Data Processing

by Dr. Matinee Kiewkanya
Adapted into English by Dr. Prakarn Unachak

ภาควิชาวิทยาการคอมพิวเตอร์
COMPUTER SCIENCE DEPARTMENT, CMU
คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

204100 IT AND MODERN LIFE

OUTLINES

1. Terminology

2. Data Categories and Desirable Properties

3. Types of Data Processing

4. Steps in Data Processing

5. Data Organization

6. Examples of Database

ภาควิชาวิทยาการคอมพิวเตอร์
COMPUTER SCIENCE DEPARTMENT, CMU
คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

1. Terminology

1. Terminology

Data, or Raw Data

- ❑   Facts about people, place or other objects of interest

- ❑   Examples: numbers, texts, pictures, temperatures, students' scores
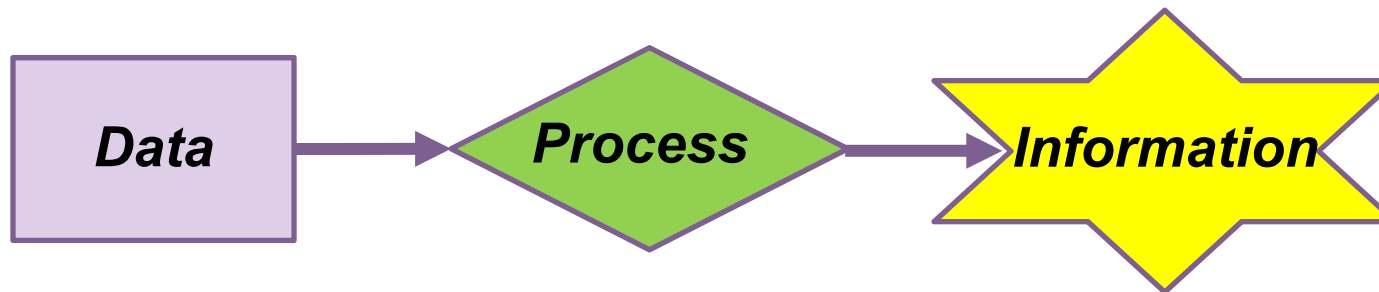
Information

- ❑   End product of data processing

- ❑   Usable for the intended purpose

- ❑   Examples: grades, average score, weather forecast

- ❑   One person's information can be other's raw data!

1. Terminology (cont.)

Data Processing *is a process that summarizes, analyzes or otherwise converts data into information which can be used in decision process*

2. Data Categories and Desirable Properties

2. Data Categories and Desirable Properties

2.1 Categories of Data

   2.2.1 Categorized by Source

   2.2.2 Categorized by Data Type

2.2 Desirable Qualities of Data

2.1.1 Categories of Data (by Sources)

1. *Primary Data are data directly gathered from the original sources of data. Can be gathered by direct observation, interviewing, questionnaires, surveying, or even those collected by electronic or mechanical means such as card reader, or pollution sensors*

2. *Secondary Data are data already collected and published by others. These data are publicly available, or can be accessed with permission. Examples are various statistics collected by government agencies such as import/export numbers in a year, or population in each country at a certain time*

ภาควิชาวิทยาการคอมพิวเตอร์
COMPUTER SCIENCE DEPARTMENT, CMU
คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

2.1.2 Categories of Data (by Data Types)

1. Text Data _are data that consists of characters such as A-Z, a-z, other symbols such as $ ( ) ? + - * % , or numbers_ but they do not represent numerical values, and are not used in calculation

Examples:

- **_Name_**
- **_Address_**
- **_Telephone Number_**
- **_Student ID_**

2.1.2 Categories of Data (by Data Types)

2. Numeric Data *are data consists* <u>*only of number*</u> *they represent numerical values, and can be used in calculation and other comparison*

Examples

- **Ages**

- **Salary**

- **Price**

2.1.2 Categories of Data (by Data Types)

3. Image Data represent images such as a drawing or a photograph. As computer file, they can be made of pixels or components of the image. *Examples:* Photo on ID card, image files

4. Audio Data are recorded sound. *Examples:* phone call recording, music songs

5. Video Data are animations, essentially series of images ordered by time. They can be accompanied by audio data. *Examples:* surveillance video clips, show recording

ภาควิชาวิทยาการคอมพิวเตอร์
COMPUTER SCIENCE DEPARTMENT, CMU
คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

2.2 Desirable Qualities of Data

1. **Accuracy**

   ➢ *Incorrect data can lead to incorrect decision, lack of credibility, or even more serious damage (in Medicine, for example)*

2. **Timeliness**

   ➢ *Data need to be up-to-date, for most accurate outlook*

   ➢ *Need update as often as possible, but have to balance with the cost of collection*

   ➢ *Ex: US Census every 10 years*

2.2 Desirable Qualities of Data (cont.)

3.   **Completeness**

  ➢   *Data need to cover everything we intend to use*

4.   **Compactness**

  ➢   *Regarding data storage*

  ➢   *How to use as little space as possible to hold the data without any loss of meaning?*

