



CS 412 Intro. to Data Mining



Chapter 2. Getting to Know Your Data



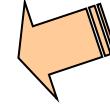
Jiawei Han, Computer Science, Univ. Illinois at Urbana-Champaign, 2017





Chapter 2. Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary



Types of Data Sets: (1) Record Data

ໜ້າຍຕາງກີ່ມີຄວາມສ່ວນໃຫ້ດັ່ງ

- Relational records
 - Relational tables, highly structured
 - Data matrix, e.g., numerical matrix, crosstabs

	China	England	France	Japan	USA	Total
Active Outdoors Crochet Glove		12.00	4.00	1.00	240.00	257.00
Active Outdoors Lycra Glove		10.00	6.00		323.00	339.00
InFlux Crochet Glove	3.00	6.00	8.00		132.00	149.00
InFlux Lycra Glove		2.00			143.00	145.00
Triumph Pro Helmet	3.00	1.00	7.00		333.00	344.00
Triumph Vertigo Helmet		3.00	22.00		474.00	499.00
Xtreme Adult Helmet	8.00	8.00	7.00	2.00	251.00	276.00
Xtreme Youth Helmet		1.00			76.00	77.00
Total	14.00	43.00	54.00	3.00	1,972.00	2,086.00

Person:

Pers_ID	Surname	First_Name	City
0	Miller	Paul	London
1	Ortega	Alvaro	Valencia
2	Huber	Urs	Zurich
3	Blanc	Gaston	Paris
4	Bertolini	Fabrizio	Rom

no relation

Car:

Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	150000	4
105	Renault	1998	2000	3
106	Renault	2001	7000	3
107	Smart	1999	2000	2

- Transaction data

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

	team	coach	y	pla	ball	score	game	n	wi	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2		
Document 2	0	7	0	2	1	0	0	3	0	0		
Document 3	0	1	0	0	1	2	2	0	3	0		

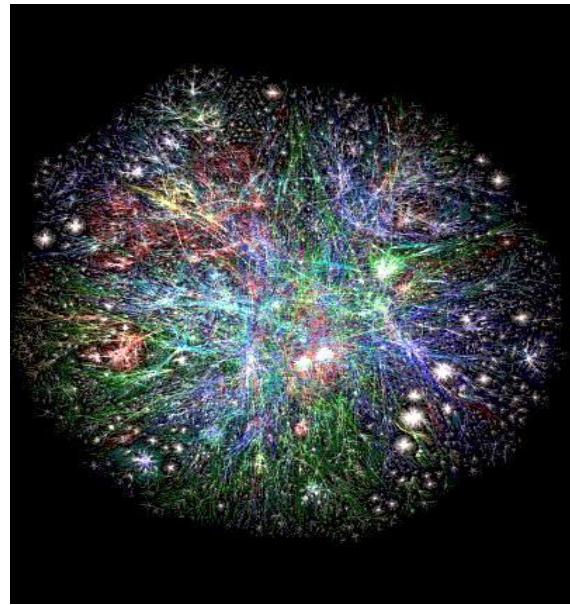
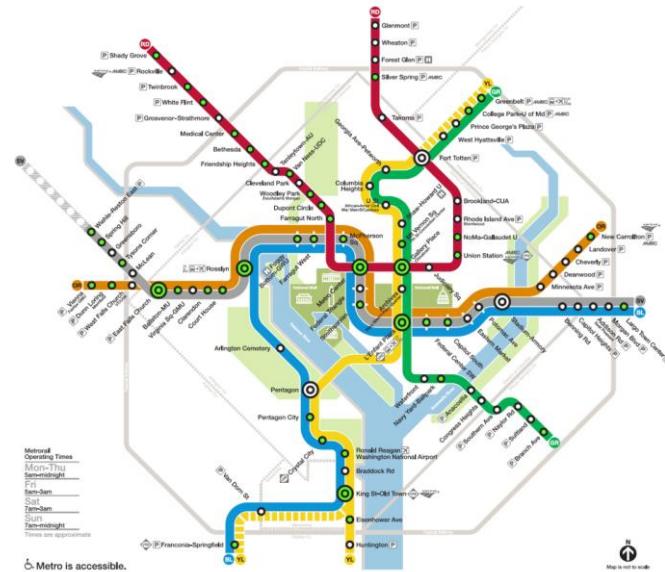
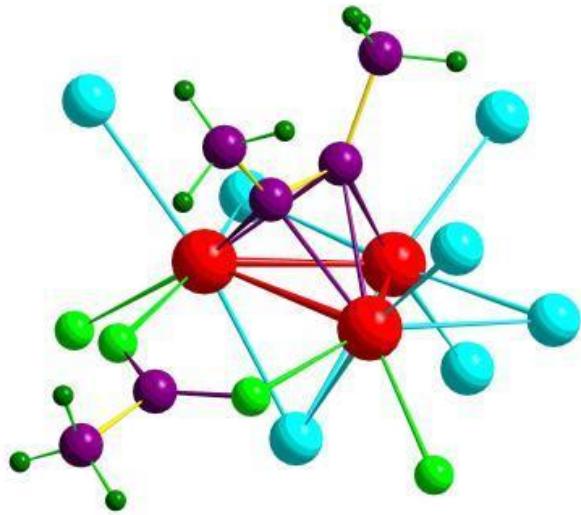
ຈຳນວດ

