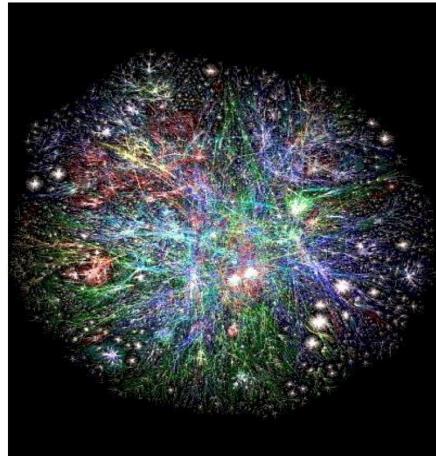
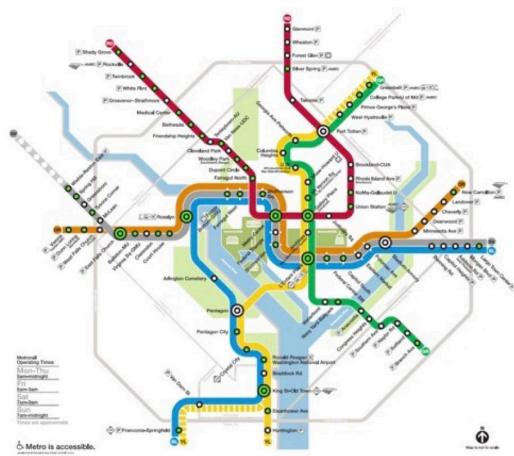


1. ประเภทของข้อมูลและออบเจกต์ (Data Objects and Attribute Types)

เนื้อหาส่วนแรกปัจจุบันเกี่ยวกับโครงสร้างของข้อมูล:

- **ประเภทของชุดข้อมูล (Types of Data Sets):**

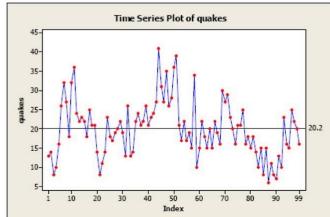
- **Record Data:** ข้อมูลแบบระเบียบ เช่น ตารางในฐานข้อมูล (Relational tables) หรือ Data matrix
- **Graphs and Networks:** ข้อมูลแบบกราฟและเครือข่าย เช่น โครงสร้างเว็บ (World Wide Web), โครงสร้างโมเลกุล, หรือ Social networks



- **Ordered Data:** ข้อมูลที่มีลำดับ เช่น ข้อมูลวิดีโอ, ข้อมูลอนุกรมเวลา (Time-series), หรือ ข้อมูลพันธุกรรม (Genetic sequence)

- ❑ Video data: sequence of images

- ❑ Temporal data: time-series



- ❑ Sequential Data: transaction sequences

```

Human: CTTTTGAGG ... TGTTCGACAAATGGCTCTTCTATTCAGAGCTGCCCA
Chimpanzee: GTTTTGAGG ... ATGTCATTAATGGCTCTTCTATTCAGAGCTGCCCA
Macaque: GTTTTGAGG ... ATGTCATTAATGGCTCTTCTATTCAGAGCTGCCCA

Human: GACAAATTCTGCTAGCGCTTGTCTATTATCTGTTTCTAACCTGAAATTGGAGTGT
Chimpanzee: GACAAATTCTGCTAGCGCTTGTCTATTATCTGTTTCTAACCTGAAATTGGAGTGT
Macaque: GACAAATTCTGCTAGCGCTTGTCTATTATCTGTTTCTAACCTGAAATTGGAGTGT

Human: GATCTTGAGACTTAA ... TCTGAAATAAACTAGCCTGATTTATTTTCTAACCTGAA
Chimpanzee: CACTGGAGACTAAACTGTAATAAAAATAGCTGATTTATTTTCTAACCTGAA
Macaque: TACTGGAGACTAAACTGTAATAAAAATAGCTGATTTATTTTCTAACCTGAA

Human: CAGAAATGAGATTAGCAAATTAGCTCTCTAGAACTATTTGATTTGCTATATGCTGA
Chimpanzee: CAGAAATGAGATTAGCAAATTAGCTCTCTAGAACTATTTGATTTGCTATATGCTGA
Macaque: CAGAAATGAGATTAGCAAATTAGCTCTCTAGAACTATTTGATTTGCTATATGCTGA

Human: CCCCTGGTTGAAATTTTCTGTTGAGCCCTATGTCACCTTCATAAGCCAGCTATAGA ...
Chimpanzee: CCCCTGGTTGAAATTTTCTGTTGAGCCCTATGTCACCTTCATAAGCCAGCTATAGA ...
Macaque: CCCCTGGTTGAAATTTTCTGTTGAGCCCTATGTCACCTTCATAAGCCAGCTATAGA ...

Human: GACAGGTTAGTTAAAAAAACATTTTTATGCTAGTTTTTGCAAGGTTTTAAATTTC
Chimpanzee: GACAGGTTAGTTAAAAAAACATTTTTATGCTAGTTTTTGCAAGGTTTTAAATTTC
Macaque: GACAGGTTAGTTAAAAAAACATTTTTATGCTAGTTTTTGCAAGGTTTTAAATTTC

```

```

Human: H A T T T S T L S K
Chimpanzee: AACTTGCGCGCTGTGGTAA ...
Macaque: AACCTGGTGGCATGTGGTAA ...