  ➢   *May need encoding*

5.   **Fitness of Purpose**

  ➢   *What are we collecting/processing these data for?*

  ➢   *Need to survey the users of data/information (just you, or larger organization) what they need*

3. Types of Data Processing

3.1 Manual Data Processing

❑ *"Processing by hands" done solely by human*

❑ *Can be assisted by scratch pads or abacus*

❑ *Used for small data size, or something you don't need in a long time*

Examples:

- *Counting money*

- *Using abacus to calculated total expenses*

3.2 Mechanical Data Processing

❑ **Semi-automatic Data Processing**

❑ **Still performed by human, but with helps of (more) sophisticated tools**

❑ **Used for moderate data size, or something you don't right away**

Examples

- *Using typewriter to type out a report*
- *Summing up numbers with a calculator*
- *Using an accounting machine*

3.3 Electronic Data Processing (EPD)

❑ **Electronic Data Processing is data processing with helps of electronic tools, or, data processing by computers**

❑ **Used for very large data size, with a lot of repetitive steps**

❑ **And/or need to be accurate and done very quickly**

❑ **And/or very complex works**

3.3 Electronic Data Processing (cont.)

There are 2 main types of EPD:

1) Batch Processing

- ❑ **Some amount of data (a batch) are collected over a period of time (day, week, month, year,...)**

- ❑ **At the end of the period, the batch is then processed**

- ❑ **Sometime called Off-Line System**

Examples

- **Collections of daily sales data, which is used to update the inventory at the end of the day**

- **Performing regional survey, whose results are then collected and processed on the national level**

3.3 Electronic Data Processing (cont.)

2) Real-time Processing

❑ **Processing happens right when a point of data occurs**

❑ **Input units (that collect data) and output units (that display information) are connected to main processing units at all time of operation, allowing data processing to happen at any time**

❑ **Sometime called On-Line System**

Examples

- **Air ticketing system**
- **Deposit/withdrawal with an ATM**

4. Steps in Data Processing

4. Steps in Data Processing

4.1 Data Preparation

4.2 Procesing

4.3 Results Management and Presentation

4.1 Data Preparation

❑ **Collecting and managing data so that they have appropriate format and structure for the processing, to get the information we need**

Steps in Data Preparation

4.1.1 Collection

4.1.2 Grouping

4.1.3 Encoding

4.1.4 Editing

4.1.5 Recording

4.1.1 Collection

A few examples of data collection

❖ Observation or Exploration

❖ Measurement

❖ Interview

❖ Questionnaire

    ○ Paper survey

    ○ Online survey

An Example of Paper Survey Form

แบบสำรวจความคาดหวังของผู้ปกครองในหลักสูตรวิทยาศาสตรบัณฑิต

สาขาวิทยาการคอมพิวเตอร์

<u>ตอนที่ 1</u> ข้อมูลทั่วไปของผู้ปกครอง

คำชี้แจง : กรุณาทำเครื่องหมาย ✔ ลงในช่อง ☐ หรือเติมคำตอบลงในช่องว่างที่กำหนดให้

1. เพศ ☐ หญิง ☐ ชาย
2. อายุ .... ปี
3. ภูมิลำเนา จังหวัด ....
4. ที่อยู่ปัจจุบัน จังหวัด ....
5. สถานภาพสมรส ☐ โสด ☐ สมรส ☐ หม้าย/หย่า/แยก
6. ระดับการศึกษา
   ☐ ต่ำกว่าประถมศึกษา ☐ ประถมศึกษา
   ☐ มัธยมศึกษา ☐ อนุปริญญา
   ☐ ปริญญาตรี ☐ ปริญญาโท
   ☐ ปริญญาเอก
7. อาชีพ ....
8. รายได้ต่อเดือน .... บาท

ตอนที่ 2 ความคาดหวังต่อการเรียนการสอนในหลักสูตรวิทยาศาสตรบัณฑิต สาขาวิทยาการ
คอมพิวเตอร์

คำชี้แจง : โปรดอ่านข้อคำถาม แล้วทำเครื่องหมาย ✔ ลงในช่องหลังข้อความตามระดับความคาดหวัง
ของท่านในแต่ละหัวข้อดังนี้

1=คาดหวังน้อยที่สุด	2=คาดหวังน้อย	3=คาดหวังปานกลาง
4=คาดหวังมาก	5=คาดหวังมากที่สุด	

