7BUIS030W
Data System Concepts and Fundamentals

Lecture -2

Lecture-2 Outline

Using Data, Database, data redundancy, data inconsistency, database normalisation, structured data, Relational database model, Entity relationship diagrams (ERD)

Using data

A large amount of data is captured and used in many areas of life.

Large scale extraction, storage , processing and handling of data are performed by various environments

Eg: Big Data, Data Mining

Big data and Data mining

Big data :

Big data can be defined as large or complex data sets.

Big data is often used for making predictions based on patterns that can be seen in the data.

Data mining:

To mine data and extract information from it.

Corporations use Data Mining to set goals and help chart the course for a business.

Big Data is a vast entity of data and Data Mining is a tool to sieve through it for better utilization.

Big data and Data mining Examples

- Fraud detection
- Business clarity.
- Customer statistics alteration.
- Money laundering
- Market analysis
- Health care data analysis

Big data and Data mining on Retail Industry

To ensure they are retaining their customers and getting new business, many shops now use big data to understand and target their audience in the right way.

- If you shop at any of the major supermarkets the likelihood is you will belong to at least one (if not many) of the loyalty schemes, from Tesco's Clubcard or Sainsbury's Nectar Card.
- These schemes allow the supermarket to collect data on you every time you shop and swipe.
- By tracking these shopping habits (what you buy, when you buy, and how much you buy) the shops are able to target specific promotions or vouchers at you.
- This data can also be a great way for the supermarket to understand market trends and pricing.
- This same data is also being used on your digital shop, the shops are tracking what you buy so they can offer recommended products and tailor adverts to suit you.

Big data and Data mining on Retail Industry

- Big data is also helping to improve customer service and experience when interacting directly with these brands.
- As with any business supermarkets receive complaints and compliments each day.
- By tracking these comments and analysing the data from them, supermarkets can learn where to improve, what products may be having issues and even how to make sure they retain existing customers.
- Many of the stores will also have multiple departments, by collecting the data they can ensure their ‘profile’ on you is in one place and your experience is seamless.
- Big data can also help the supermarkets predict shopping trends and buyer habits.
- This data helps the store not just know what to stock but how to place them in the store, for example if two products are typically bought together should they be placed in the same, or near, aisles?

Big data and Data mining on Retail Industry

By **mining big data using a team of data scientists**, retailers and suppliers will better understand what you want, when and how and they will develop their range, pricing and distribution models to meet those needs.

Data mining, or knowledge discovery from data (KDD), is the process of uncovering trends, common themes or patterns in “big data”.

For example, an early form of data mining was used by companies to analyze huge amounts of scanner data from supermarkets

Big data and Data mining on Retail Industry

This analysis revealed when people were most likely to shop, and when they were most likely to buy certain products, like wine or baby products.

This enabled the retailer to maximize revenue by ensuring they always had enough product at the right time in the right place.

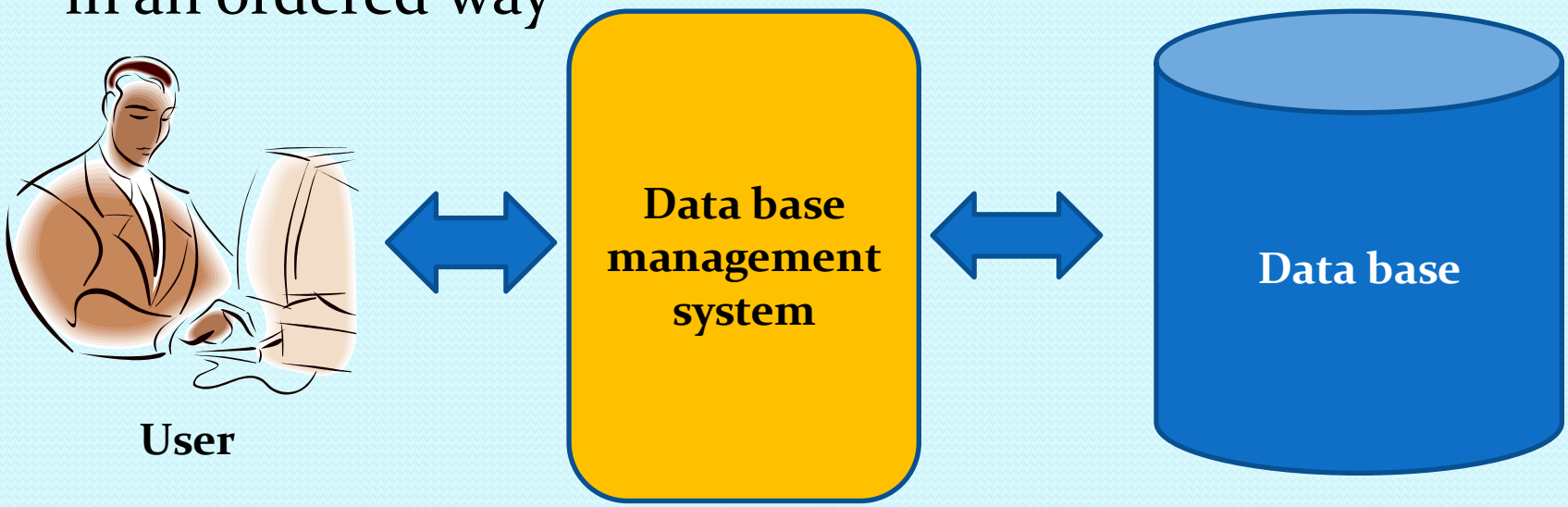
Although commonly used in large businesses and organizations, any kind of data can be mined, from any type of database.