Data ສະໝັບຊັບຄວາມ

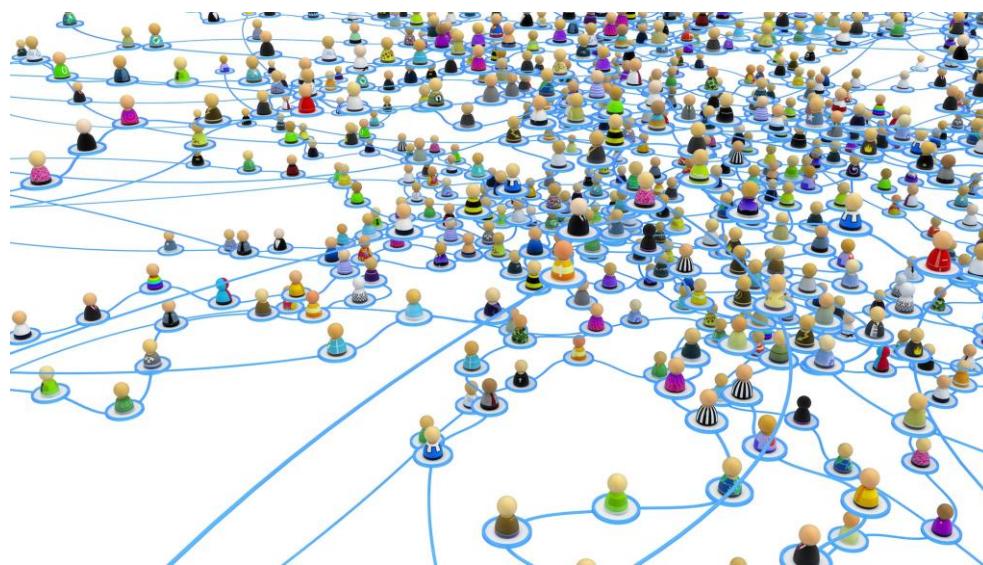
- Document data: Term-frequency vector (matrix) of text documents