ข้อคำถาม	ระดับความคาดหวัง				
	1	2	3	4	5
1. รายวิชามีความหลากหลาย น่าสนใจ และทันสมัย					
2. จำนวนหน่วยกิตตลอดหลักสูตร มีความเหมาะสม (130 หน่วยกิต)					
3. ระยะเวลาตลอดหลักสูตร มีความเหมาะสม (4 ปี)					
4. มีสื่อและเทคโนโลยีภายในห้องเรียนที่มีมาตรฐานและทันสมัย					
5. มีการศึกษาดูงานนอกสถานที่					
6. มีการฝึกงานหรือสหกิจศึกษาที่เหมาะสม					
7. มีการทำโครงงานวิจัยระดับปริญญาตรี					
8. มีการฝึกทักษะการเขียน การพูด เพื่อนำเสนอผลงานสู่สาธารณชน					
9. มีกระบวนการเรียนการสอนที่ส่งเสริมคุณธรรม จริยธรรม					
10. มีกิจกรรมสร้างรายได้ให้แก่นักศึกษาระหว่างเรียน					
11. มีกิจกรรมเสริมหลักสูตรฯ ที่เป็นประโยชน์ต่อนักศึกษา					
12. มีระบบช่วยเหลือนักศึกษาทั้งด้านวิชาการและด้านอื่น ๆ					

An Example of Paper Survey Form (Translated)

Parent Expectation Survey for Computer Science Graduate

<u>Section 1</u>: *Parent's General Information*

Notice: Please check ✓ *on the appropriate box* ☐, *or fill in the details under provided spaces*

1. **Gender** ☐ *Female*   ☐ *Male*

2. **Age**   ....*Years*

3. **Place of birth**   *Province* ....

4. **Current Address**   *Province* ....

5. **Marital Status**   ☐ *Single*   ☐ *Married*   ☐ *Divorce/Widowed/Separated*

6. **Education**

    ☐ *Below Elementary Level*   ☐ *Elementary Level*

    ☐ *Secondary Level*   ☐ *Associated Degree*

    ☐ *Bachelor Degree*   ☐ *Master Degree*

    ☐ *Doctoral Degree*

7. **Occupation**   ....

8. **Monthly Income**   ....*Baht*

ภาควิชาวิทยาการคอมพิวเตอร์
COMPUTER SCIENCE DEPARTMENT, CMU
คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

Section 2: Expectation Regarding Computer Science Bachelor Curriculum

Notice: Please check ✓ on the box after each statement that match your expectation level, where:
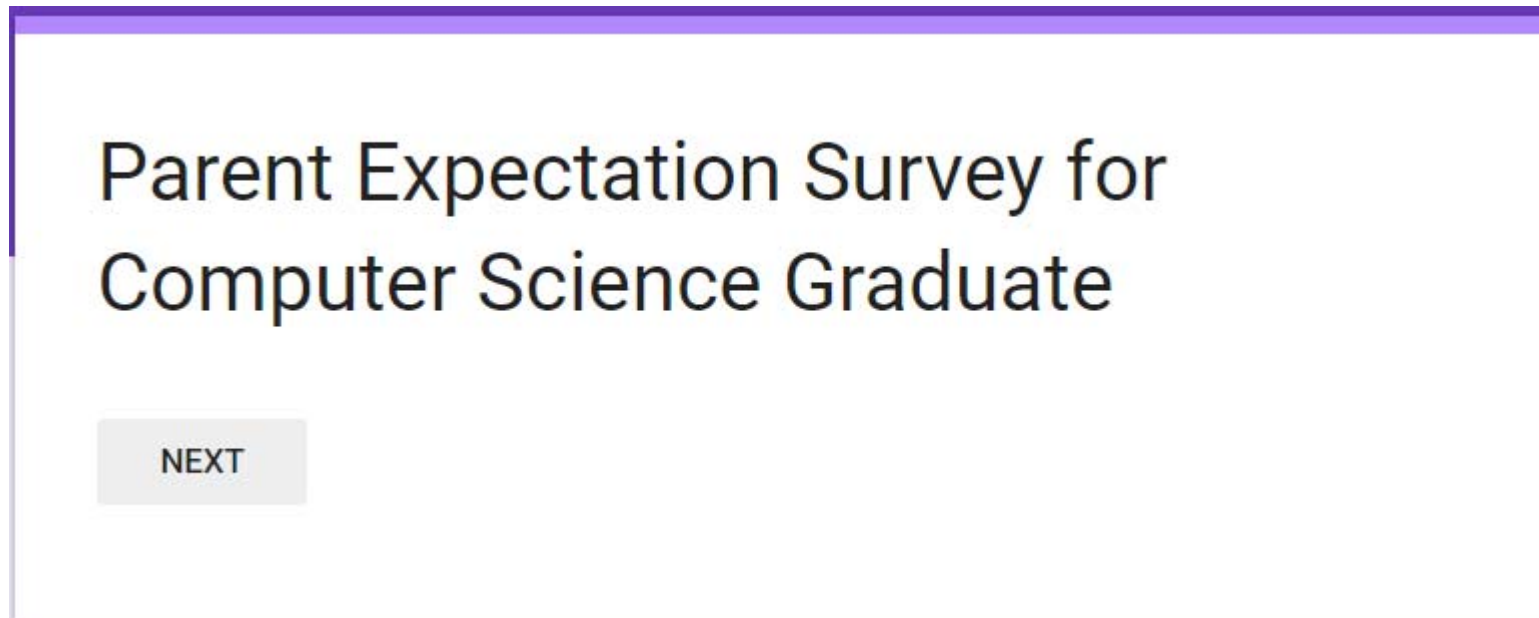
1 = least expected	2 = little expected	3 = moderately expected
4 = highly expected	5 = most highly expected	