```

```

Human: AAC TAAAC AAC TG AG TACA ...
Chimpanzee: AAC TGGCGCGCTGTGGTAA ...
Macaque: AAC TGGCGCGCTGTGGTAA ...

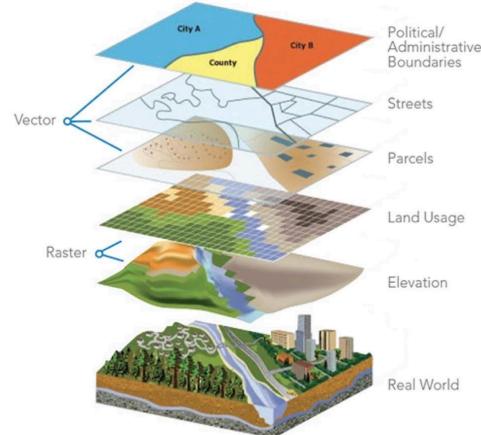
```

- ❑ Genetic sequence data

6

- Spatial, image and multimedia Data

- ❑ Spatial data: maps



- ❑ Image data:

- ❑ Video data:

7

- **Data Objects:** หมายถึง เอบนที (Entity) ในข้อมูล เช่น ลูกค้า, สินค้า, หรือผู้ป่วย ซึ่งอาจเรียกว่า samples, examples, หรือ tuples โดยถูกอธิบายด้วยคุณลักษณะ (Attributes)
- **ประเภทของ Attribute:**
 - **Nominal:** ข้อมูลที่เป็นกลุ่มหรือชื่อเรียก เช่น สีผม, สถานะการแต่งงาน
 - **Binary:** มี 2 สถานะ (0 หรือ 1) และเป็น Symmetric (ความสำคัญเท่ากัน เช่น เพศ) และ Asymmetric (ความสำคัญไม่เท่ากัน เช่น ผลตรวจนครุ ซึ่งมักให้ผล Positive เป็น 1)
 - **Ordinal:** ข้อมูลที่มีลำดับชั้นแต่ไม่ทราบระยะห่างที่แน่นอน เช่น ขนาด (เล็ก, กลาง, ใหญ่) หรือยศทหาร

- **Numeric:** แบ่งเป็น *Interval* ("ไม่มีศูนย์แท้ เช่น อุณหภูมิของศาส泽ลเชียส") และ *Ratio* ("มีศูนย์แท้ เช่น น้ำหนัก, จำนวนเงิน")
- **Discrete vs Continuous:** ข้อมูลแบบไม่ต่อเนื่อง (นับจำนวนได้) vs ข้อมูลต่อเนื่อง (ทศนิยม)

2. การบรรยายข้อมูลด้วยสถิติพื้นฐาน (Basic Statistical Descriptions of Data)

เพื่อทำความเข้าใจลักษณะของข้อมูลทั้งในด้านแนวโน้มเข้าสู่ส่วนกลางและการกระจายตัว:

- **การวัดแนวโน้มเข้าสู่ส่วนกลาง (Central Tendency):**
 - **Mean:** ค่าเฉลี่ยเลขคณิต (รวมถึงแบบถ่วงน้ำหนักและแบบตัดค่าสุดโต่ง หรือ Trimmed mean)
 - **Median:** ค่ามัธยฐาน คือค่าที่อยู่ตรงกลางเมื่อเรียงข้อมูล เหมาะสำหรับข้อมูลที่มีค่าเบี่ยงเบน (Skewed)
 - **Mode:** ฐานนิยม คือค่าที่เกิดขึ้นบ่อยที่สุด
- **การวัดการกระจายตัว (Dispersion):**
 - **Variance & Standard Deviation:** ความแปรปรวนและส่วนเบี่ยงเบนมาตรฐาน เป็นค่าที่ใช้วัดว่าข้อมูลกระจายออกจากค่าเฉลี่ยมากเพียงใด
 - **Quartiles & Boxplots:** การใช้ค่าวอร์ไชล์ (Q1, Q3) และ Inter-quartile range (IQR) เพื่อสร้างแผนภูมิกล่อง (Boxplot) ซึ่งช่วยระบุค่าผิดปกติ (Outliers) ได้
- **การแสดงผลด้วยกราฟ (Graphic Displays):** เช่น Histogram (แสดงความถี่), Quantile plots (แสดงการกระจาย), และ Scatter plot (ดูความสัมพันธ์ของข้อมูลสองตัวแปร)