Types of Data Sets: (2) Graphs and Networks

- Transportation network
- World Wide Web

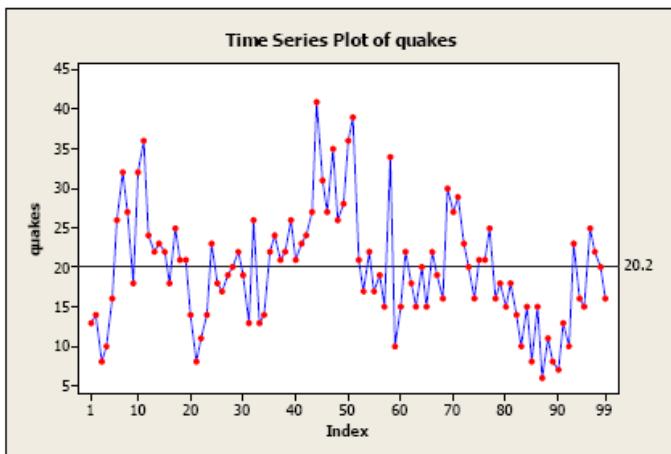
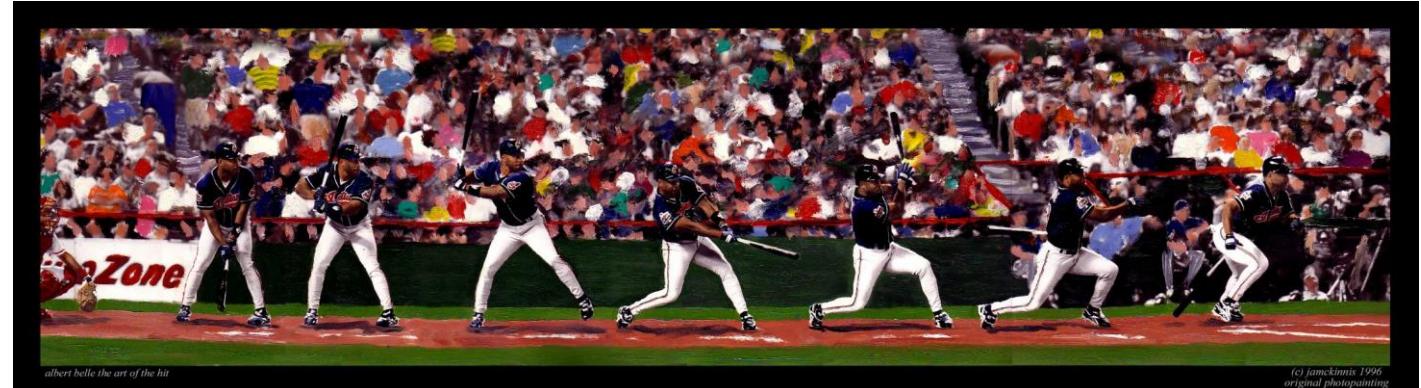


- Molecular Structures
- Social or information networks



Types of Data Sets: (3) Ordered Data

- ## ▢ Video data: sequence of images



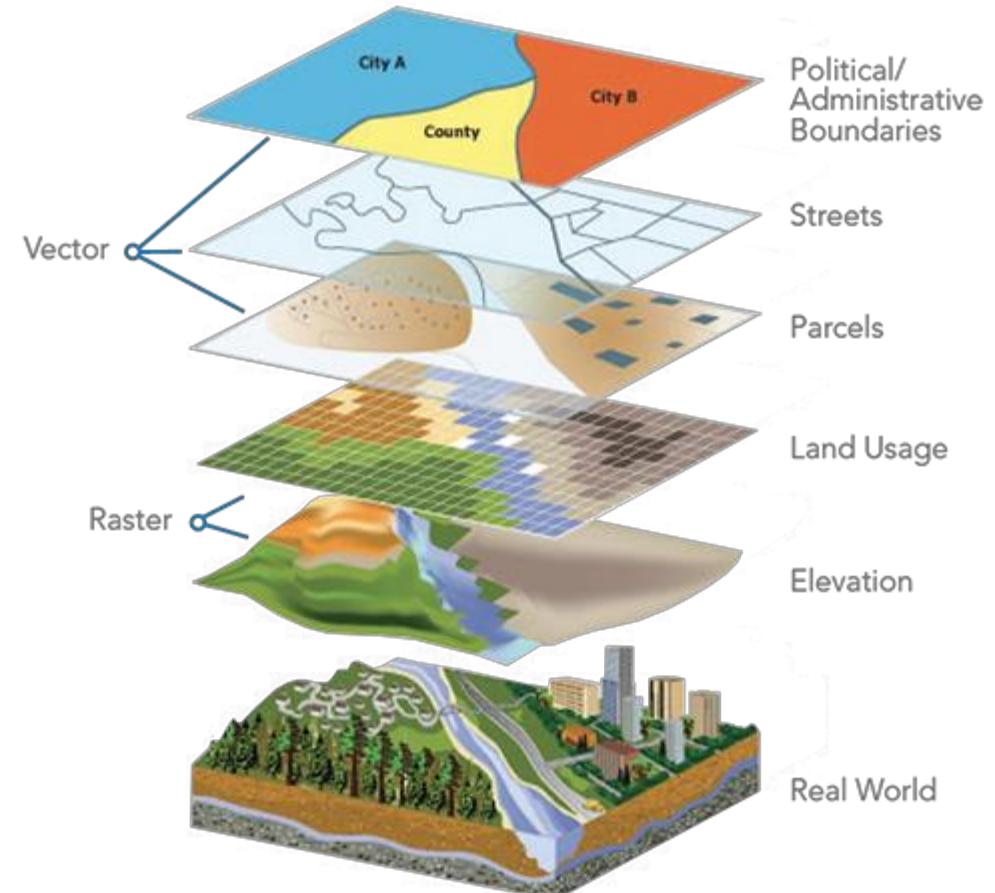
- ## ❑ Sequential Data: transaction sequences

