CHAPTER 8

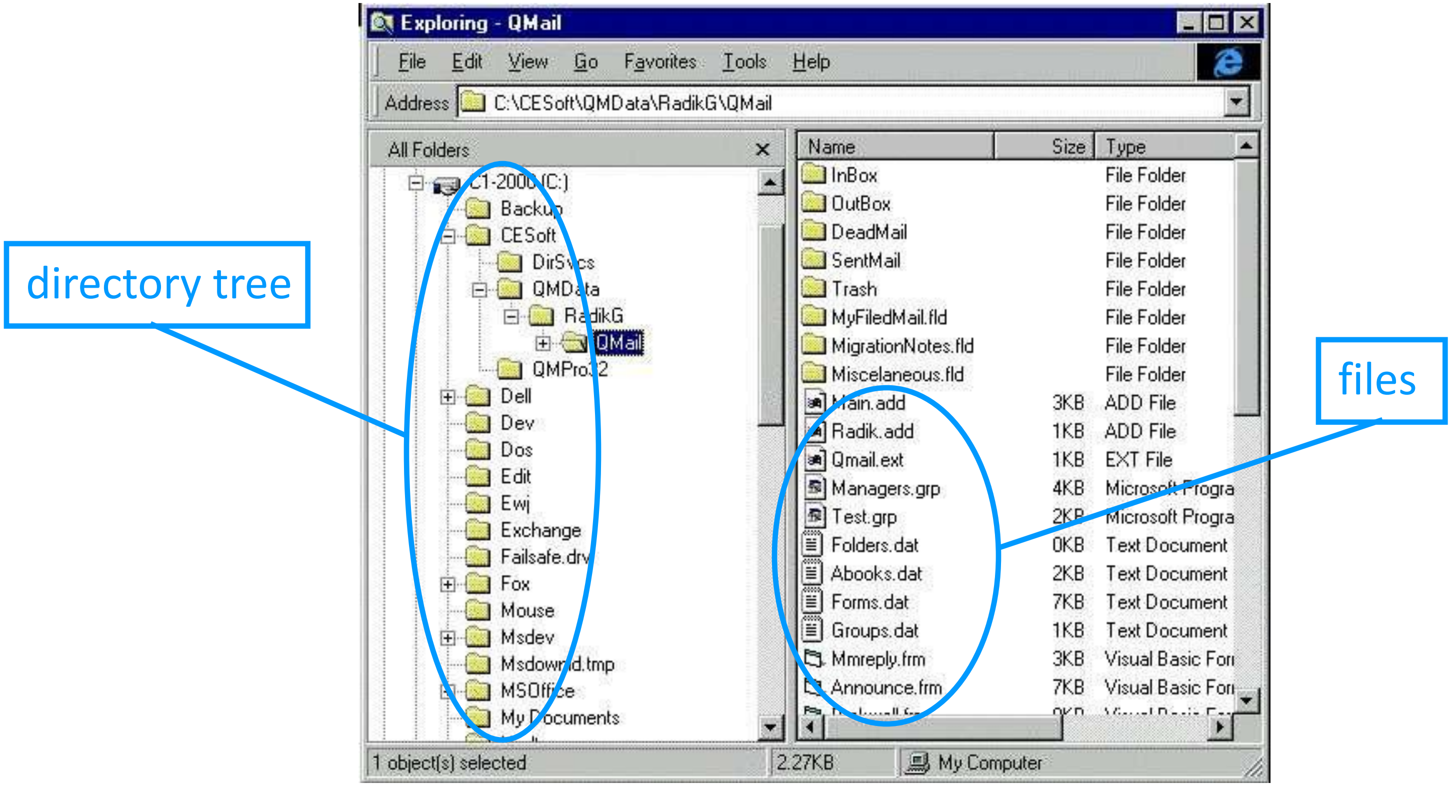
File Structures

Reference: Computer Science an Overview
Author: J. Glenn Brook Shear
6th Edition

- Abstractions of the actual data organization on *mass storage*
- Again: differences between *conceptual* and *actual* data organization

8.1: Files, Directories & the Operating System

- OS storage structure:
  - conceptual hierarchy of *directories* and *files*



directory tree

files

Address C:\CESoft\QMData\RadikG\QMail

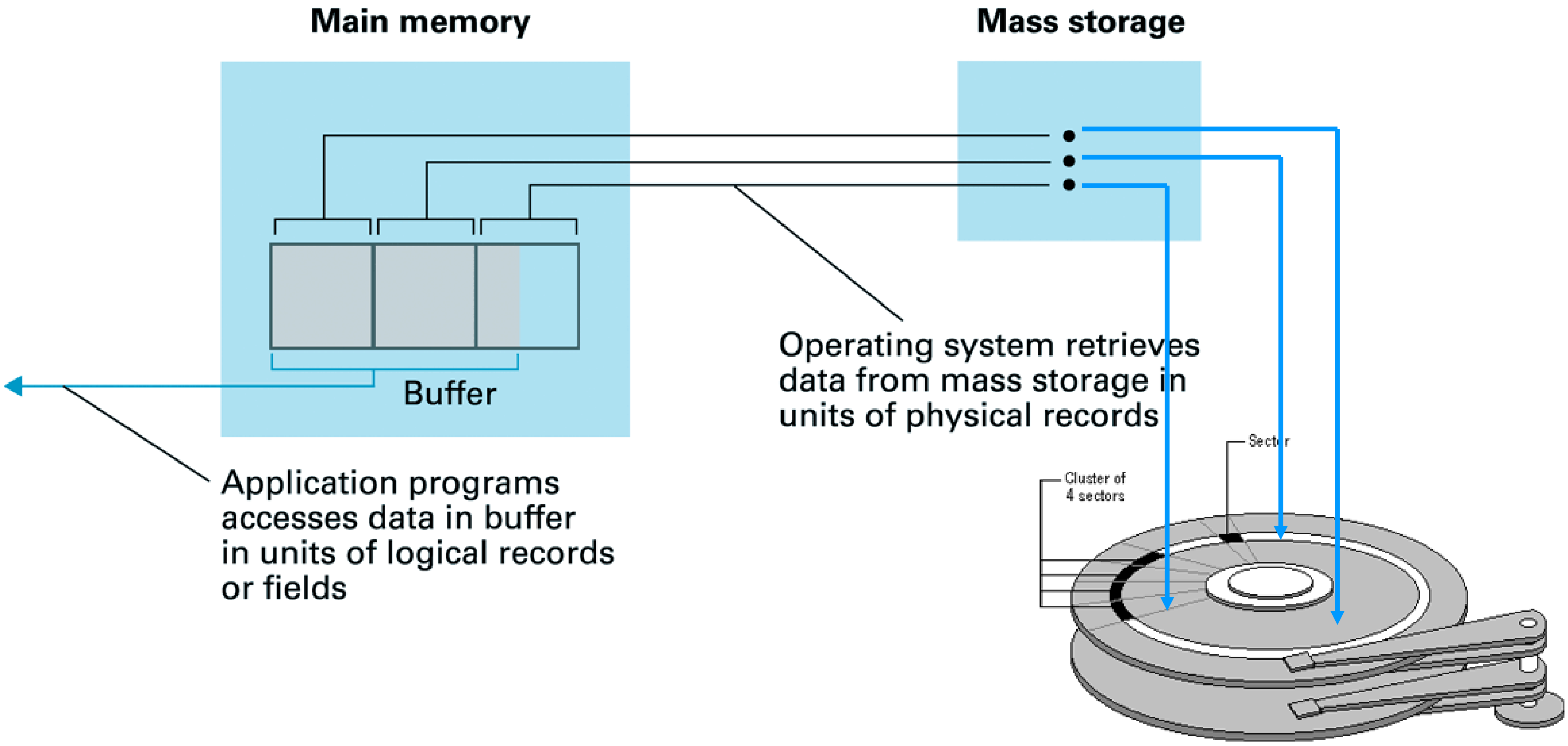
```
C1-2000 (C:)
    Backup
    CESoft
        DirSvcs
        QMData
            RadikG
                QMail
        QMPro32
    Dell
    Dev
    Dos
    Edit
    Ewj
    Exchange
    Failsafe.drv
    Fox
    Mouse
    Msdev
    Msdownld.tmp
    MSOffice
    My Documents
```

Name	Size	Type
InBox		File Folder
OutBox		File Folder
DeadMail		File Folder
SentMail		File Folder
Trash		File Folder
MyFiledMail.fld		File Folder
MigrationNotes.fld		File Folder
Miscelaneous.fld		File Folder
Main.add	3KB	ADD File
Radik.add	1KB	ADD File
Qmail.ext	1KB	EXT File
Managers.grp	4KB	Microsoft Progra
Test.grp	2KB	Microsoft Progra
Folders.dat	0KB	Text Document
Abooks.dat	2KB	Text Document
Forms.dat	7KB	Text Document
Groups.dat	1KB	Text Document
Mmreply.frm	3KB	Visual Basic For
Announce.frm	7KB	Visual Basic For

1 object(s) selected | 2.27KB | My Computer

8.1: Files: Conceptual vs. Actual View

- View at OS-level is conceptual
  - actual storage may differ significantly!

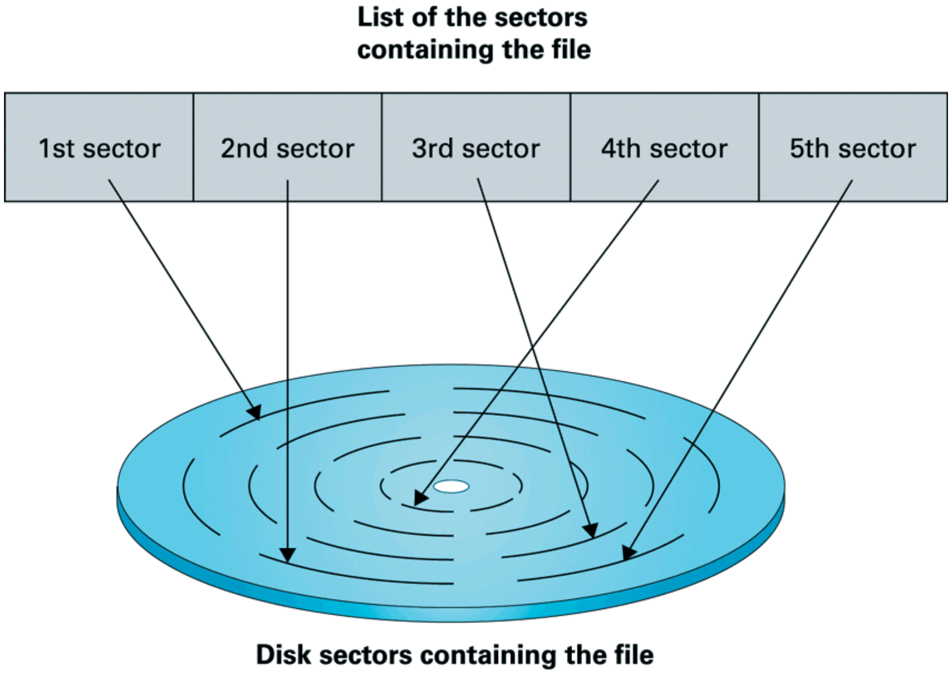


Operating system retrieves data from mass storage in units of physical records

Application programs accesses data in buffer in units of logical records or fields

8.2: Sequential Files

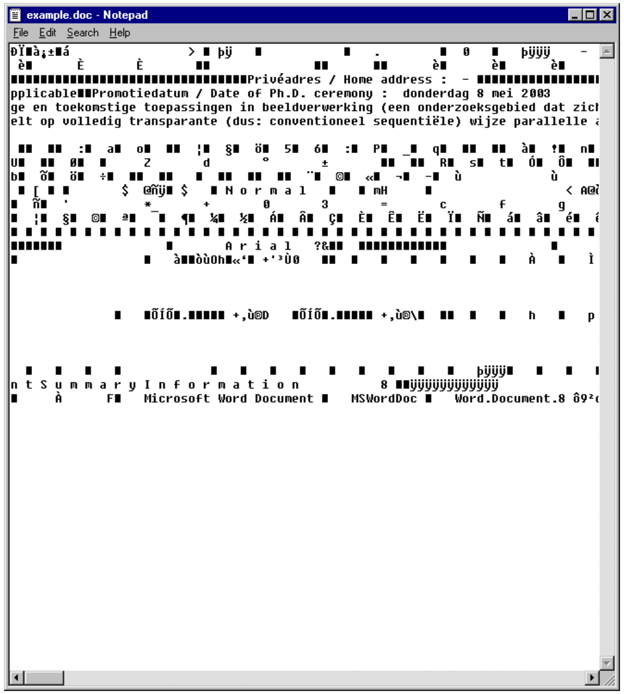
- To ‘remember’ where data resides on disk, the OS maintains a list of sectors for each file


List of the sectors containing the file
1st sector
2nd sector
3rd sector
4th sector
5th sector
Disk sectors containing the file

