที่มีการเรียงต่อกัน (a – z, A – Z, 0 – 9, และอักขระพิเศษ) ที่เรียกว่า สายอักขระ (string)
  - Integers—เป็นข้อมูลตัวเลขจำนวนเต็มที่สามารถเป็นทั้งตัวเลขจำนวนเต็มบวกและจำนวนเต็มลบ ที่ซึ่งจัดเก็บอยู่ในลักษณะทั้งแบบ signed และ unsigned
  - Floating point—เป็นข้อมูลจำนวนจริง หรือเลขทศนิยม

อ้างอิง

- https://en.wikibooks.org/wiki/Data_Science:_An_Introduction/Definitions_of_Data