Statement	Expectation Level				
	1	2	3	4	5
1. Courses are diverse, interesting, and up-to-date					
2. Appropriate number of credit in the curriculum (130 credits)					
3. Appropriate length of the curriculum (4 years)					
4. Classroom media and technology are up to standard and up-to-date					
5. There are out-of-classroom activities					
6. There are appropriate work training or cooperative opportunity					
7. There are bachelor-level research opportunity					
8. There are writing and public speaking training					
9. There are course about ethics					
10. There are working opportunity during the course of studies					
11. There are useful extracurricular activities					
12. There are support systems, academically and other issues					

Example of Online Survey

* Google Form Example

Parent Expectation Survey for Computer Science Graduate

NEXT

Section 1: Parent General Information

1. Gender

○ Female

○ Male

Multiple choice

2. Age

Your answer

Short answer

3. Place of Birth, Province

Choose ▼

Drop-down

4. Current Address, Province

Choose ▼

5. Marital Status

○ Single

○ Marrieds

○ Divorce/Widowed/Separated

6. Educational Level

○ Elementary

○ Secondary

7. Occupation

Your answer

8. Monthly Income

Your answer

BACK   NEXT

Section 2: Expectation Regarding Computer Science Bachelor Curriculum

Notice: Please select our expectation level for each statement, where:
1 = least expected   2 = little expected   3 = moderately expected
4 = highly expected   5 = most highly expected

Statements

	1	2	3	4	5
Courses are diverse, interesting, and up-to-date	○	○	○	○	○
Appropriate number of credit in the curriculum (130 credits)	○	○	○	○	○
Appropriate length of the curriculum (4 years)	○	○	○	○	○
Classroom media and technology are up to standard and up-to-date	○	○	○	○	○
There are out-of-classroom activities	○	○	○	○	○
There are appropriate work training or cooperative opportunity	○	○	○	○	○
There are bachelor-level research opportunity	○	○	○	○	○

ภาควิชาวิทยาก
COMPUTER SCIENC
คณะวิทยาศาสตร์ มหาวิทย

31

4.1.2 Grouping

❑ **Raw data you have collected may be divided into groups so that:**

1. **You can focus only on groups that interest you**

2. **You can gain more insights on each group of data**

Examples:

➢ **Grouping survey answers by respondent's genders**

➢ **Grouping parents by province they are currently living in**

4.1.3 Encoding

❑ *Encoding means assigning (shorter) codes to represent data*

❑ *The code can be a combination of characters, numbers, or both*

❑ *Encoding can save storage space, and allowing faster data access*

❑ *For numerical data, you can use rescaling, changing units of data (1000 kg instead of kg, for example), so that you will need smaller number to represent data*

Example of Raw Data

Gender	Age	Birth Province	Current Province	Marital Status	Education	Occupation	Monthly Income (Baht)
Male	35	Chiang Mai	Chiang Mai	Single	Bachelor	Police	35,000
Female	40	Lamphun	Lamphun	Married	Master	Accountant	40,000
Female	32	Lampang	Chiang Mai	Married	Secondary	Farmer	20,000
Male	45	Chiang Rai	Chiang Rai	Divorced	Master	Pilot	200,000

* Example data from parent expectation survey

Example of Encoding and Rescaling

Field Name	Encoding/Rescaling	Example
Gender	Can be encoded	F = Female M = Male
Age	-	
Birth Province	Can be encoded Use 01-77 to represent all provinces, sorted alphabetically	01 = Amnat Charoen 02 = Ang Thong 03 = Bueng Kan ....
Current Province	Can be encoded Use 01-77 to represent all provinces, sorted alphabetically	01 = Amnat Charoen 02 = Ang Thong 03 = Bueng Kan ....

ภาควิชาวิทยาการคอมพิวเตอร์
COMPUTER SCIENCE DEPARTMENT, CMU
คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

Example of Encoding and Rescaling (cont.)

Field Name	Encoding/Rescaling	Example
Marital Status	Can be encoded	1 = Single 2 = Married 3 = Divorced/ Separated/ Widowed
Educational Level	Can be encoded	1 = Below Elementary 2 = Elementary 3 = Secondary 4 = Associated Degree 5 = Bachelor Degree 6 = Master Degree 7 = Doctoral Degree

Example of Encoding and Rescaling (cont.)