- Result: *sequential view* of scattered set of data

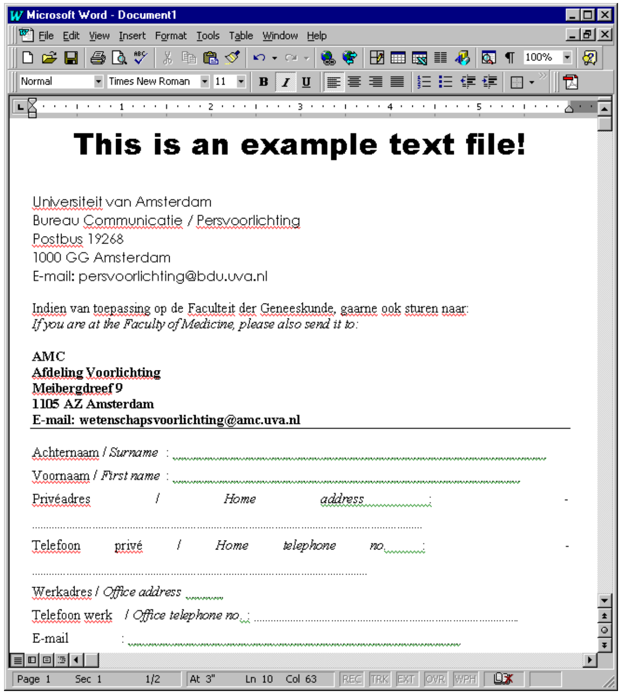
8.2: Text Files

- Sequential file consisting of long string of encoded characters (e.g. ASCII-code)
  - But: character-string still interpreted by word processor!



```
ÐÏ■à¡±■á                 > ■ þÿ   ■            ■   .        ■  0   ■   þÿÿÿ    -
 è■       È       È       ■■                ■■      ■■       è■       è■       è■
■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■Privéadres / Home address :  - ■■■■■■■■■■■■■■■■■
pplicable■■Promotiedatum / Date of Ph.D. ceremony :  donderdag 8 mei 2003
ge en toekomstige toepassingen in beeldverwerking (een onderzoeksgebied dat zich
elt op volledig transparante (dus: conventioneel sequentiële) wijze parallelle a

 ■■  ■■  :■  a■  o■  ■■  ¦■  §■  ö■  5■  6■  :■  P■  _■  q■  ■■  ■■  à■  !■  n■
U■  ■■  Ø■  ■     Z       d       °       ±       ■■  ■■  R■  s■  t■  Ó■  Ô■  ■■
b■  õ■  ö■  ÷■  ■■  ■■   ■  ■■  ■■  ■■  ¨■  ©■  «■  ¬■  -■  ù           ù
 ■ [ ■ ■       $  @ñÿ■ $   ■ N o r m a l   ■    ■ mH     ■                  < A@à
■  ñ■  '        *_      +       0       3       =       c       f       g
■  ¦■  §■  ©■  ª■  ¯■  ¶■  ¼■  ½■  Á■  Â■  Ç■  È■  Ê■  Ë■  Ï■  Ñ■  á■  â■  é■  é
■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■
■■■■■■■           ■       A r i a l   ?&■■  ■■■■■■■■■■■■               ■
■                 ■   à■■òùOh■«'■ +'³Ù0   ■■  ■   ■   ■   ■   ■   ■   À   ■   Ì



             ■   ■ÕÍÕ■.■■■■■ +,ù®D   ■ÕÍÕ■.■■■■■ +,ù®\■  ■■  ■   ■   h   ■   p



 ■   ■   ■   ■            ■   ■   ■   ■   ■   ■   ■   ■   ■   þÿÿÿ■   ■   ■   ■
n t S u m m a r y I n f o r m a t i o n          8 ■■ÿÿÿÿÿÿÿÿÿÿÿÿ
■     À      F■   Microsoft Word Document ■   MSWordDoc ■   Word.Document.8 ô9²q
```

File in "Notepad"


Microsoft Word - Document1
File Edit View Insert Format Tools Table Window Help
100%
Normal
Times New Roman
11
B I U
Page 1 Sec 1 1/2 At 3" Ln 10 Col 63 REC TRK EXT OVR WPH

This is an example text file!

Universiteit van Amsterdam
Bureau Communicatie / Persvoorlichting
Postbus 19268
1000 GG Amsterdam
E-mail: persvoorlichting@bdu.uva.nl

Indien van toepassing op de Faculteit der Geneeskunde, gaarne ook sturen naar:
If you are at the Faculty of Medicine, please also send it to:

AMC
Afdeling Voorlichting
Meibergdreef 9
1105 AZ Amsterdam
E-mail: wetenschapsvoorlichting@amc.uva.nl

Achternaam / *Surname* : ....

Voornaam / *First name* : ....

Privéadres / *Home address* : -
....

Telefoon privé / *Home telephone no.* : -
....

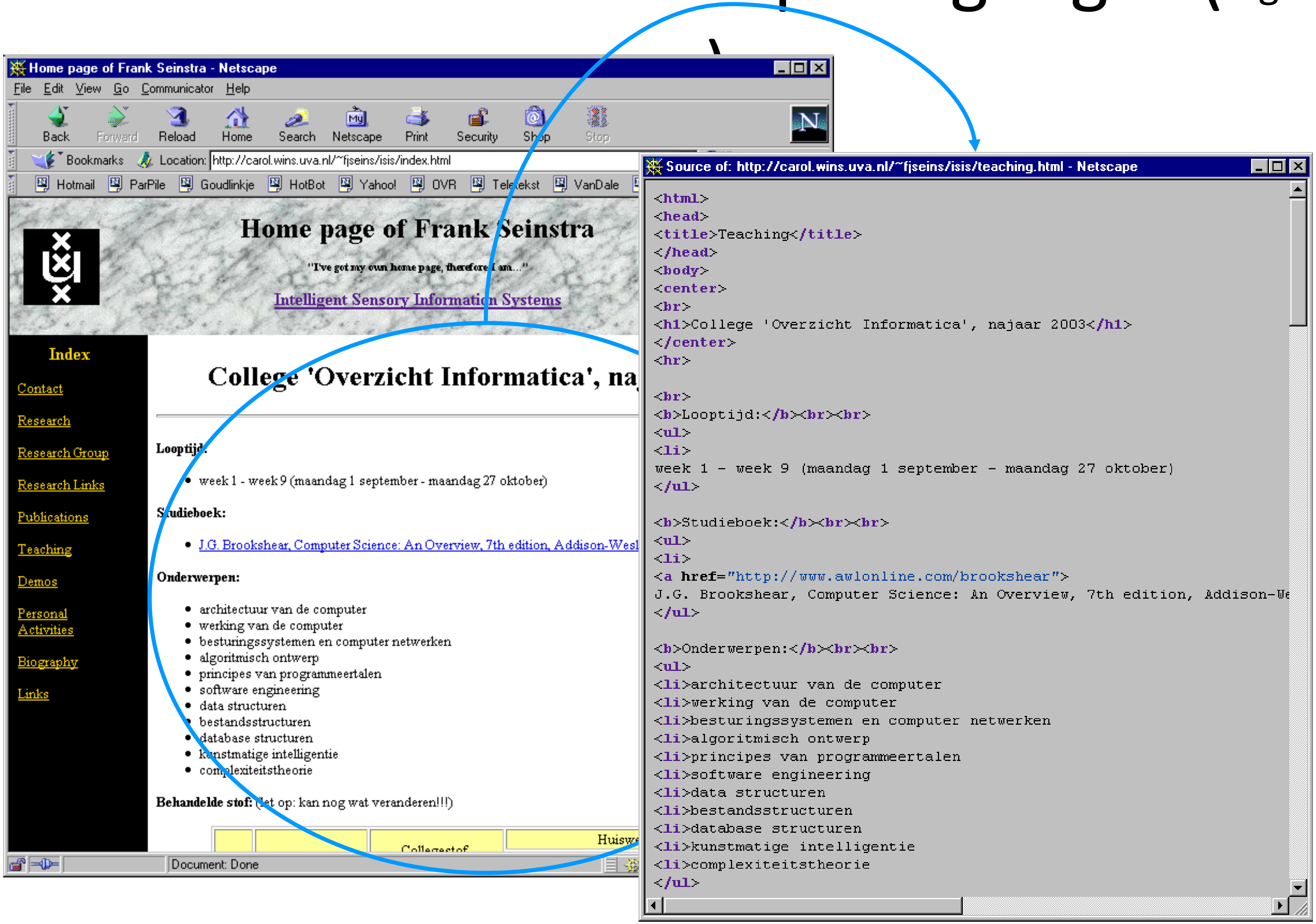
Werkadres / *Office address* ....

Telefoon werk / *Office telephone no.* : ....

E-mail : ....

Same file in "MS Word"

8.2: Text files & Markup Languages (e.g.)



Home page of Frank Seinstra - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://carol.wins.uva.nl/~fjseins/isis/index.html
Hotmail ParPile Goudlinkje HotBot Yahoo! OVR Teletekst VanDale
Document: Done
Source of: http://carol.wins.uva.nl/~fjseins/isis/teaching.html - Netscape

Home page of Frank Seinstra

"I've got my own home page, therefore I am..."

Intelligent Sensory Information Systems

Index

Contact

Research

Research Group

Research Links

Publications

Teaching

Demos

Personal Activities

Biography

Links

College 'Overzicht Informatica', na

Looptijd:

- week 1 - week 9 (maandag 1 september - maandag 27 oktober)

Studieboek:

- J.G. Brookshear, Computer Science: An Overview, 7th edition, Addison-Wesl

Onderwerpen:

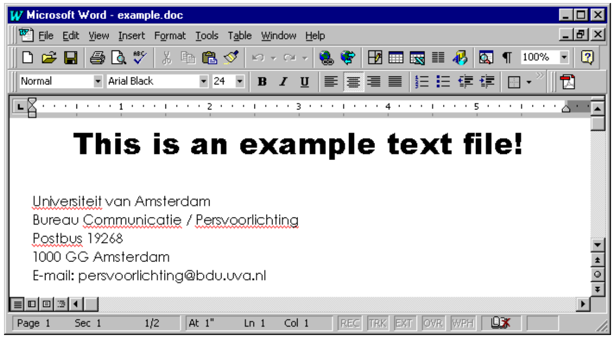
- architectuur van de computer
- werking van de computer
- besturingssystemen en computer netwerken
- algoritmisch ontwerp
- principes van programmeertalen
- software engineering
- data structuren
- bestandsstructuren
- database structuren
- kunstmatige intelligentie
- complexiteitstheorie

Behandelde stof: (let op: kan nog wat veranderen!!!)

		Collegestof	Huisw

```
<html>
<head>
<title>Teaching</title>
</head>
<body>
<center>
<br>
<h1>College 'Overzicht Informatica', najaar 2003</h1>
</center>
<hr>

<br>
<b>Looptijd:</b><br><br>
<ul>
<li>
week 1 - week 9 (maandag 1 september - maandag 27 oktober)
</ul>

<b>Studieboek:</b><br><br>
<ul>
<li>
<a href="http://www.awlonline.com/brookshear">
J.G. Brookshear, Computer Science: An Overview, 7th edition, Addison-We
</ul>

<b>Onderwerpen:</b><br><br>
<ul>
<li>architectuur van de computer
<li>werking van de computer
<li>besturingssystemen en computer netwerken
<li>algoritmisch ontwerp
<li>principes van programmeertalen
<li>software engineering
<li>data structuren
<li>bestandsstructuren
<li>database structuren
<li>kunstmatige intelligentie
<li>complexiteitstheorie
</ul>
```

8.2: From actual storage to conceptual


Microsoft Word - example.doc
File Edit View Insert Format Tools Table Window Help
Normal
Arial Black
24
100%
Page 1 Sec 1 1/2 At 1" Ln 1 Col 1 REC TRK EXT OVR WPH

This is an example text file!