Database

Databases are used to organise data in a clear and consistent way.

A dictionary is a physical database!

A database is a data store in which the data is organised in an ordered way



Uses of Database

Computer Database:

- Computers are ideal for storing and manipulating data.
- Enables fast processing of data in a structured way.
- Computer Databases are very powerful tools used in all areas of computing.
- One of the main benefits of computer databases is that they make it easy to store information so it is quick and easy to find.

Uses of Database

Social Networking sites
- User information including photographs and location

Media applications
-Media files , eg: iTunes, Google Music

Retail industry
-Market analysis, consumer research

Weather pattern predictions
-Software models to predict natural calamities

Database Software

Database softwares:

- **Oracle**
- **MySQL**
- **SQL server**

Programming languages:

Databases can also be created and organised using programming languages.

- **SQL**
- **Visual Basic**
- **Delphi**

Using **programming languages** means that you can customise a database to do exactly as you want.

Database Design

- When designing a database, it is important to decide what the structure will be before you start adding data.
- It is difficult to change a database structure once it has been set in place.
- If data was not organised, it would be difficult to work with.

Database structure

A **database** is organised using a set of key components. These include:

- entities - each recorded item
- attributes - details about the entity
- field - columns used to capture attributes
- record - one row of details about an entity
- table - a set of fields and records
- primary key - unique number for an entity

Database Design

Methodologies:-

- Identify the **entities**
- Identify **attributes**
- Creating database table/s
- Establish **relationships**
- Determine the **multiplicities**
- Determine the **keys**

Database design

Case study-01- Part 1

Consider the scenario of a University where a database has to be created to store the details of its students

Identify the Entity

An entity is any item that has its details stored as data.

Entity should have independent existence and has to be uniquely identified

Entities can be a physical object such as a person, book or a concept such as customer transactions.

This database is about students.

It stored details about the students .

Hence the student is the entity in this case.

Database Design

Identify the attributes

The properties or details of the student that are relevant to the university should be identified.

In this case it would be name, date of birth, year of study, tutor group. These are the attributes of the entity student.

Attributes are items of data stored about an entity

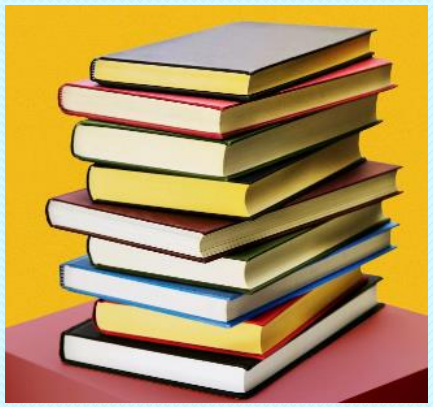
A person is an entity with attributes including age, height and nationality, among many others.

When you design a database you need to think about which attributes you want to store.