Field Name	Encoding/Rescaling	Example
Occupation	For this example, we use open-ended question for occupation, making encoding impossible. However, if we limited number of possible answers, we can use encoding here.	
Monthly Income	Can be rescaled	Use 1,000 Baht as unit Dividing real value by 1,000

Example of Encoded and Rescaled Data

Gender	Age	Birth Province	Current Province	Marital Status	Education	Occupation	Monthly Income (1,000 Baht)
M	35	14	14	1	5	Police	35
F	40	54	54	2	6	Accountant	40
F	32	53	14	2	3	Farmer	20
M	45	13	13	3	6	Pilot	200

Encoded

Rescaled

ภาควิชาวิทยาการคอมพิวเตอร์
COMPUTER SCIENCE DEPARTMENT, CMU
คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

4.1.4 Editing

❑ *The process of making sure the data are accurate, so that data in storage can be used without any error down the line. If a mistake is found, they must be corrected*

In editing, we perform:

❖ *Verification ─ is the data entered the same as data collected?*

❖ *Validation ─ does the data make sense?*

  ○ *Range Check: Is the field value in reasonable range?*

  ○ *Relation Check: Does the value make sense, considering the other related field values?*

❖ *Correction ─ fix the mistake*

4.1.5 Recording

❑ **_Recording_ _is the preparation of data in a way that they can be read and processed by computers. Data is then will be stored on a computer, either on your PC, or a server._**

**Examples:**

➢ **Prepare data and store them in a database system**

4.2 Processing

Data Processing is where we perform certain computation on *raw data* to obtain some *information*

There are many data processing techniques, such as:

- ❖ *Calculating* certain numerical values of the data
- ❖ *Updating* some part of data
- ❖ *Sorting* the data in certain order
- ❖ *Searching* for some item that we want
- ❖ *Classification* of data, showing a group we want to see
- ❖ *Summarizing* obtained information to show some insights into data

4.2 Processing (cont.)

Calculating *is the process of using mathematical methods, such as addition, taking average, or finding standard deviation, on the data.*

Examples:

- ➢ *Finding average ages of the users*
- ➢ *Finding standard deviation of the test scores*

Updating *is changing the values of the data to a more up-to-date values.*

Examples:

- ➢ *Increasing everyone's salary numbers by 10%*
- ➢ *Changing bank account balance after a deposit/withdrawal*

4.2 Processing (cont.)

Sorting *is arranging items in the data in a particular order, for ease of searching and management*

Examples:

- ➢ **Sorting employees by names**
- ➢ **Sorting students by their test scores**

Searching *is finding an item (or more) in the data that if a criteria*

Examples:

- ➢ **Searching for students with lower-than-average test scores**
- ➢ **Searching for employees in the accounting department and has salary over 50,000 THB**

4.2 Processing (cont.)

Classification **is grouping of data items by certain characteristic, which can be a result of**

another data processing method

Examples:

➢ **Displaying athletes' data, grouped by types of sports**

➢ **Grading (grouping students' data by their scores)**

Summarizing **is the act of using information we have gained from various data**

processing methods to show a description of the data as a whole

Examples:

➢ **Grades summary for a course**

➢ **Test scores summary, divided by sections, with maximum and minimum scores,**

mean, and standard deviation

4.3 Results Management and Presentation

Result management is the storage and maintenance of gained information. This include making backups for future use
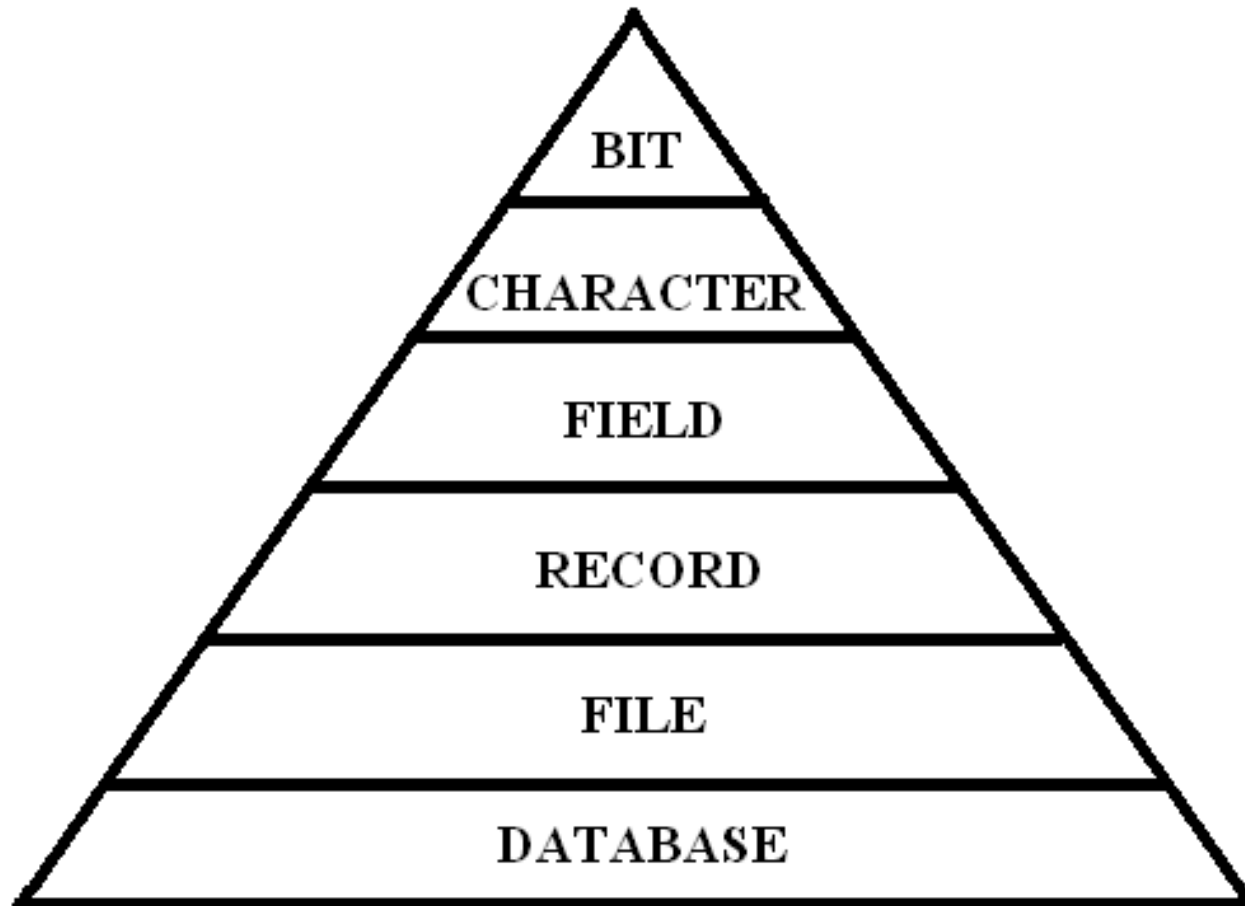
Presentation is taking gained information and display them to those involved (clients, executives, etc.) This can be tables of figures, or charts, etc. This also includes how to disseminate information as quickly as possible. So that even distance user can get the information in a timely manner

5. Data Organization

Data Organization

Data Organization (cont.)

❑ **Bit** *is the smallest unit of data that is understandable by a computer. A bit is a binary (base two) digit, with value of either 0 or 1*

❑ **Character** *is a symbol of data, which can be a digit (0-9) ,an alphabetical character (A-Z, a-z), or even a special symbol (? ! $). It takes 1 byte (8 bits) to store a character*

Examples:

➤ *A is represented by 0100 0001 (equals 65 in decimal)*

➤ *B is represented by 0100 0010 (equals 66 in decimal)*

➤ *C is represented by 0100 0011 (equals 67 in decimal)*

➤ *D is represented by 0100 0100 (equals 68 in decimal)*

Data Organization (cont.)

❑ *Field* **consists of one more characters. Field will have a meaning**

Examples:

A name field with the value of Kanowaan

An age field with the value of 19

An grade field with the value of A

❑ *Record* **is a collection of fields combining, representing an item of data.**