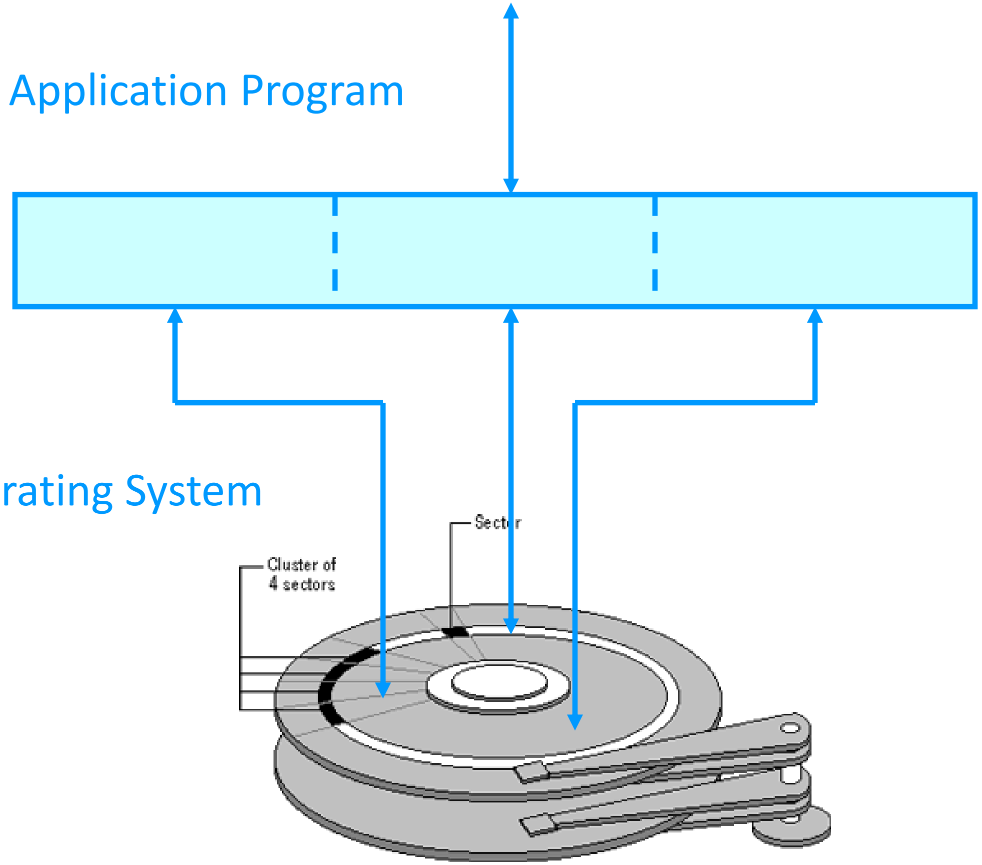
Universiteit van Amsterdam
Bureau Communicatie / Persvoorlichting
Postbus 19268
1000 GG Amsterdam
E-mail: persvoorlichting@bdu.uva.nl

conceptual view

Interpretation by Application Program

Sequential buffer

sequential view

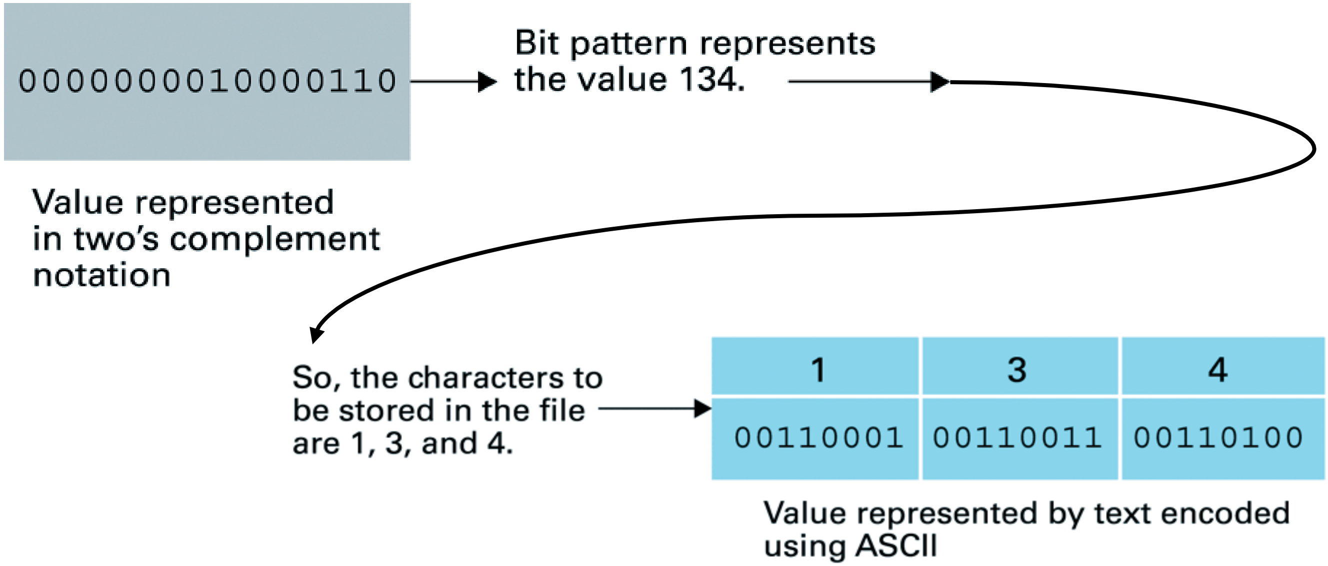


Assembly by Operating System

actual storage

8.2: Data Conversion

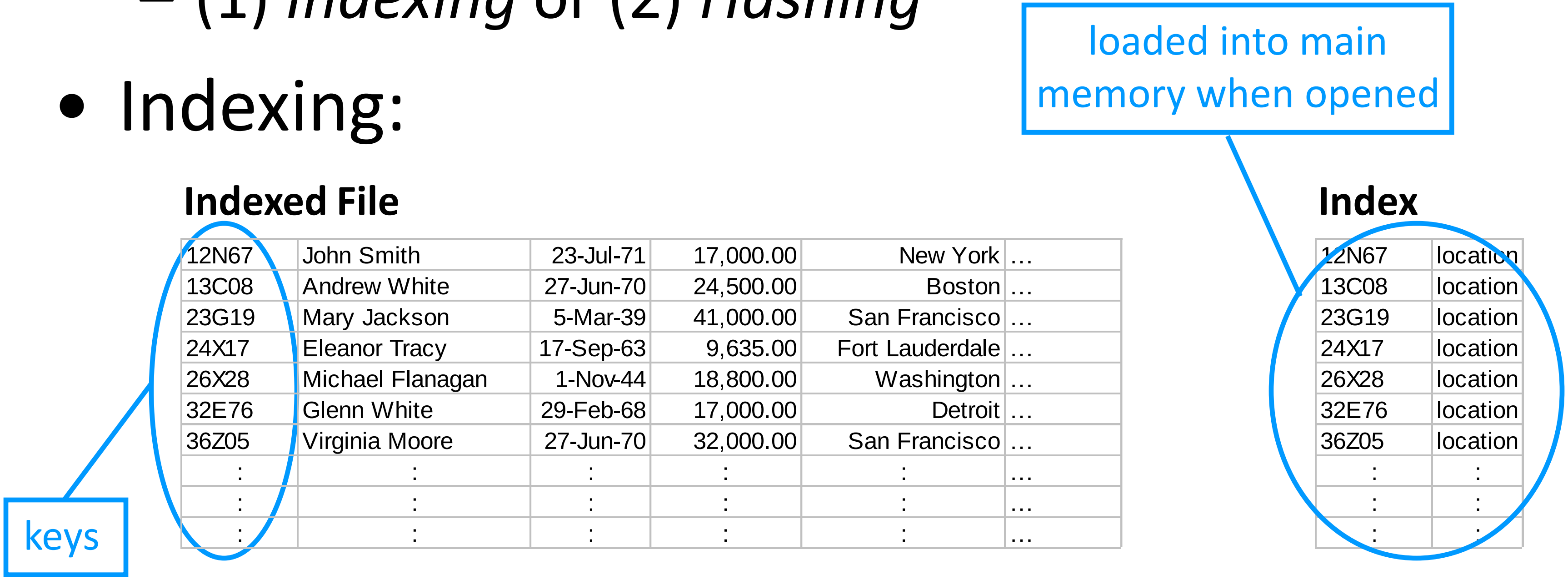
- When programming: note that data transfer to/from file may involve data conversion:
  - e.g., from two's complement notation to ASCII:


0000000010000110
Bit pattern represents the value 134.
Value represented in two's complement notation
So, the characters to be stored in the file are 1, 3, and 4.
1
3
4
00110001
00110011
00110100
Value represented by text encoded using ASCII

- So: again it's about the *interpretation* of data

8.3: Quick File Access

- Disadvantage of sequential files:
  - no quick access to particular file data
- Two techniques to overcome this problem:
  - (1) *Indexing* or (2) *Hashing*
- Indexing:



loaded into main memory when opened

Indexed File

12N67	John Smith	23-Jul-71	17,000.00	New York	…
13C08	Andrew White	27-Jun-70	24,500.00	Boston	…
23G19	Mary Jackson	5-Mar-39	41,000.00	San Francisco	…
24X17	Eleanor Tracy	17-Sep-63	9,635.00	Fort Lauderdale	…
26X28	Michael Flanagan	1-Nov-44	18,800.00	Washington	…
32E76	Glenn White	29-Feb-68	17,000.00	Detroit	…
36Z05	Virginia Moore	27-Jun-70	32,000.00	San Francisco	…
:	:	:	:	:	…
:	:	:	:	:	…
:	:	:	:	:	…

Index

12N67	location
13C08	location
23G19	location
24X17	location
26X28	location
32E76	location
36Z05	location
:	:
:	:
:	:

keys

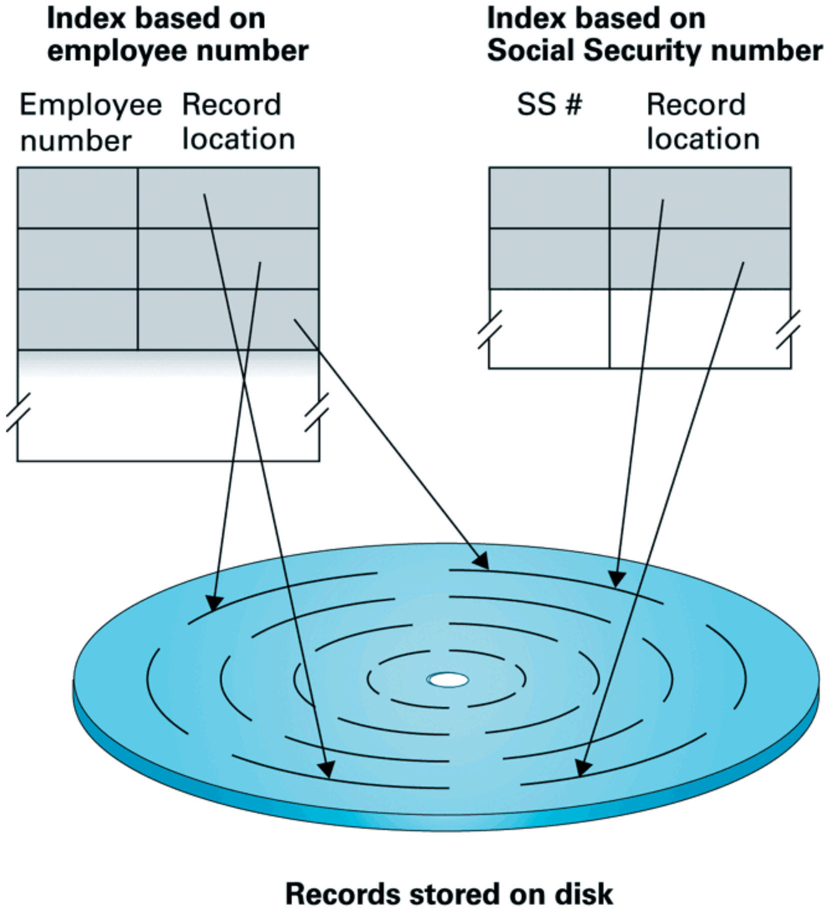
Chapter 8 - Problem 10

Why is a '*patient identification number*' a better choice for a key field than the last name of each patient?

- If key unique:
  - additional sequential search never required
- Patient's last name is not always unique

8.3: Inverted Files

- Variation to (single) indexing: inverted file

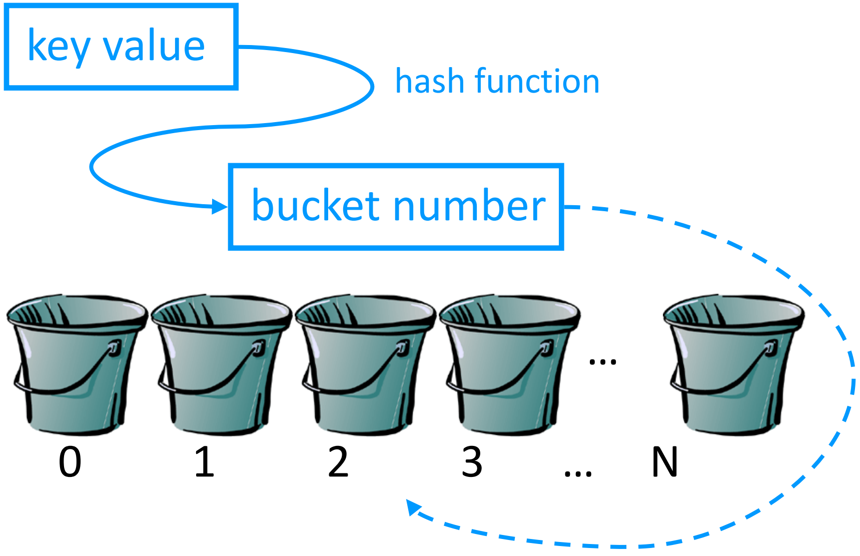


8.4: Hashing

- Disadvantage of indexing is… the index
  - requires extra space

- Solution: ‘*hashing*’
  - finds position in file using a key value (as in indexing)…
  - … simply by identifying location *directly from the key*