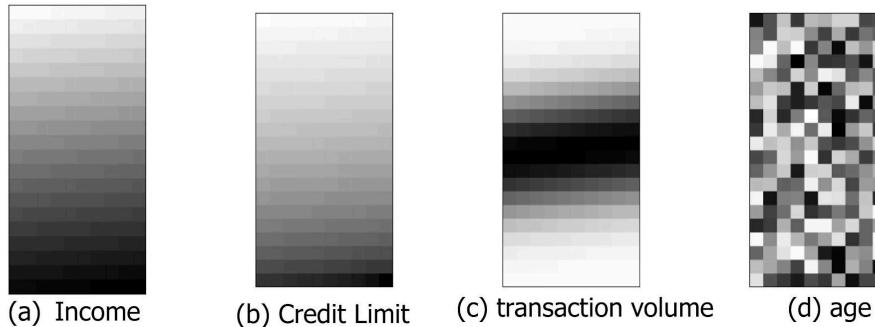
3. การสร้างภาพข้อมูล (Data Visualization)

การแปลงข้อมูลเป็นภาพเพื่อช่วยค้นหาแพทเทิร์น ความผิดปกติ หรือโครงสร้างข้อมูล:

- **วัดคุณภาพสังเคราะห์:** ช่วยให้เห็นภาพรวมเชิงคุณภาพ (Qualitative overview) ของข้อมูลขนาดใหญ่ และค้นหาความสัมพันธ์ที่ซ่อนอยู่
- **เทคนิคต่างๆ:**
 - **Pixel-oriented:** ใช้สีของพิกเซลแทนค่าของข้อมูล

Pixel-Oriented Visualization Techniques

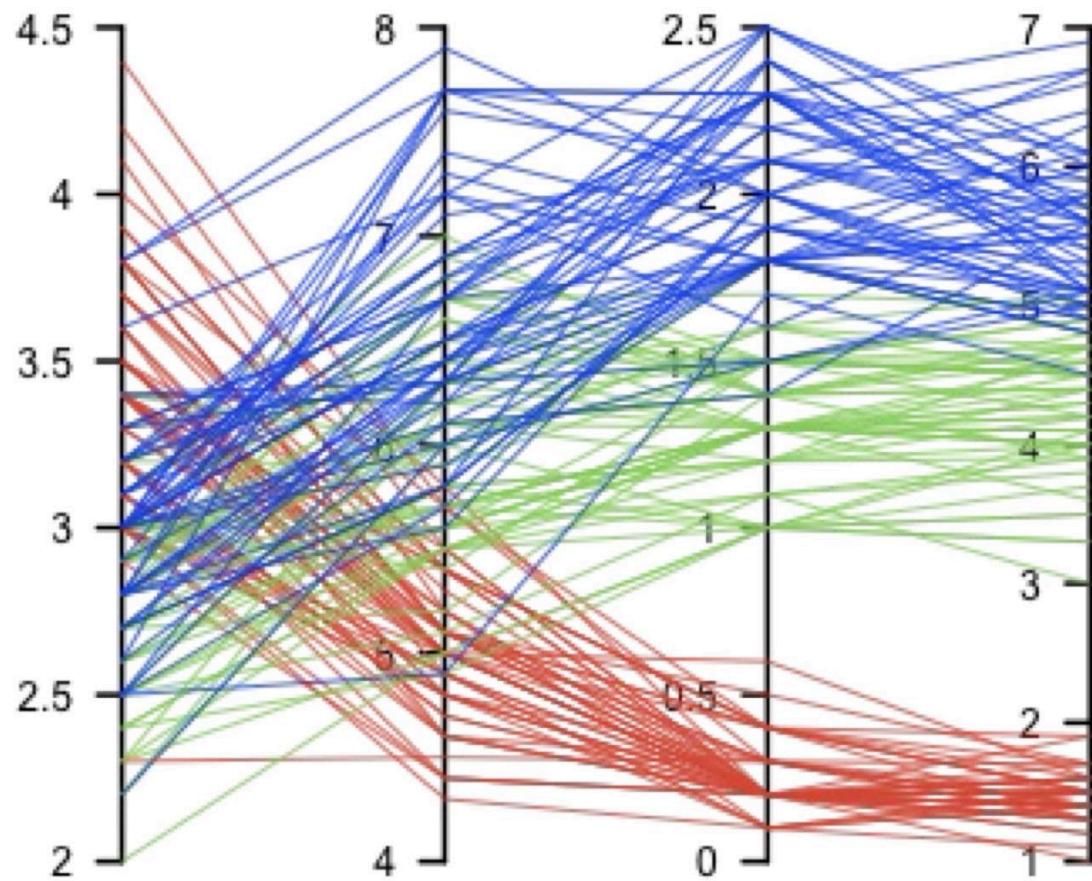
- ❑ For a data set of m dimensions, create m windows on the screen, one for each dimension
- ❑ The m dimension values of a record are mapped to m pixels at the corresponding positions in the windows
- ❑ The colors of the pixels reflect the corresponding values