- ## Genetic sequence data

ចំណុះអេដ្ឋស៊ីបៀវតានការងារក្នុងលាង x,y

Types of Data Sets: (4) Spatial, image and multimedia Data

□ Spatial data: maps



□ Image data:

□ Video data:

Important Characteristics of Structured Data

- Dimensionality → ເຄີເມນ້ອຫັນກໍ 2,3,4,5
- Curse of dimensionality
- Sparsity → ສາມາດຮັບກຳລັງຂອ້ງລູກ, ຂັ້ນລູກເປັນ 0 ເພື່ອຈະໄຟສັນໄວ
- Only presence counts
- Resolution → ເກີນຂອ້ງລູກໃສ່ສັນແລ້ວ
- Patterns depend on the scale
- Distribution → ວັດຄ່າລາງວ່າສິ້ນນີ້ສ່າງເຫັນວັດ
- Centrality and dispersion



Data Objects



- Data sets are made up of data objects → ດັບກຳຂອບຂໍ້ມູນນີ້ແກ່ລາຍງານ Data
ຂົ້ນສົກລະຕົກ: ຕົກ
- A **data object** represents an entity
- Examples:
 - sales database: customers, store items, sales
 - medical database: patients, treatments
 - university database: students, professors, courses
- Also called *samples*, *examples*, *instances*, *data points*, *objects*, *tuples*
- Data objects are described by **attributes** → ຂໍ້ມູນນີ້ຈະຖືກອໍານວຍເຫັນ attributes
- Database rows → data objects; columns → attributes



Attributes

- Attribute (or dimensions, features, variables)
 - A data field, representing a characteristic or feature of a data object.
 - E.g., *customer_ID, name, address*
- Types:
 - Nominal (e.g., red, blue) : ចំណែកប៉ាន្ត់ទៅតាមរយៈលេខ
 - Binary (e.g., {true, false}) : ចំណែកប៉ាន្ត់ទៅតាមរយៈលេខ
 - Ordinal (e.g., {freshman, sophomore, junior, senior})
 - Numeric: quantitative (+, -, ×, ÷ និងគោលគោល)
 - Interval-scaled: 100°C is interval scales
 - Ratio-scaled: 100°K is ratio scaled since it is twice as high as 50°K
- Q1: Is student ID a nominal, ordinal, or interval-scaled data?
- Q2: What about eye color? Or color in the color spectrum of physics?
↳ Numeric



Attribute Types

□ Nominal: categories, states, or “names of things”

- Hair_color = {auburn, black, blond, brown, grey, red, white}
- marital status, occupation, ID numbers, zip codes

□ Binary : เนื่องจาก Nominal เต็มไปด้วย 0 กับ 1 ใช้กันได้ทุกอย่าง

- Nominal attribute with only 2 states (0 and 1)
- Symmetric binary: both outcomes equally important
 - e.g., gender Left, Right-handed / Coke, Pepsi / Hot, Cold
- Asymmetric binary: outcomes not equally important.
 - 2 ค่าความสำคัญไม่เท่ากัน เช่น ตาราง Covid ระบาดใน 2 ค่า ใจ狠 1000 คน ใจ肠软 10 คน
 - e.g., medical test (positive vs. negative)
 - Convention: assign 1 to most important outcome (e.g., HIV positive)

□ Ordinal → ไม่สามารถนำมารวมกันได้, สามารถรتبลำดับได้

- Values have a meaningful order (ranking) but magnitude between successive values is not known
- Size = {small, medium, large}, grades, army rankings



Numeric Attribute Types

- ❑ Quantity (integer or real-valued)

- ❑ Interval 0 ແຫ່ງ ການດິນສອ 0 ແກ່ງ ດັ່ງ 0 ແຫ່ງ

- ❑ Measured on a scale of **equal-sized units**
- ❑ Values have order
 - ❑ E.g., *temperature in C° or F°, calendar dates*
- ❑ No true zero-point