- How?
  - define set of ‘*buckets*’ & ‘*hash function*’ that converts keys to bucket numbers



key value

hash function

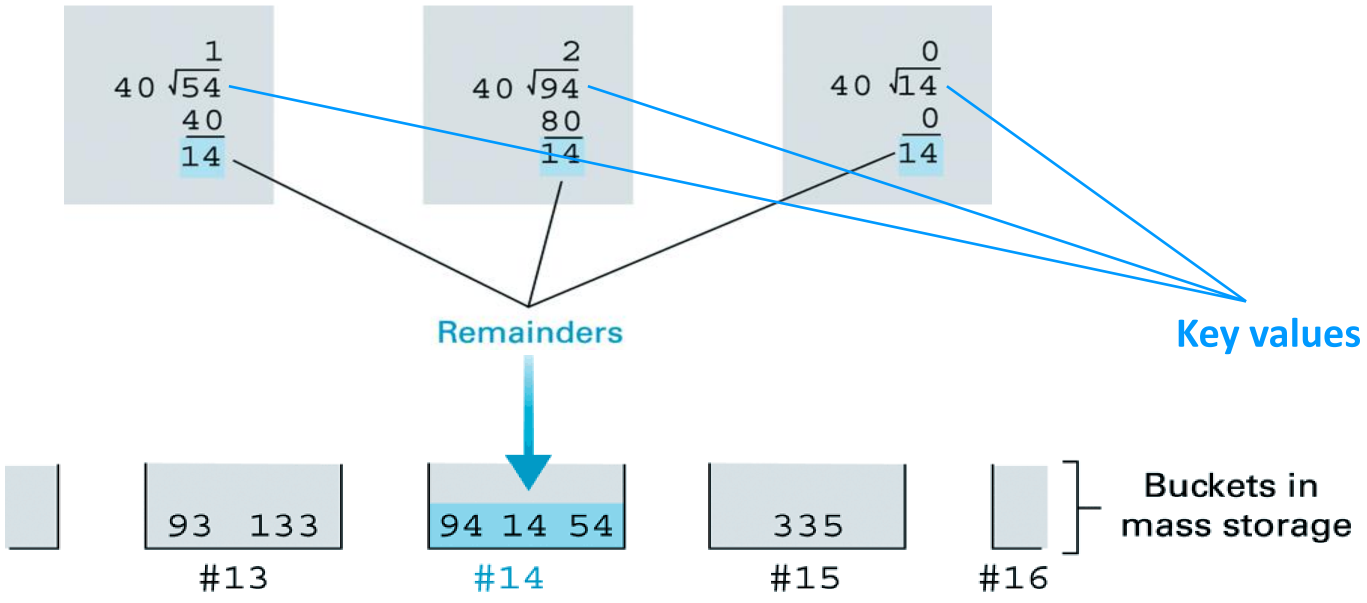
bucket number

…

0 1 2 3 … N

8.4: Hash Function: Example

- If storage space divided into *40 buckets* and hash function is *division*:
  - key values 14, 54, & 94 all map onto same bucket (collision)



Chapter 8 - File Structures: Conclusions

- File Structures:
  - abstractions of actual data organization on mass storage
- Changes of ‘view’:
  - actual storage -> sequential view by OS -> conceptual view presented to user
- Quick access to particular file data by
  - (1) indexing (many forms)
  - (2) hashing (requires no index, *but requires bucket search*!)

End of the Lecture

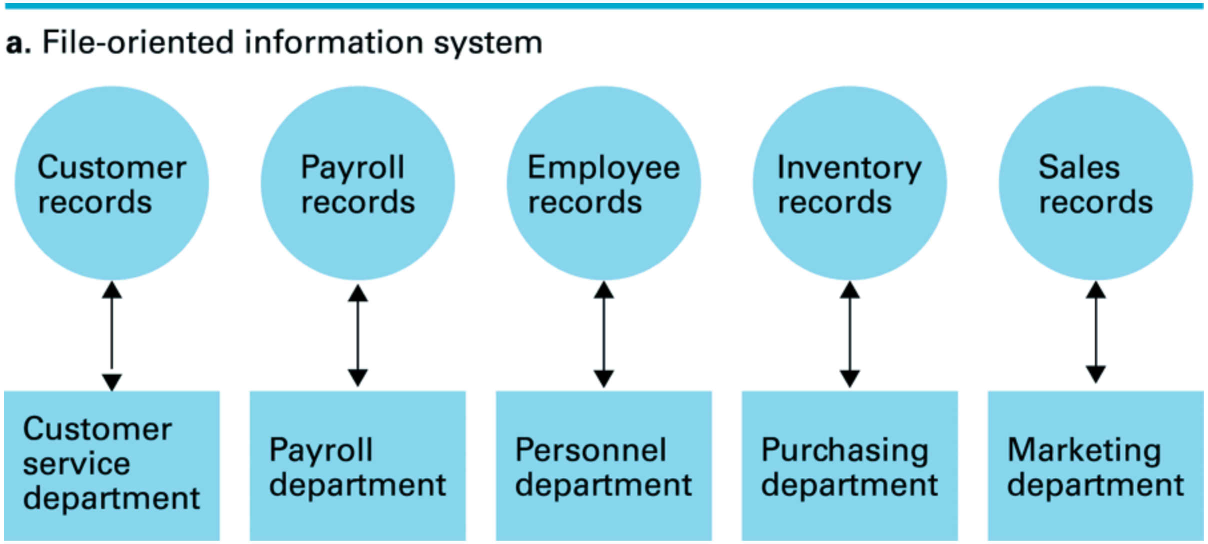
CHAPTER 9

Database Structures

- (Large) integrated collections of data that can be accessed quickly

9.1: Historical Perspective

- Originally: departments of large organizations stored all data separately in *flat files*



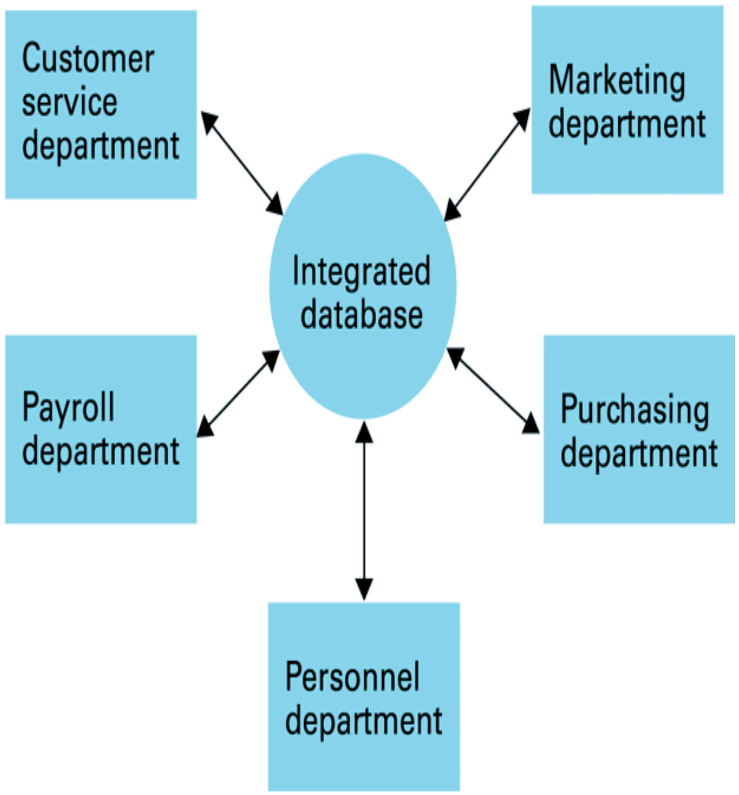
- Problems: redundancy & inconsistencies

9.1: Integrated Database System

- Better approach: integrate all data in a single system, to be accessed by all departments.
- Schema and Subschema

Example: University student and faculty records

b. Database-oriented information system



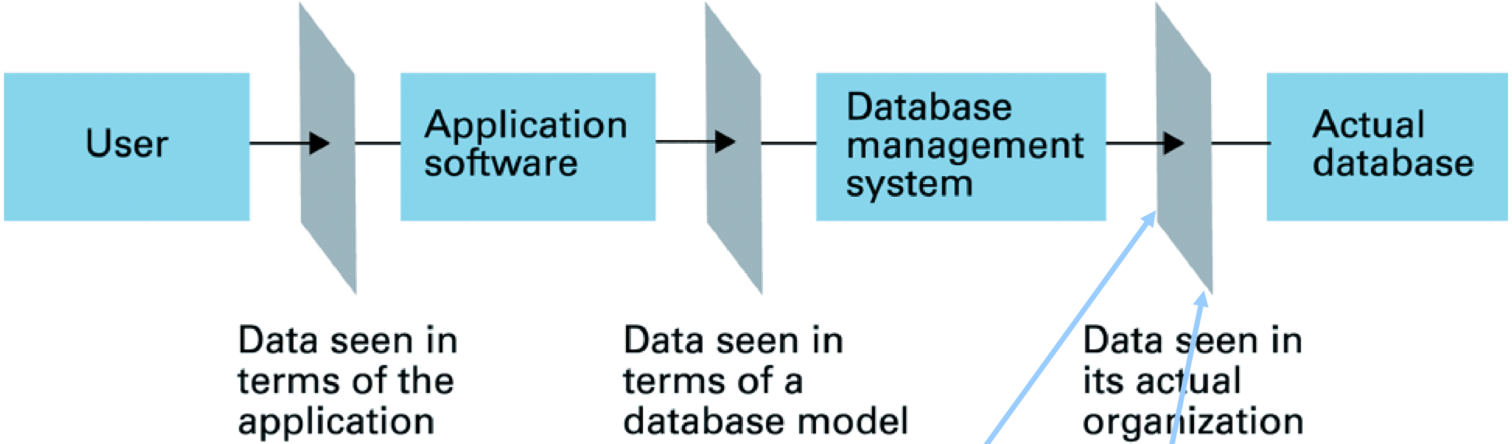
9.1: Disadvantages of Data Integration

- Control of access to sensitive data?

- Misinterpretation of integrated data?

- What about the **right** to hold/collect/interpret data?

9.2: Conceptual Database Layers



User

Application software

Database management system

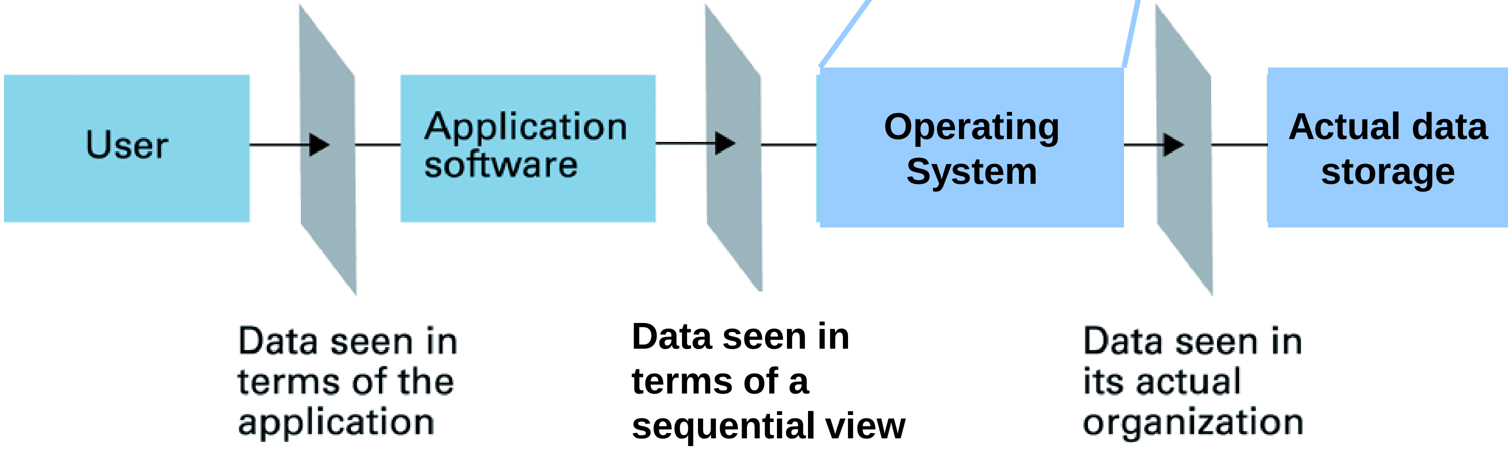
Actual database

Data seen in terms of the application

Data seen in terms of a database model

Data seen in its actual organization

- Compare:



User

Application software

Operating System

Actual data storage

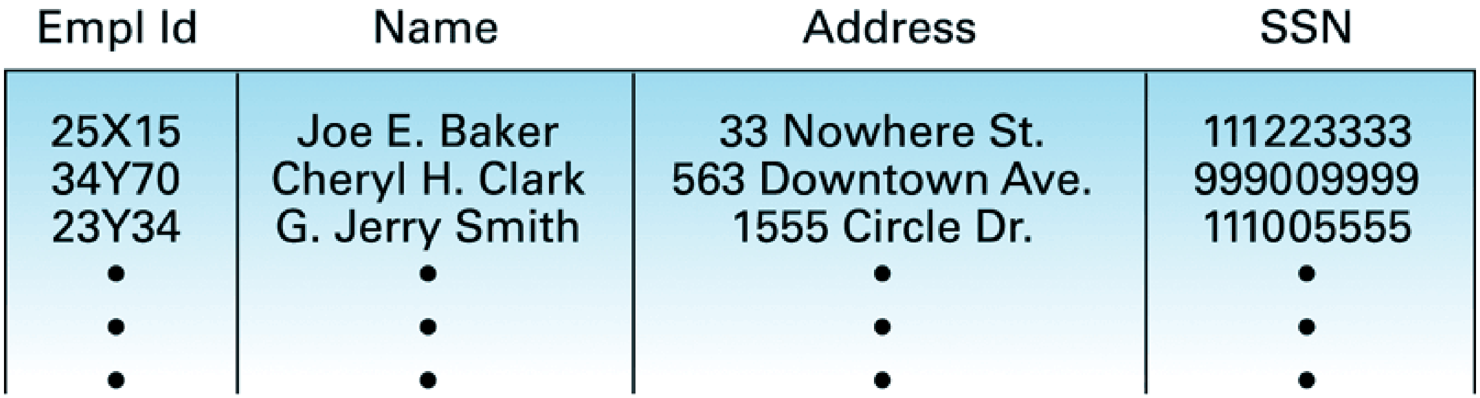
Data seen in terms of the application

Data seen in terms of a sequential view

Data seen in its actual organization

9.3: The Relational Model

- Relational Model
  - – shows data as being stored in rectangular tables, called *relations*, e.g.:

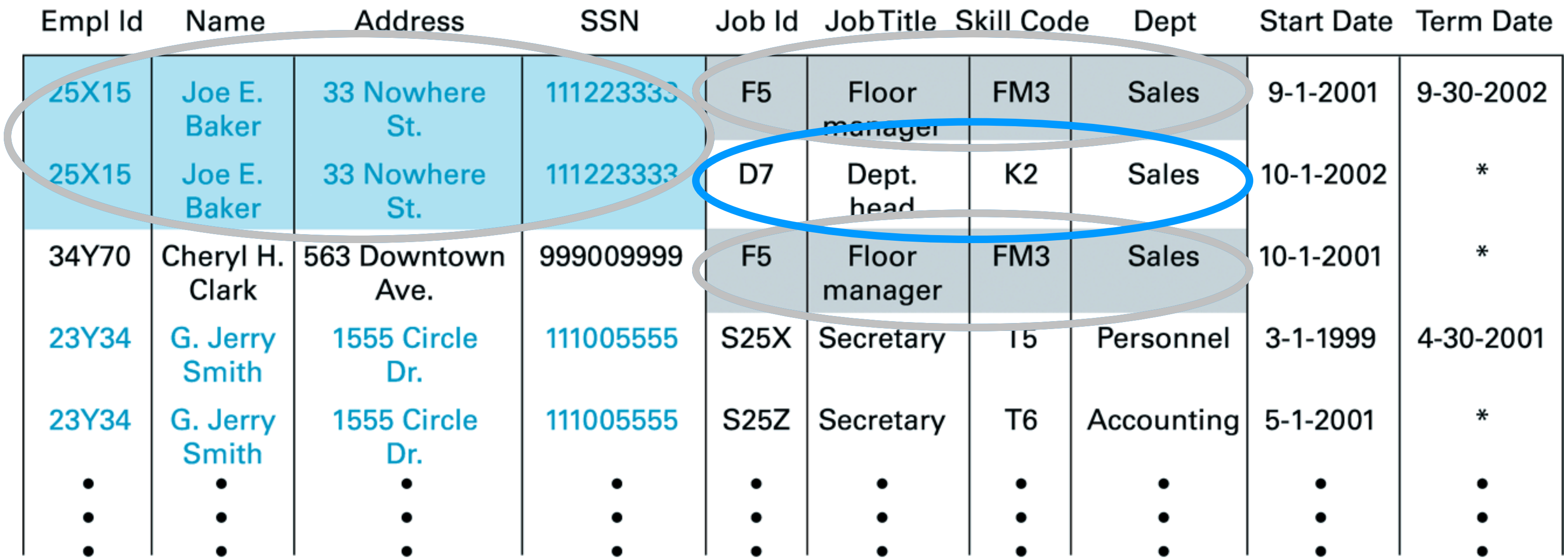


Empl Id	Name	Address	SSN
25X15	Joe E. Baker	33 Nowhere St.	111223333
34Y70	Cheryl H. Clark	563 Downtown Ave.	999009999
23Y34	G. Jerry Smith	1555 Circle Dr.	111005555
●	●	●	●
●	●	●	●
●	●	●	●

  - – row in a relation is called ‘*tuple*’
  - – column in a relation is called ‘*attribute*’

9.3: Issues of Relational Design

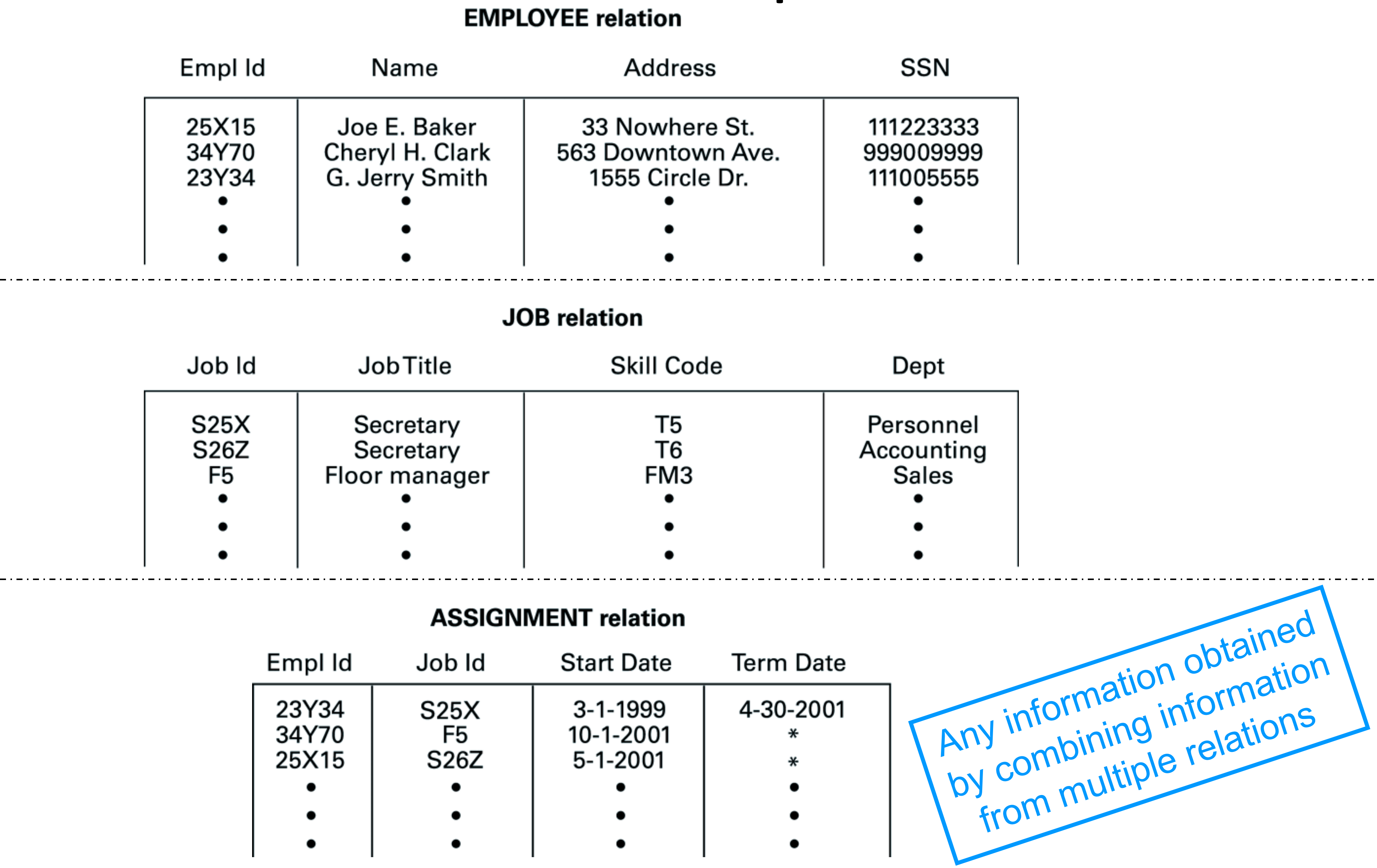
- So, *relations* make up a relational database…
- … but this is not so straightforward:



Empl Id	Name	Address	SSN	Job Id	Job Title	Skill Code	Dept	Start Date	Term Date
25X15	Joe E. Baker	33 Nowhere St.	111223333	F5	Floor manager	FM3	Sales	9-1-2001	9-30-2002
25X15	Joe E. Baker	33 Nowhere St.	111223333	D7	Dept. head	K2	Sales	10-1-2002	*
34Y70	Cheryl H. Clark	563 Downtown Ave.	999009999	F5	Floor manager	FM3	Sales	10-1-2001	*
23Y34	G. Jerry Smith	1555 Circle Dr.	111005555	S25X	Secretary	T5	Personnel	3-1-1999	4-30-2001
23Y34	G. Jerry Smith	1555 Circle Dr.	111005555	S25Z	Secretary	T6	Accounting	5-1-2001	*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

- Problem: more than one concept combined in single relation

9.3: Redesign by extraction of 3 concepts



EMPLOYEE relation

Empl Id	Name	Address	SSN
25X15	Joe E. Baker	33 Nowhere St.	111223333
34Y70	Cheryl H. Clark	563 Downtown Ave.	999009999
23Y34	G. Jerry Smith	1555 Circle Dr.	111005555
•	•	•	•
•	•	•	•
•	•	•	•

JOB relation

Job Id	JobTitle	Skill Code	Dept
S25X	Secretary	T5	Personnel
S26Z	Secretary	T6	Accounting
F5	Floor manager	FM3	Sales
•	•	•	•
•	•	•	•
•	•	•	•

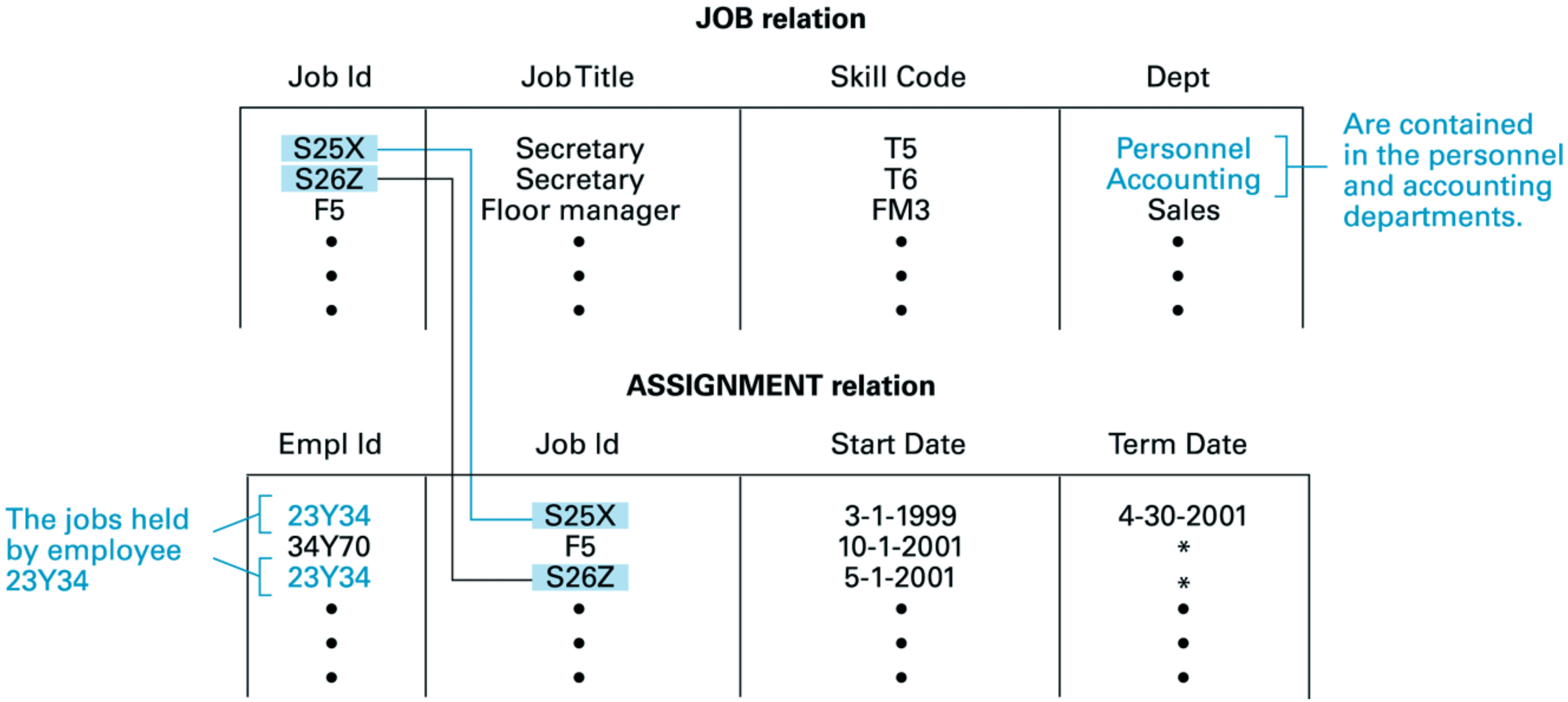
ASSIGNMENT relation

Empl Id	Job Id	Start Date	Term Date
23Y34	S25X	3-1-1999	4-30-2001
34Y70	F5	10-1-2001	*
25X15	S26Z	5-1-2001	*
•	•	•	•
•	•	•	•
•	•	•	•

Any information obtained by combining information from multiple relations

9.3: Example:

- Finding all departments in which employee 23Y34 has worked:



JOB relation

Job Id	Job Title	Skill Code	Dept
S25X	Secretary	T5	Personnel
S26Z	Secretary	T6	Accounting
F5	Floor manager	FM3	Sales
•	•	•	•
•	•	•	•
•	•	•	•

Are contained in the personnel and accounting departments.