Examples: data about one parent, one student, or one piece of merchandize
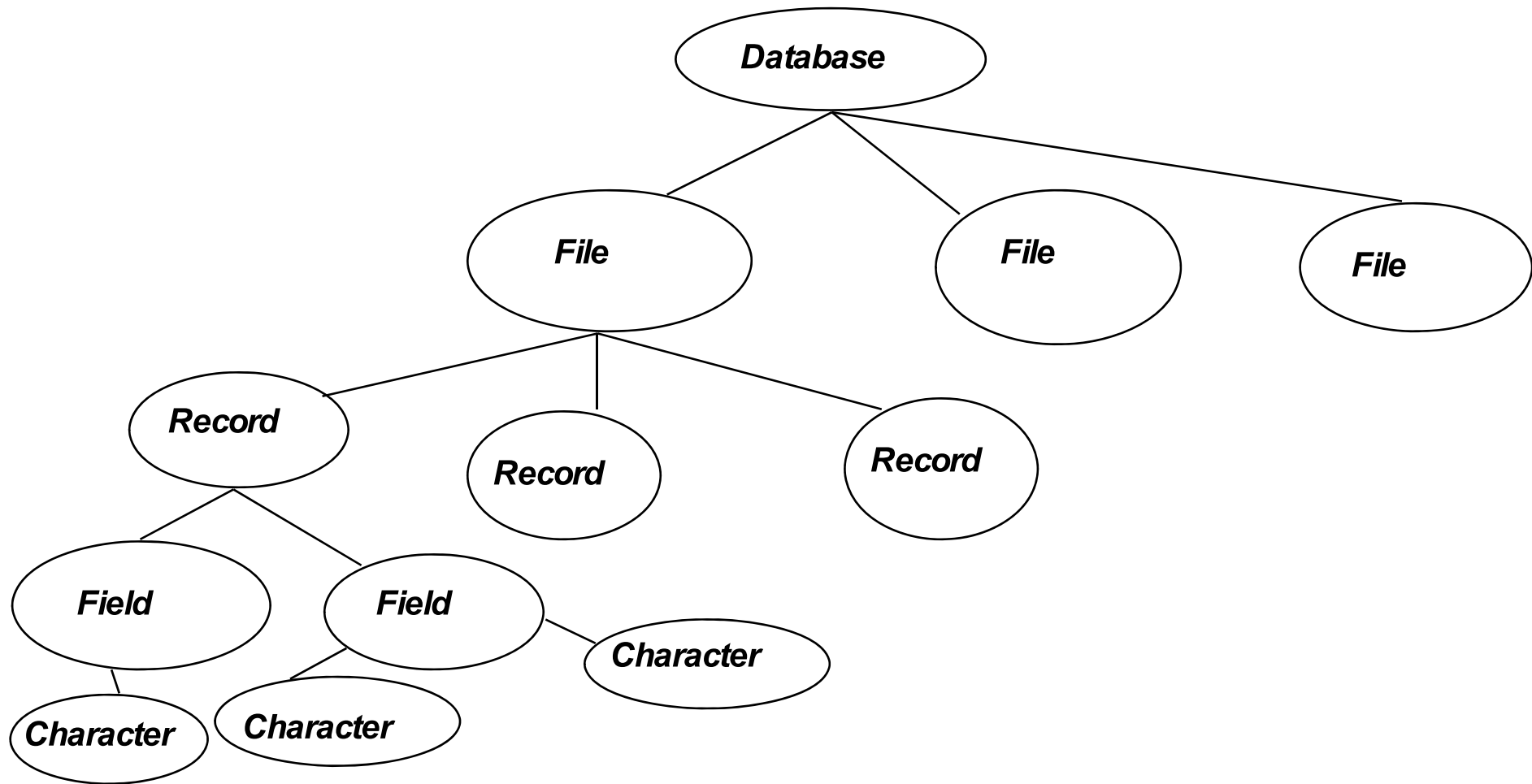
Data Organization (cont.)

❑ *File is a collection of records of the same type. For example, a student file will consists of multiple records, each representing a student. Sometime, a file is called a* table*.*

❑ *Database is a collection of related files. For example, a student database can consists of a student personal file, an advisor file, a major file, etc.*

ภาควิชาวิทยาการคอมพิวเตอร์
COMPUTER SCIENCE DEPARTMENT, CMU
คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

Database Components Diagram

Examples of File, Record, and Field

Field

Gender	Age	Birth Province	Current Province	Marital Status	Education	Occupation	Monthly Income (1,000 Baht)
M	35	14	14	1	5	Police	35
F	40	54	54	2	6	Accountant	40
F	32	53	14	2	3	Farmer	20
M	45	13	13	3	6	Pilot	200

Record

File

This example has 4 records and 8 fields

6. Example Databases

Example 1

Data from Parent Expectation Survey for

Computer Science Graduate

Parent File

Field Name	Description	Type	Example
Id	ID for parents	Char(4)	0001
Sex	Gender of the parent	Char(1)	M
Age	Age of the parent	Int	35
Province1	Birth Province	Char(2)	14
Province2	Currently-living Province	Char(2)	14
Marital_Status	Marital status	Char(1)	1
Education	Educational level	Char(1)	5
Occupation	Occupation	Varchar(20)	Police
Income	Monthly Income (1,000 THB)	Int	35

ภาควิชาวิทยาการคอมพิวเตอร์
COMPUTER SCIENCE DEPARTMENT, CMU
คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

Parent File (cont.)

Field Name	Description	Type	Example
Answer1	Answer for question 1	Char(1)	3
Answer2	Answer for question 2	Char(1)	3
….			
Answer12	Answer for question 12	Char(1)	4

Parent File, with Data

Id	Sex	Age	Province1	Province2	Marital_Status	Education	Occupation	Income	Answer1	Answer2	....	Answer12
0001	M	35	14	14	1	5	Police	35	3	3		4
0002	F	40	54	54	2	6	Accountant	40	4	4		4
0003	F	32	53	14	2	3	Farmer	20	3	4		5
0004	M	45	13	13	3	6	Pilot	200	2	2		3
…….												

ภาควิชาวิทยาการคอมพิวเตอร์
COMPUTER SCIENCE DEPARTMENT, CMU
คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

Example 2

Student Database

Example of Student Database

Student
Student_Id
Firstname
Lastname
Address
…….
Advisor_Id
Major_Id

Advisor
Advisor_Id
Firstname
Lastname
Office_Room
…….