Entities and Attributes

Hotel

Attributes
HotelId
TravellerRatings
Ranking
Award
location

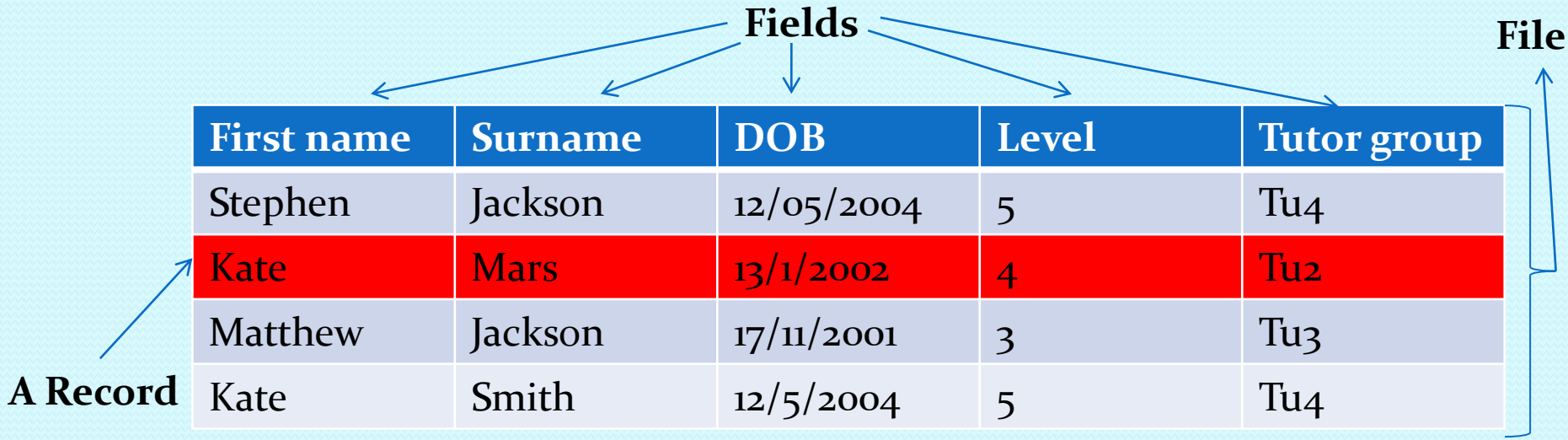


Book

Attributes
Book name
Author name
year
Publisher

Database table design

Once the entity and attributes are identified , a structure of storing the data can be set up. The structure is in the form of a table.



First name	Surname	DOB	Level	Tutor group
Stephen	Jackson	12/05/2004	5	Tu4
Kate	Mars	13/1/2002	4	Tu2
Matthew	Jackson	17/11/2001	3	Tu3
Kate	Smith	12/5/2004	5	Tu4

In the table, same items of information about each student are being stored. These items of information are called **Fields**

All the information about a particular student is called a **Record**.

Each row of the table represents one record.

All records in a table forms a **File**

Data types

When you create a database you need to set data types for each **field**.

For example, in a film database you might need alphabetical characters for 'Titles', but numbers for 'Duration'.

Fields are usually restricted to a certain data type.

The actual units of data that are entered into a database give the attributes for each entity. These units of data are also called **data elements.**

Data structure

The structure of the database is also called the **schema** or **dictionary**.

When you design a database you need to create a schema to explain what type of data is being stored.

There are key areas to consider about the data:

- **data type** - the data types are used in each field
- **field size** - the maximum or minimum size of entry
- **validation** - the rules for accepting data
- **key field** - the field which is the primary key

Data types

Data typing is a way of classifying data values that have common properties.

Different kinds of data values also need different amounts of **memory** to store them, and have different operations that can be performed upon them.

Data types

The most commonly-supported data types are:

integers (whole numbers), for example: 4, 27, 65535

floating point numbers (with decimal points, sometimes called real numbers, or floats), for example: 4.2, 27.4, 56.8

characters, for example: a, F, 3, $, £, #

character strings (ordered sequences of characters), for example: abc, def456, 3erf78!@

Boolean values, for example: 'True' or 'False'

Data Redundancy

Case study 01-Part-2

Suppose the database requires information on the books lend by each student

The table will now show information on the details of the books the students have borrowed.

What changes will be made in the table?

First name	Surname	DOB	Level	Tutor group	Book name	author	Borrow date
Stephen	Jackson	12/5/2004	5	Tu4	Computing	J.Smith	15/10/2019
Kate	Mars	13/1/2002	4	Tu2	Big Data	J.Smith	13/12/2019
Matthew	Jackson	17/11/2001	3	Tu3	Statistics	F.Greg	8/01/2020
Kate	Smith	12/5/2004	5	Tu4	Data Science	S.Sinda	21/01/2020
Kate	Smith	12/5/2004	5	Tu4	Gardening	S.Morr	21/01/2020

Data Redundancy

When the student borrows multiple books , the same student details have to be entered multiple times.

Also if the same book is borrowed by another person, the same book information have to be re-entered

Having the same data entered more than once into the same table is called **Data Redundancy**

How to avoid Data redundancy??

In order to avoid repeated data entry, new fields can be added to the table

These fields might be used by some data but in most cases it will never be used

Highly inefficient structure!!