ASSIGNMENT relation

Empl Id	Job Id	Start Date	Term Date
23Y34	S25X	3-1-1999	4-30-2001
34Y70	F5	10-1-2001	*
23Y34	S26Z	5-1-2001	*
•	•	•	•
•	•	•	•
•	•	•	•

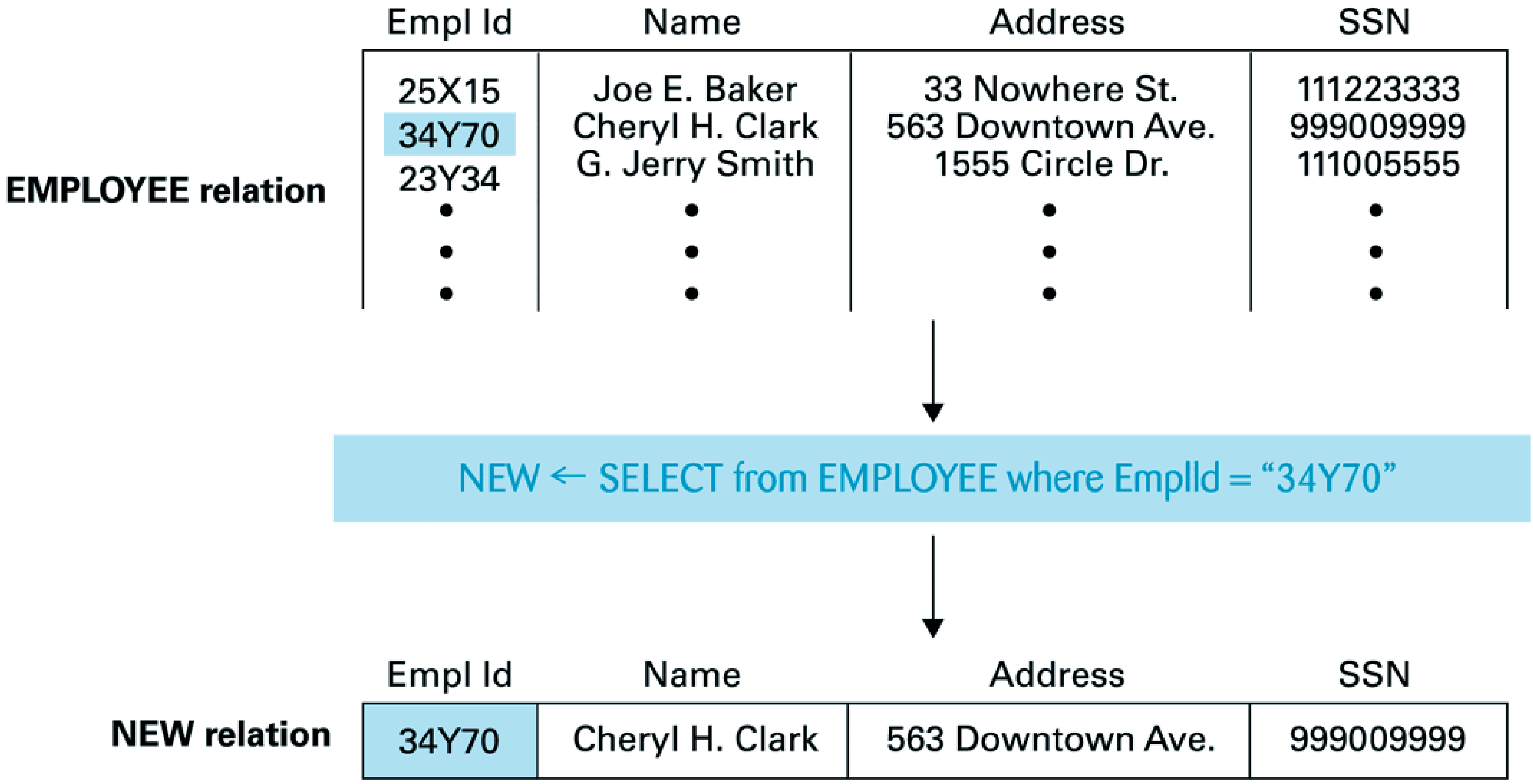
The jobs held by employee 23Y34

9.3: Relational Operations

- Extracting information from a relational database by way of *relational operations*
  - Most important ones:
    - (1) extract tuples (rows) : SELECT
    - (2) extract attributes (columns) : PROJECT
    - (3) combine relations : JOIN

- Such operations on relations produce other relations
  - so: they can be used in combination, to create complex database requests (or '*queries*')

9.3: The SELECT operation



EMPLOYEE relation

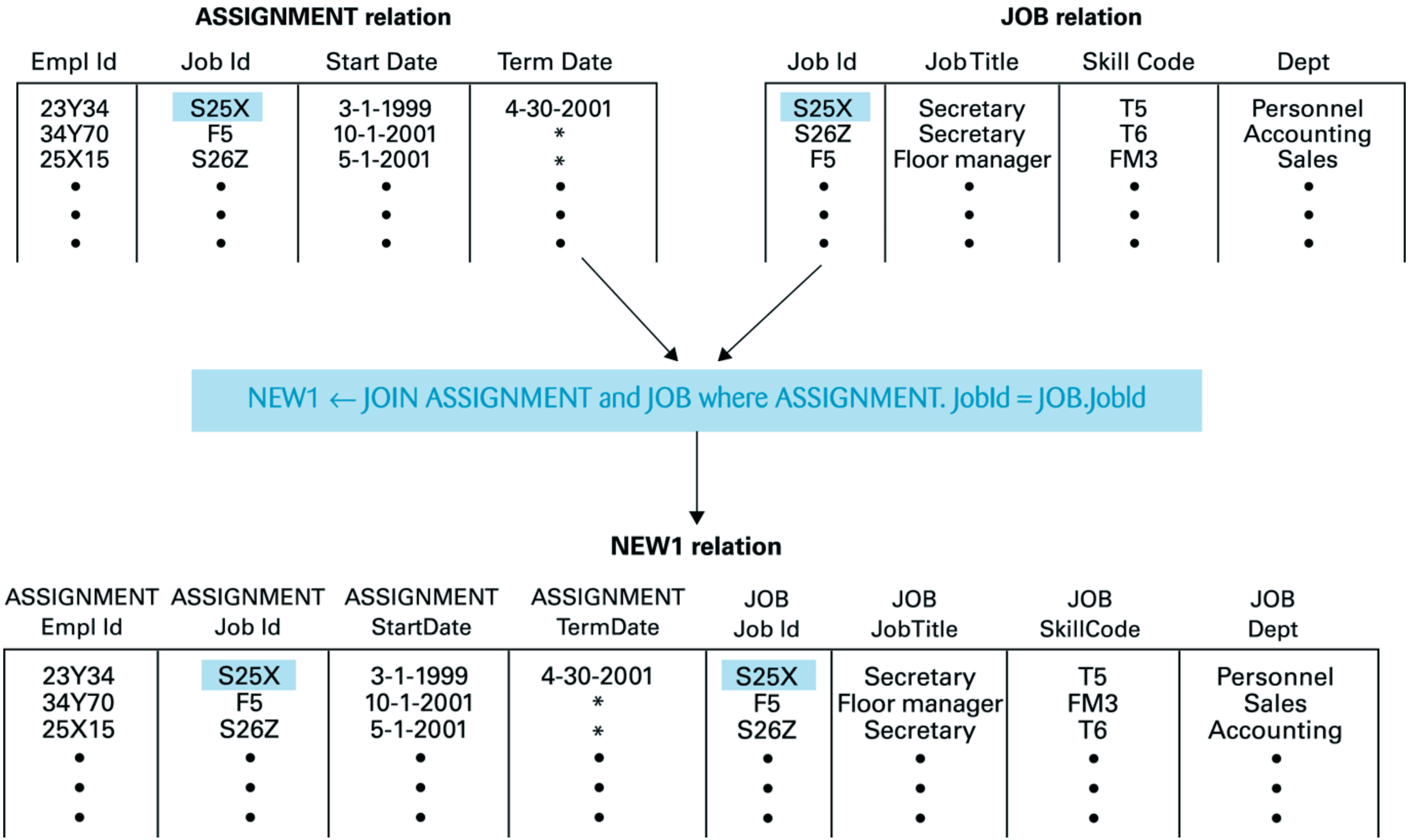
Empl Id	Name	Address	SSN
25X15	Joe E. Baker	33 Nowhere St.	111223333
34Y70	Cheryl H. Clark	563 Downtown Ave.	999009999
23Y34	G. Jerry Smith	1555 Circle Dr.	111005555
⋮	⋮	⋮	⋮

NEW ← SELECT from EMPLOYEE where EmplId = "34Y70"

NEW relation

Empl Id	Name	Address	SSN
34Y70	Cheryl H. Clark	563 Downtown Ave.	999009999

9.3: The JOIN operation



ASSIGNMENT relation

Empl Id	Job Id	Start Date	Term Date
23Y34	S25X	3-1-1999	4-30-2001
34Y70	F5	10-1-2001	*
25X15	S26Z	5-1-2001	*
•	•	•	•
•	•	•	•
•	•	•	•

JOB relation

Job Id	Job Title	Skill Code	Dept
S25X	Secretary	T5	Personnel
S26Z	Secretary	T6	Accounting
F5	Floor manager	FM3	Sales
•	•	•	•
•	•	•	•
•	•	•	•

NEW1 ← JOIN ASSIGNMENT and JOB where ASSIGNMENT. JobId = JOB.JobId

NEW1 relation

ASSIGNMENT	ASSIGNMENT	ASSIGNMENT	ASSIGNMENT	JOB	JOB	JOB	JOB
Empl Id	Job Id	StartDate	TermDate	Job Id	JobTitle	SkillCode	Dept
23Y34	S25X	3-1-1999	4-30-2001	S25X	Secretary	T5	Personnel
34Y70	F5	10-1-2001	*	F5	Floor manager	FM3	Sales
25X15	S26Z	5-1-2001	*	S26Z	Secretary	T6	Accounting
•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•

Chapter 9 - Database Structures: Conclusions

- Database Structures:
  - (large) integrated collections of data that can be accessed quickly
- Database Management System
  - provides high-level view of actual data storage (database model)
- Relational Model most often used
  - relational operations: SELECT, PROJECT, JOIN, …
  - high-level language for database access: SQL

Background: Relational Database

- Ted Codd Mathematician paper
- “A relational Model of data for large shared data banks”
- Chamberlin inspired by Codd’s Symposium and convinced IBM to create R system group and to fund a research project to build a prototype of relational DB which leads to DB2 and SQL creations
- IBM focused on IMS in 1968

Background (Continued...)

- Based on Codd’s Work two professors from university of Berkeley started a project “Ingres”
- Researched Competition flared between the two groups and number of the research papers are being published. IBM did not realizing the potential of the project, published these papers publicly.

Background (Continued…)

- Larry Ellison formed a company "System Development Labs" which recruited Employees from System R and Ingres. He started developing a system based on the research papers by the funding from CIA and NAVY.
- First Structured Query Language was launched in 1979.
- IBM came up with its version in 1983, with SQL/IDS 1980
- Ellison Changed the Company name to Oracle
- In 2003, $ 7 Billion Relational DB

Chapter contents

- Good Decision requires good information and good information is derived from raw facts called data.
- *Good Decision means which delivers accurate, relevant and timely information*.
- What is DB?, What does it do? And Comparison between other Data Management Methods, Different Types of DB and Importance of DB Design.

File System

- Database is evolved from the File Systems.
- Understand the characteristics of the file system.
- Data management limitations by File system.
- Eliminations of the short comings of the file system by DBMS.

Basic Definitions

- Data: raw facts
  - Not processed yet to reveal their meaning
  - Constitute building blocks of information
  - For Examples:
    - Online Surveys
    - Online Data Entry Forms
    - Excel Sheets
    - Reports Forms
- Record keeping with the raw facts
  - Example: Students
    - » Pass 90%
    - » Fail 10 %
    - » Quick Answers

Basic Definitions (Continued...)

- Information: is produced by processing data and reveals meaning of data
  - Good, timely, relevant information key to decision making
  - Good decision making key to organizational survival
  - Example: Informed decisions to meet student grading record
  - Raw data: Storage, Processing and presentation
- Complex formatting: is required when working with complex data types such as sounds, videos ‘or’ images.
- For Example: Yes/No or Y/N

Basic Definitions (Continued...)

- Knowledge: the body of the information and facts about a specific subjects
- Knowledge implies familiarity awareness and understanding of information.
- New Knowledge can be derived from Old Knowledge.

Basic Definitions (Continued...)