Major
Major_Id
Major_Name
Description
…….
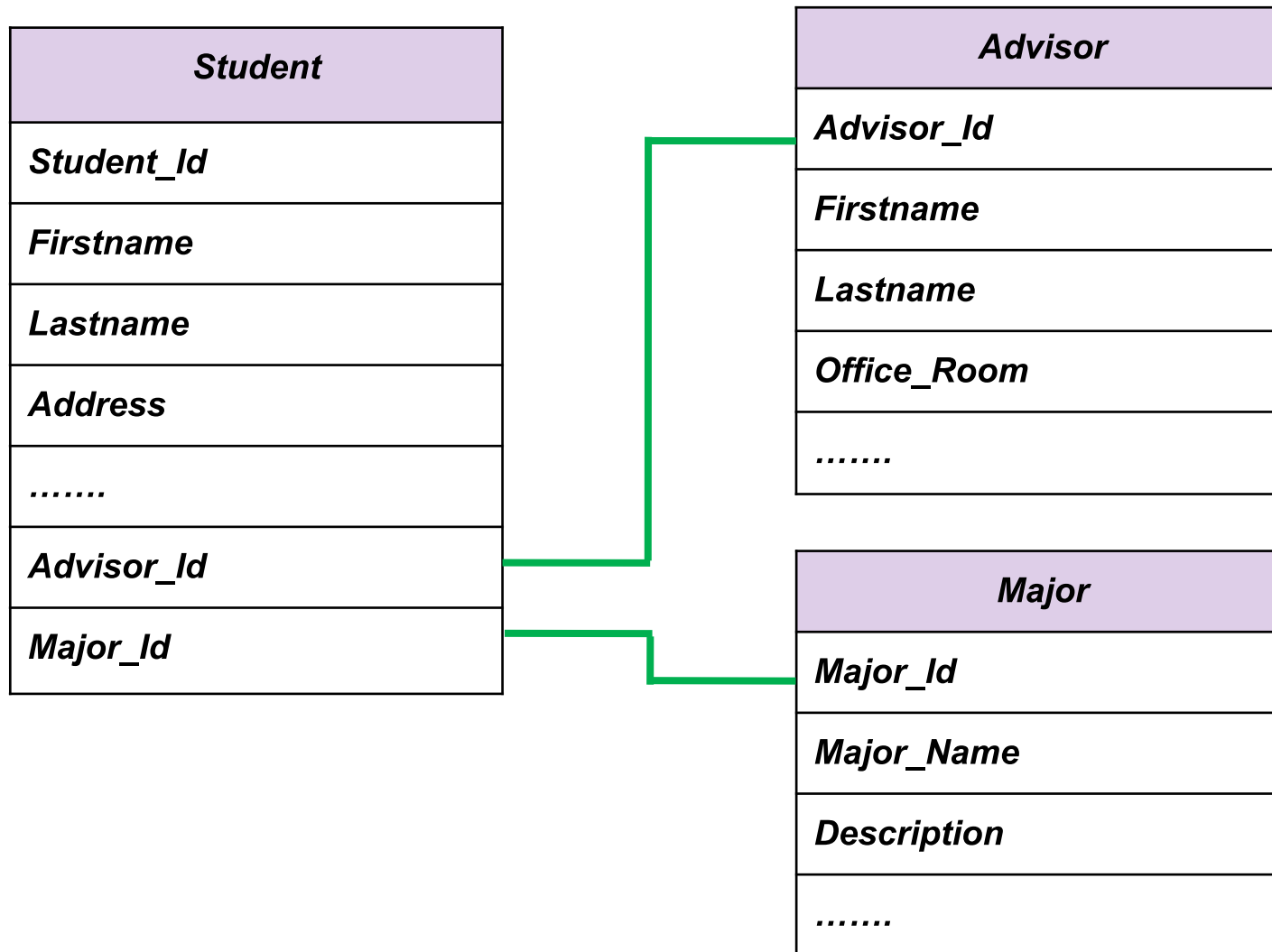
Student File

Stuent_Id	Firstname	Lastname	Address	…	Advisor_Id	Major_Id
600510001	Piyachat	Chaidee	214 Moo 1 T. Suthep Chiang Mai		0001	001
600510002	Kanokwaan	Oonchoi	176 Moo 5 T. Changpuek Chiang Mai		0001	001
600510111	Wuttisak	Pakdeenopparattr	123 Moo 7 T. Lampangluang A.Kohkhaa Lampang		0012	004
600510112	Prapha	Niyomchart	178 Moo 2 T.Koh A.Li Lamphun		0012	004
600510215	Tinapob	Wisetniyom	4 Moo 4 T.Vieng Chiang Rai		0027	011
……						

Advisor File

Advisor_Id	Firstname	Lastname	Office_Room	……
0001	Matinee	Kiewkanya	CSB 101	
0012	Prapaa	Wuttisakkriengkrai	BB 201	
0027	Wattana	Prasobsook	GB 225	
……				

* Only a few records are shown

Major File

Major_Id	Major_Name	Description	……
001	Computer Science	Studies of computational Theories…	
004	Biology	Rational studies related to all living things…	
011	Gemology	Studies of gemstone, including natural origins…	
……			

* Only a few records are shown