- ❑ Ratio

- ❑ Inherent **zero-point**
- ❑ We can speak of values as being an order of magnitude larger than the unit of measurement (10 K° is twice as high as 5 K°).
 - ❑ e.g., *temperature in Kelvin, length, counts, monetary quantities*



Discrete vs. Continuous Attributes

- ❑ **Discrete Attribute** (ກະບ່າງຄ່າ 2 ຕໍ່ໄຟສັດທຶນທຽງກລາວ ເຊັ່ນ ແລ້ວ ກຍບາລ ປິ່ນປົວເຫັນທຽງກລາວ)
- ❑ Has only a finite or countably infinite set of values
 - ❑ E.g., zip codes, profession, or the set of words in a collection of documents
- ❑ Sometimes, represented as integer variables
- ❑ Note: **Binary** attributes are a special case of discrete attributes
- ❑ **Continuous Attribute** (ລົດກລາຍ:ນ່ຳງວ່າ 2 ດ້ວຍ ສູງ 180 ກັບ ສູງ 181 ໂຕລົດສູງ 180.5)
- ❑ Has real numbers as attribute values
 - ❑ E.g., temperature, height, or weight
- ❑ Practically, real values can only be measured and represented using a finite number of digits
- ❑ Continuous attributes are typically represented as floating-point variables



Chapter 2. Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary



ເລື່ອດຳກັນສະຕິພາບ ໃຊ້ເສດຖາ data ຂອງເກີດຕາມທີ່ເຊື້ອດົວ

Basic Statistical Descriptions of Data

□ Motivation

- To better understand the data: central tendency, variation and spread

□ Data dispersion characteristics

- Median, max, min, quantiles, outliers, variance, ...

□ Numerical dimensions correspond to sorted intervals

- Data dispersion:

- Analyzed with multiple granularities of precision

- Boxplot or quantile analysis on sorted intervals

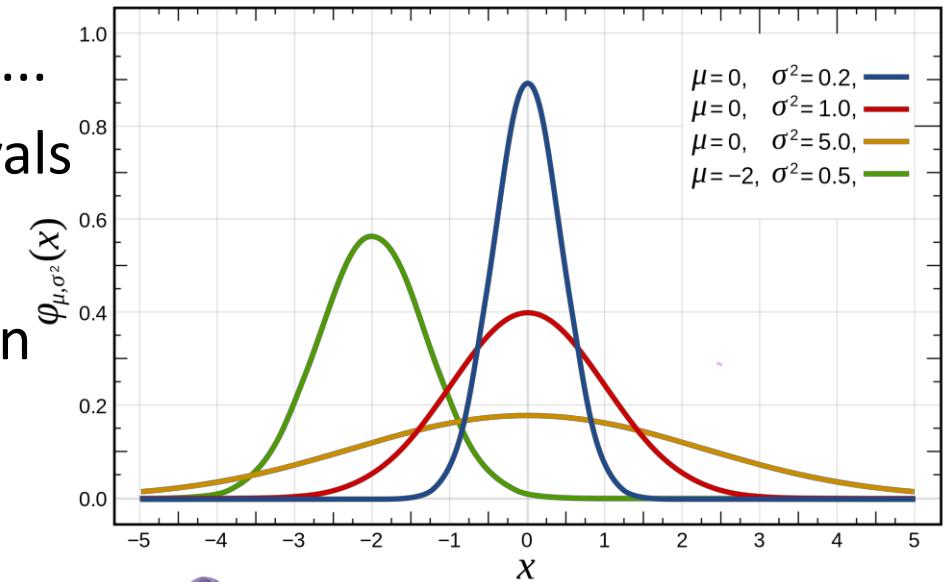
□ Dispersion analysis on computed measures

- Folding measures into numerical dimensions

- Boxplot or quantile analysis on the transformed cube

ເນື່ອງແນ່ນຈຳກຳຂອງການກົດ
ໄຫວຕົວກຳກຳຂອງການຈົກປະກັນຢ່າງໄດ້ໃນເນັດ

ກຳນົດຕົວກຳກຳຂອງການຈົກປະກັນຢ່າງໄດ້ໃນເນັດ



ຄະຫຼວງໃນນິ້ງໃນນິ້ງໃນນິ້ງ 2017
-ຮຽນໃຈ (ຫົວຍອດ)