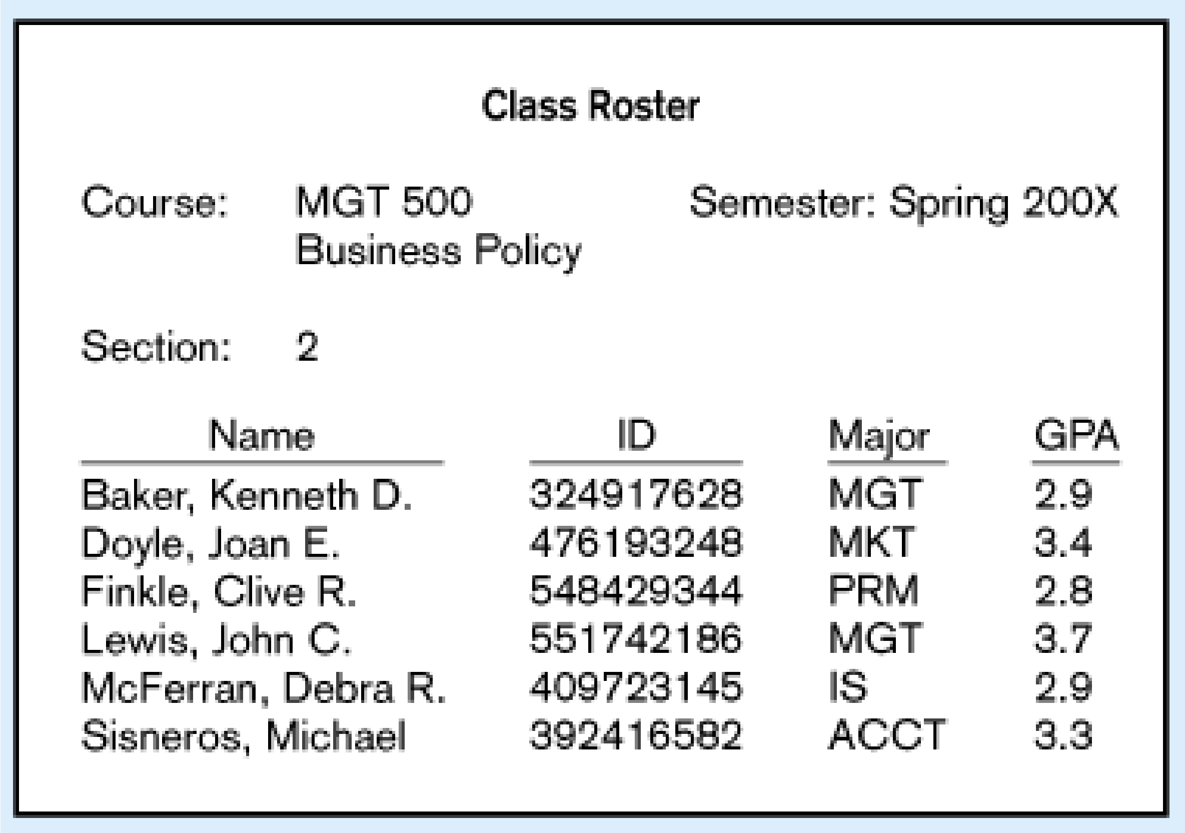
- Data Management is a discipline that focuses on the proper generation, storage and retrieval of data.
- Efficient Data Management requires computer DB.

Basic Definitions (Continued...)

- Database: shared, integrated computer structure housing:
  - End user data
  - Metadata
- Metadata provides a description of the data characteristics and set of relationships that link the data within the Database.
  - Structural Metadata -> data about data
  - Descriptive Metadata-> Content about content

An Example

- Converting data to information



Class Roster

Course: MGT 500 Business Policy

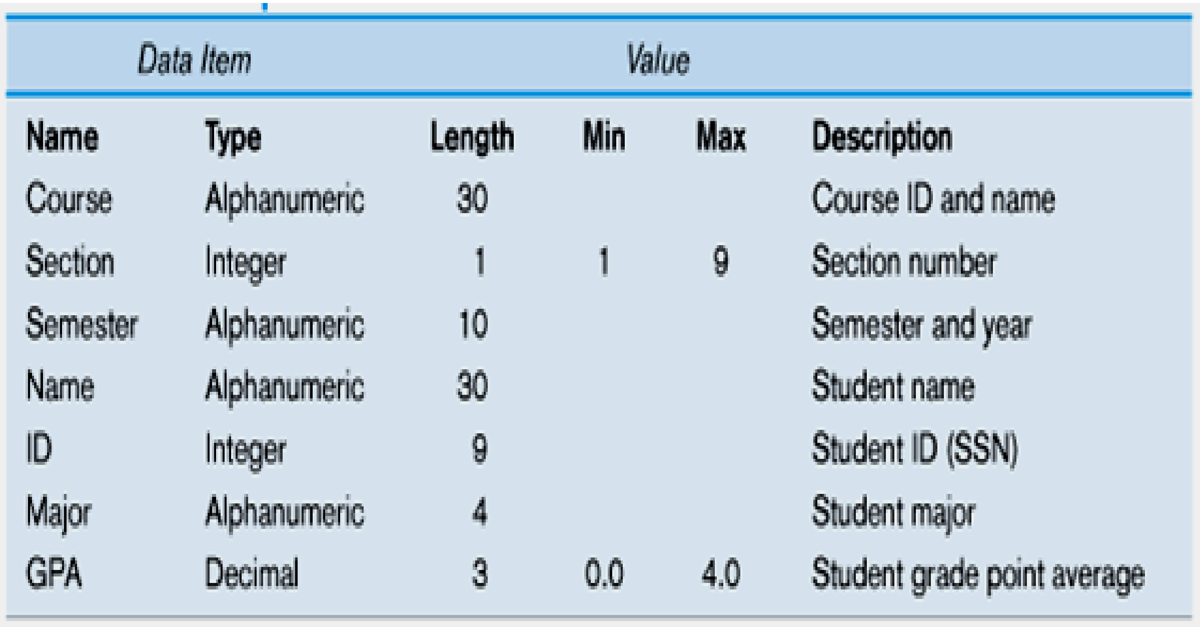
Semester: Spring 200X

Section: 2

Name	ID	Major	GPA
Baker, Kenneth D.	324917628	MGT	2.9
Doyle, Joan E.	476193248	MKT	3.4
Finkle, Clive R.	548429344	PRM	2.8
Lewis, John C.	551742186	MGT	3.7
McFerran, Debra R.	409723145	IS	2.9
Sisneros, Michael	392416582	ACCT	3.3

An Example (Cont’d)

- Metadata

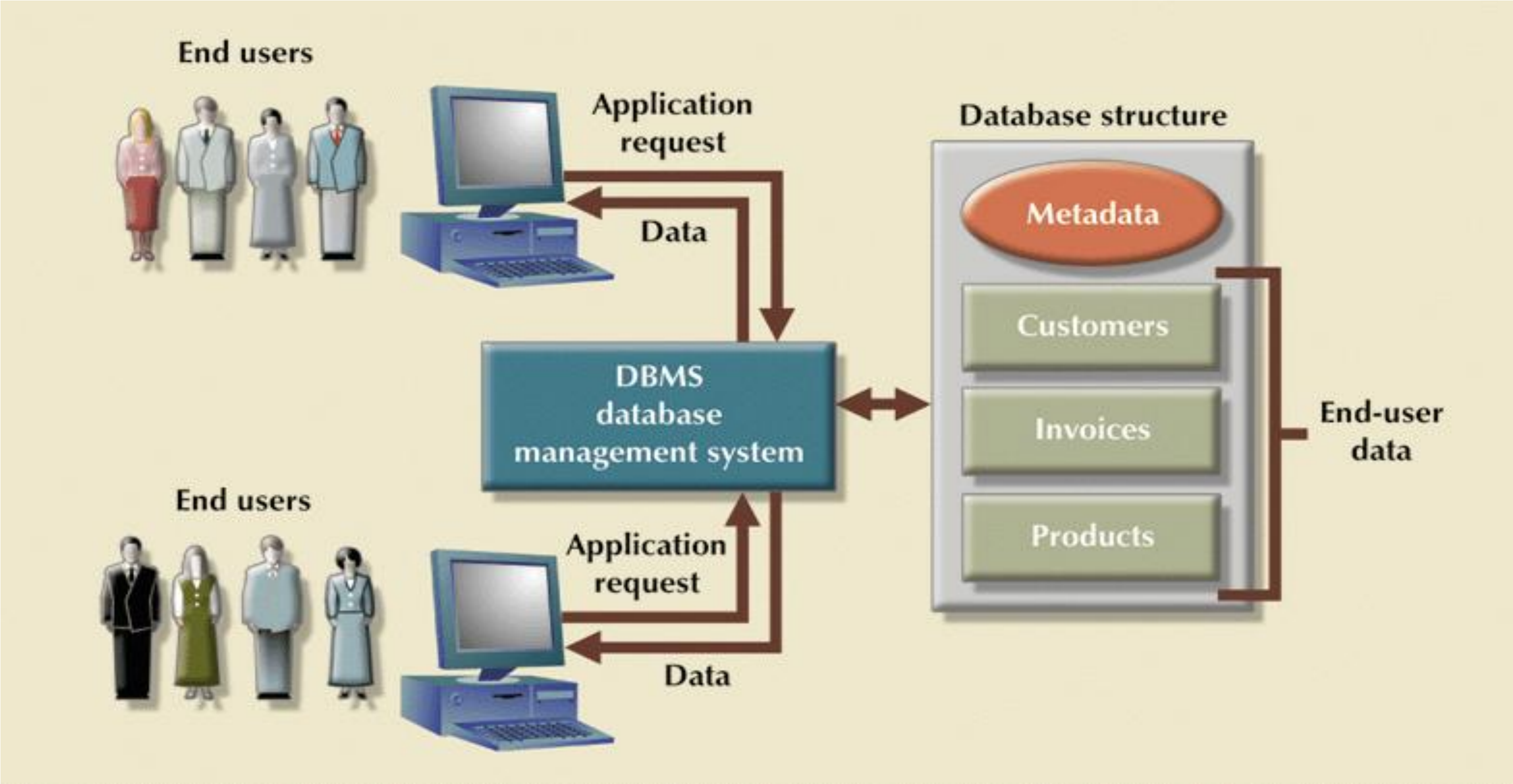


Data Item		*Value*			
Name	**Type**	**Length**	**Min**	**Max**	**Description**
Course	Alphanumeric	30			Course ID and name
Section	Integer	1	1	9	Section number
Semester	Alphanumeric	10			Semester and year
Name	Alphanumeric	30			Student name
ID	Integer	9			Student ID (SSN)
Major	Alphanumeric	4			Student major
GPA	Decimal	3	0.0	4.0	Student grade point average

What is a Database Management System (DBMS)

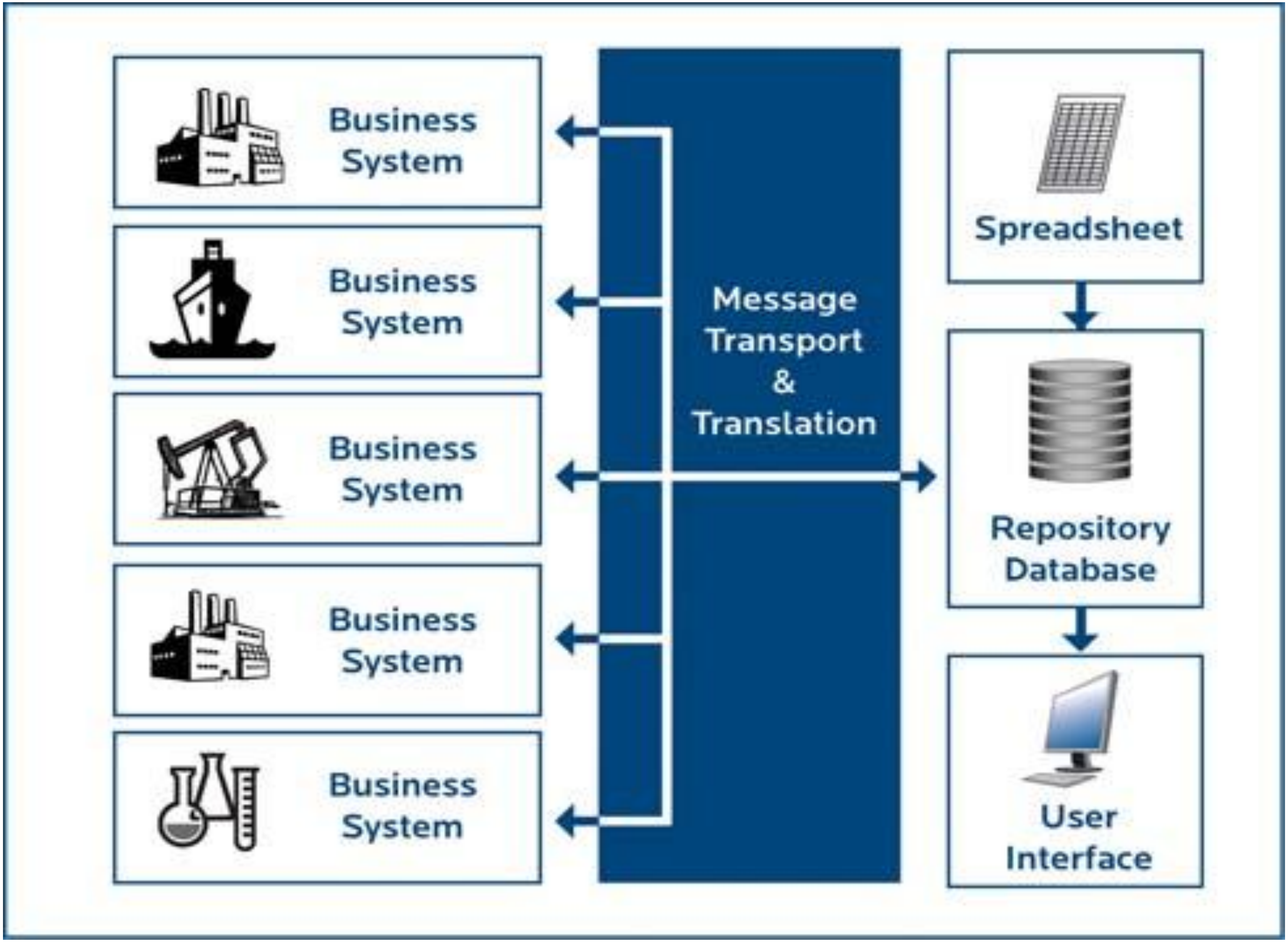
- A collection of programs that manages the database structure and controls access to the data stored in the database
  - Possible to share data among multiple applications or users
    - Example: bank and its ATM machines
  - Makes data management more efficient and effective
    - End users have better access to more and better-managed data
- DBMS hides much of the database's internal complexity from application program and End user