Data Inconsistency

Along with data redundancy repeated entry of data can also cause errors. In the following table the first name stephen was misspelt as 'Stepen'

First name	Surname	DOB	Level	Tutor group	Book name	author	Borrow date
Stephen	Jackson	12/5/2004	5	Tu4	Computing	J.Smith	15/10/2019
Kate	Mars	13/1/2002	4	Tu2	Big Data	J.Smith	13/12/2019
Matthew	Jackson	17/11/2001	3	Tu3	Statistics	F.Greg	8/01/2020
Kate	Smith	12/5/2004	5	Tu4	Data Science	S.Sinda	21/01/2020
Kate	Smith	12/5/2004	5	Tu4	Gardening	S.Morr	21/01/2020
Stepen	Jackson	12/5/2004	5	Tu4	Statistics	F.Greg	24/1/2020

The occurrence of errors when same data is entered more than once is called **Data Inconsistency**

Structured Database design

The problems of Data redundancy and data inconsistency can be solved by creating more than one data table in a single data base and linking them together.

In the part -2 of the case study -01, data redundancy occurs when the book lending information was also included in the table

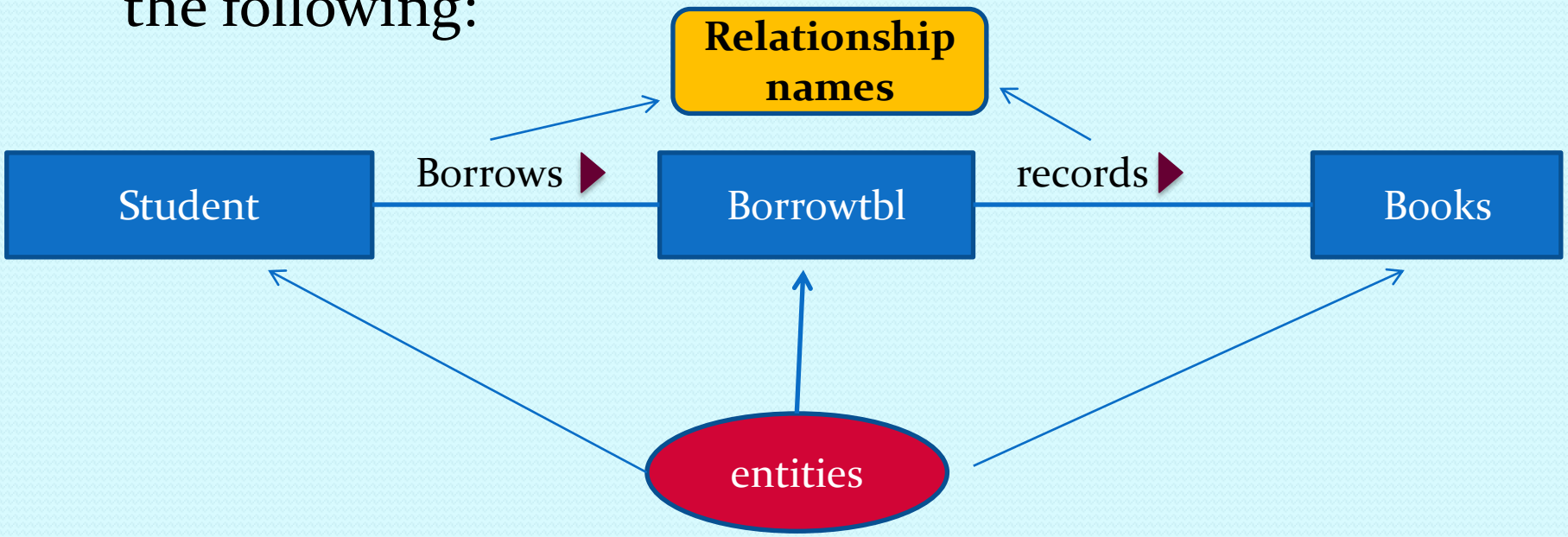
This could be avoided by creating another entity named book and provide the book name and author name as the attributes

Also the book borrowing is now recorded as another entity with book number and borrow date as the attributes

Relational Database

The three different tables created can now be linked.

The relationship between the three tables is shown as the following:



Database with multiple tables linked with relationships is called a **relational database**

Relational database

Relational **databases** allow data to be stored in a clear, organised manner across multiple tables.

Links, known as **relationships**, are formed to allow the data to be shared across the tables.

A single table database is useful for recording a limited amount of data. A large single table database can be inefficient as it takes up more space and **memory** than a relational database.

It also requires new data to be added every time you enter a new record, whereas a relational database does not.

Data redundancy can more easily be avoided in relational databases.