34

- **Parallel Coordinates:** ใช้เส้นแกนนานกันแทนแต่ละ Attribute และลากเส้นเชื่อมโยงข้อมูล

Parallel coordinate plot, Fisher's Iris data



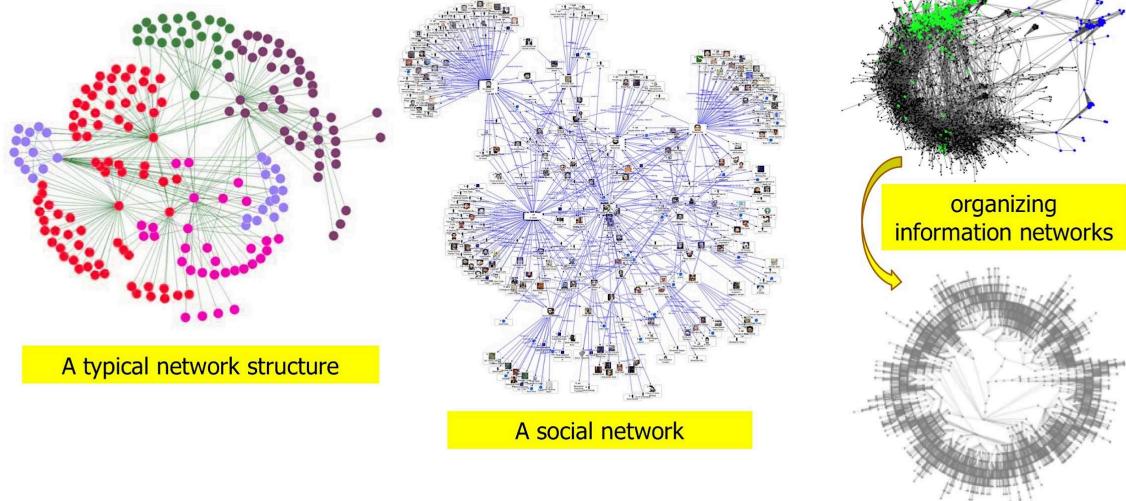
Sepal Width Sepal Length Petal Width Petal Length
— setosa — versicolor — virginica

- Tree-Map: แสดงข้อมูลแบบลำดับชั้นโดยใช้พื้นที่สีเปลี่ยนช้อนกัน
- Tag Cloud: แสดงความสำคัญของคำด้วยขนาดตัวอักษร

Newsmap: Google News Stories in 2005

Visualizing Complex Data and Relations: Social Networks

- Visualizing non-numerical data: **social and information networks**



53

4. การวัดความเหมือนและความต่าง (Measuring Data Similarity and Dissimilarity)

เป็นพื้นฐานสำคัญสำหรับงาน Data Mining เช่น การจัดกลุ่ม (Clustering):

- **Similarity vs Dissimilarity:**
 - **Similarity:** ค่าความเหมือน ยิ่งมากยิ่งเหมือน (มักอยู่ในช่วง [0, 1])
 - **Dissimilarity (Distance):** ค่าความต่าง ยิ่งน้อยยิ่งเหมือน (มักเริ่มที่ 0)
- **การวัดระยะทางสำหรับข้อมูลประเภทต่างๆ:**
 - **Numeric Data:** ใช้ Minkowski Distance ซึ่งมีกรณีพิเศษคือ Manhattan distance (L1) และ Euclidean distance (L2) รวมถึง Supremum distance (L-infinity)
 - **Binary Attributes:** ใช้ตาราง Contingency table โดยเฉพาะ Jaccard coefficient สำหรับข้อมูลแบบ Asymmetric
 - **Nominal/Categorical:** ใช้การจับคู่กันแบบง่าย (Simple matching) หรือเปลี่ยนเป็น Binary
 - **Ordinal:** เปลี่ยนค่าเป็น Rank และแปลงให้อยู่ในช่วง [0, 1] ก่อนคำนวณเหมือนข้อมูล Numeric
 - **Mixed Types:** ใช้สูตรถ่วงน้ำหนักเพื่อรวมผลการวัดจาก Attribute หลายประเภทเข้าด้วยกัน
 - **Cosine Similarity:** ใช้สำหรับวัดความคล้ายคลึงของเอกสาร (Document) โดยดูจากมุมระหว่างเวกเตอร์ความถี่ของคำ