DBMS Manages Interaction



Advantages of the DBMS

- Improved data sharing
  - Shared among users and applications
- Better Data Integration
  - Different User’s views into single data Repository
    - Repository: can be a place where multiple DBs or files are located for distribution over the network.
- Minimized Data inconsistency
  - Different versions of the same data.
    - Example: Product ID and Product Number in different departments


Business System
Business System
Business System
Business System
Business System
Message Transport & Translation
Spreadsheet
Repository Database
User Interface

Advantages of the DBMS

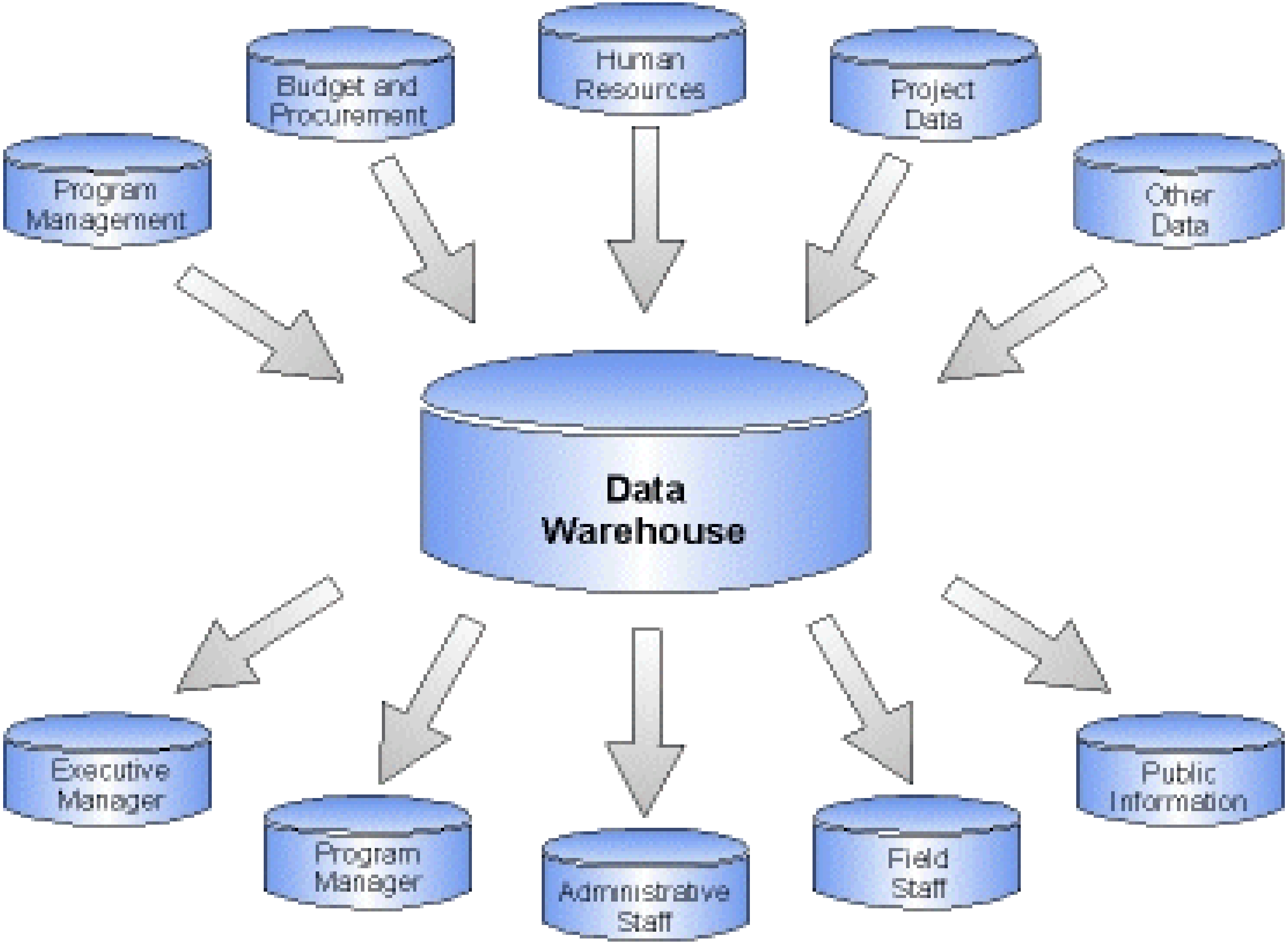
- Improved Data access
  - Quick answers to the ad hoc queries
  - Query is a complete question: a specific request for data manipulation (read or update data)
  - DBMS sends back an Answer (Query result set) to the application
- Improved Decision Making
  - Better managed data and improved data access ->to better quality information ->better decisions
- Increased End User Productivity

Types of the databases

- Single User Database: Runs on a personal Computer
- Multiuser Database: less than 50 workgroup DB, more than 50 Enterprise DB
- Location wise:
  - Single site: Centralized DB
  - Several sites: Distributed DB
- Function wise: Operational/transactional/production
  - Time Sensitive information gathered
  - Support a company 's day to day operations

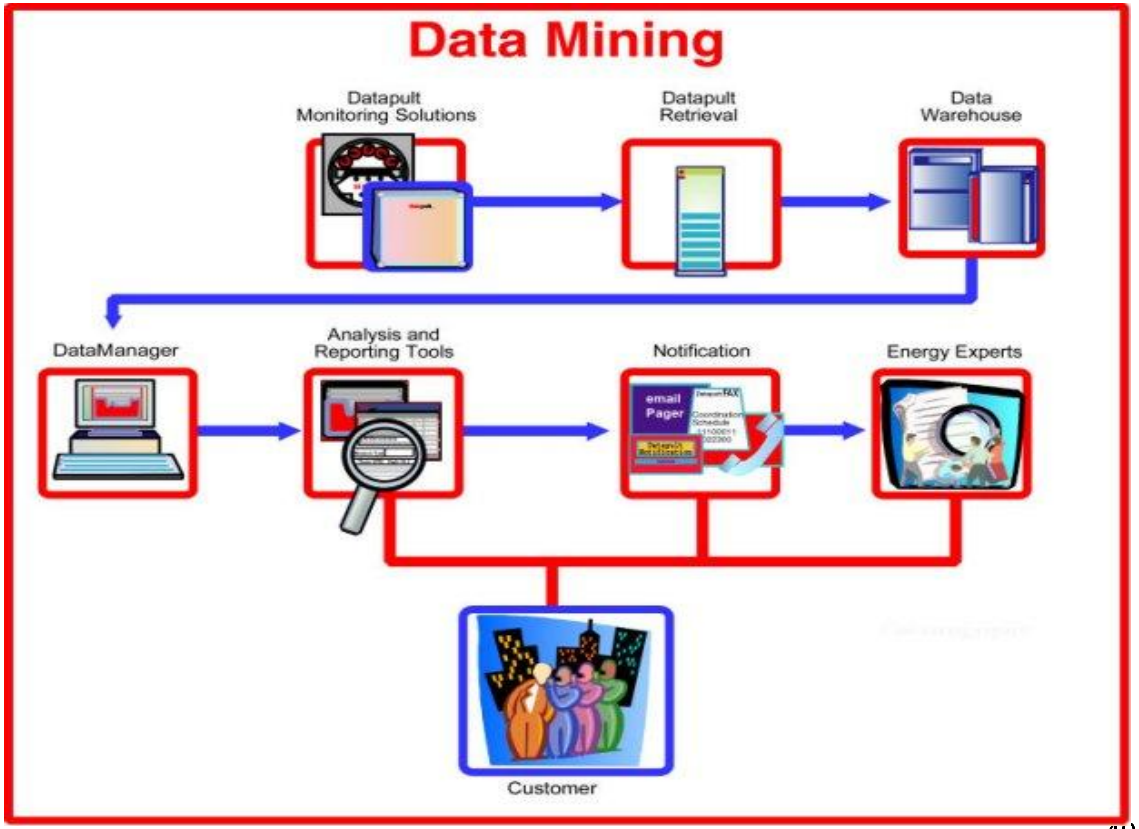
house

- A collection of data designed to support management decision making generally refers to combination of many different DBs across entire enterprise.
  - Generate information to make tactical or strategic decisions
    - Extensive data messaging
    - Historical data from operational DB
    - Examples:
      - Formulate pricing decisions
      - Sale forecast
      - Market Position

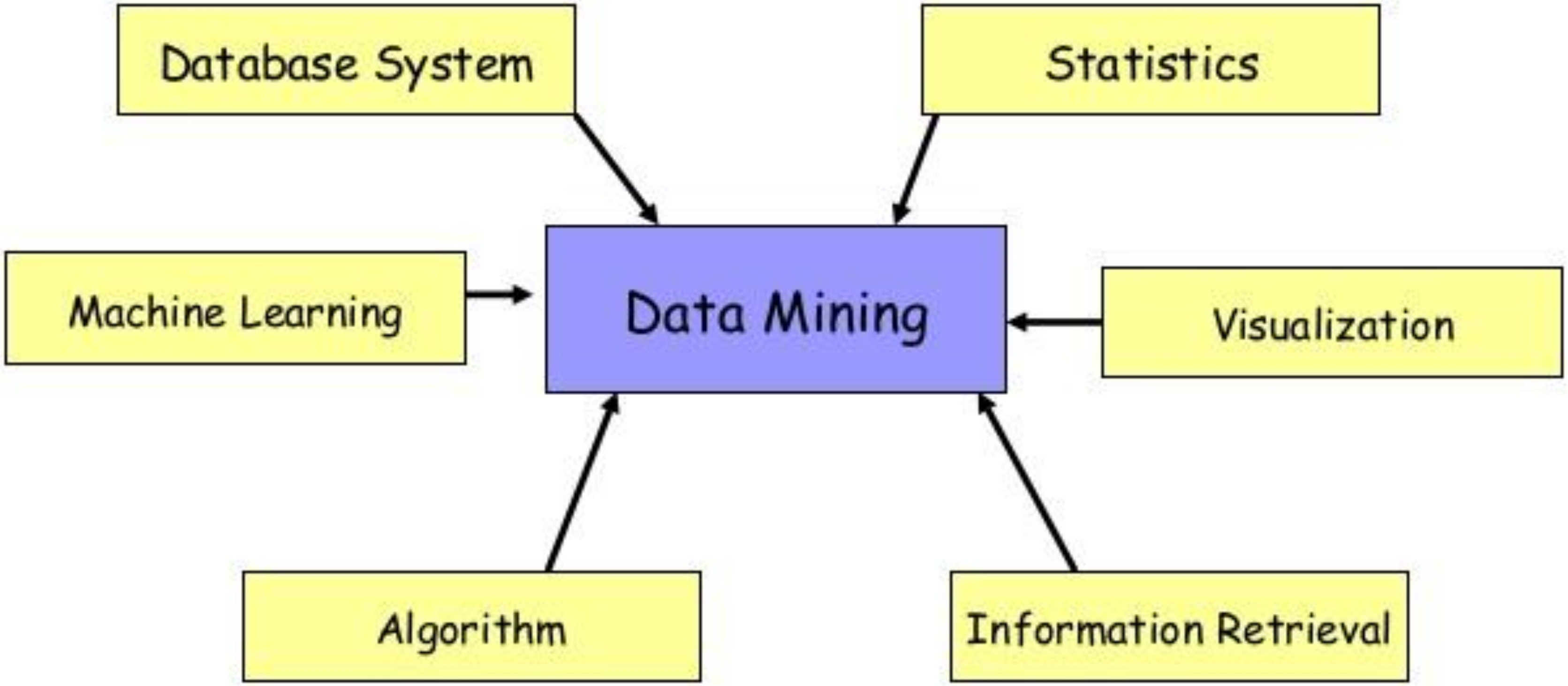

Budget and Procurement
Human Resources
Project Data
Program Management
Other Data
Data Warehouse
Executive Manager
Program Manager
Administrative Staff
Field Staff
Public Information

Data Mining Concept

- A class of database applications that look for the hidden patterns in a group of data that can be used to predict future behavior.



Disciplines Of Data Mining



Assignment

- What is database structure and Database management system?
- How do you differentiate between file system and computer file systems?
- Give an example of computer file system?
- Bring an example to describe “Database application” ?

End of the Lecture