Therefore, if you have a large set of data it is more efficient to create separate tables and connect them with relationships.

Data Normalisation

The process of structuring the data to create relational databases is called **Normalisation**

Normalisation is the process of analysing how to make databases more efficient by using separate tables to reduce **redundant data**.

When a database is normalised, data is broken down into smaller tables and relationships are used to link them.

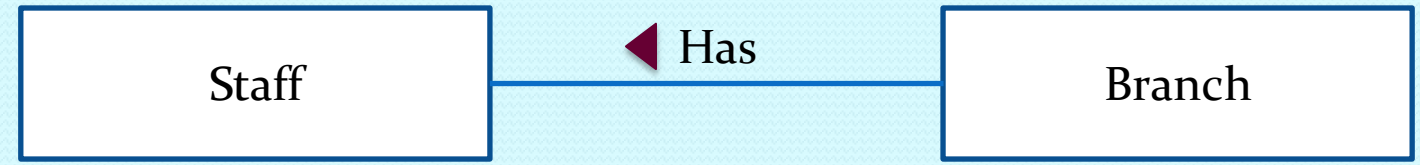
Entity-relationship Diagrams

As shown in the case study 01-Part-2, the relationships between entities can be shown using Entity-relationship diagrams

The relationships provide a set of meaningful associations among entities

Examples:

Provide the ERD of the staff and branch of a bank



ERD exercise-01

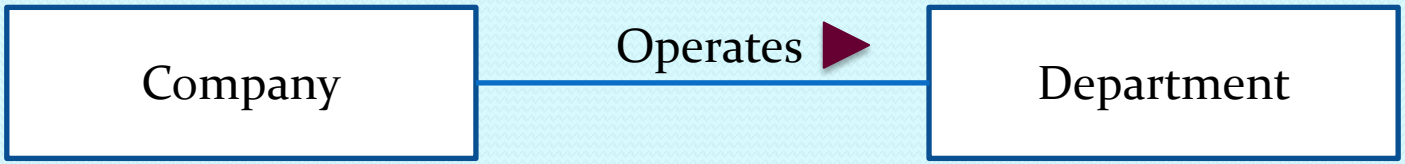
Create a basic conceptual ERD for each of the following descriptions. Make sure you include the entity names, relationship names and reading directions

1. A company operating multiple departments
2. Employees working in each department
3. Order history of a product
4. Computers in labs

ERD exercise-01 Solution Part-1

Create a basic conceptual ERD for each of the following descriptions. Make sure you include the entity names, relationship names and reading directions

1. A company operating multiple departments

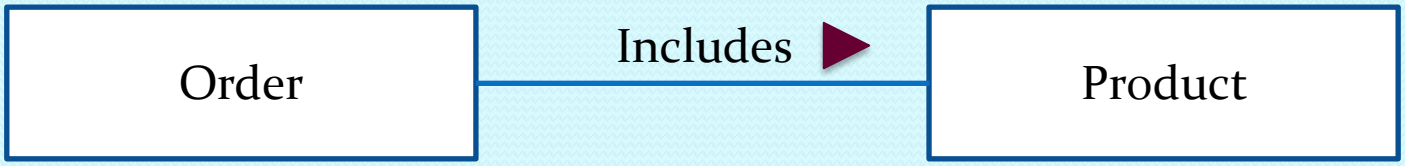


2. Employees working in each department

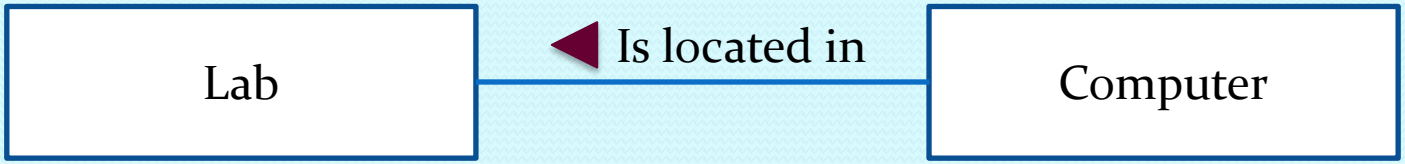
ERD exercise-01 Solution Part 2

Create a basic conceptual ERD for each of the following descriptions. Make sure you include the entity names, relationship names and reading directions

3. Order history of a product



4. Computers in labs



ERD exercise-02

For an online grocery super market that sells a large number of products to members of the general public. Customers can place order for a number of items and repeat the process until they decide to finalise their order and check out. Create a basic conceptual ERD and include entities and relationships

ERD Exercise-02 Solution