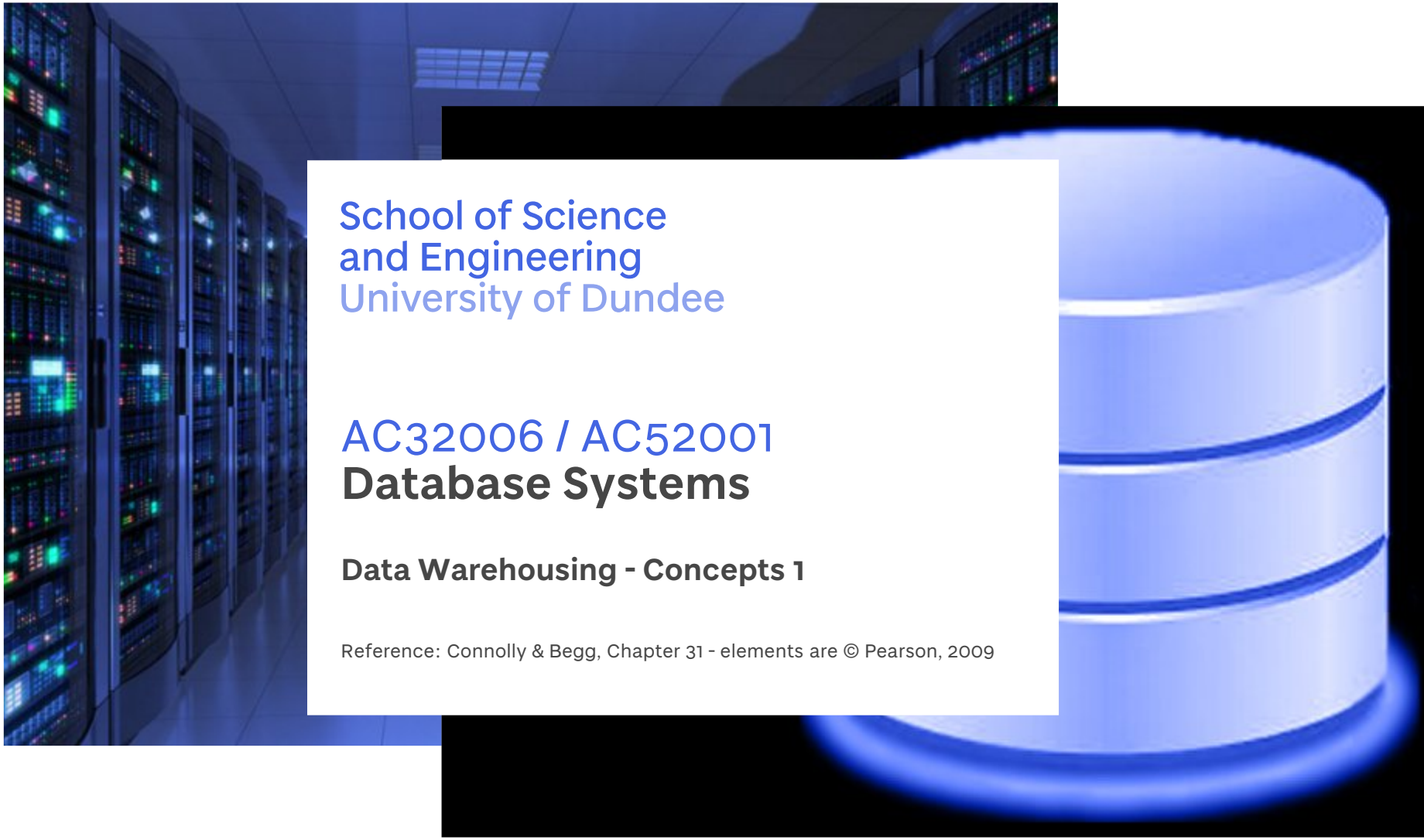


School of Science
and Engineering
University of Dundee

AC32006 / AC52001
Database Systems

Data Warehousing - Concepts 1

Reference: Connolly & Begg, Chapter 31 - elements are © Pearson, 2009

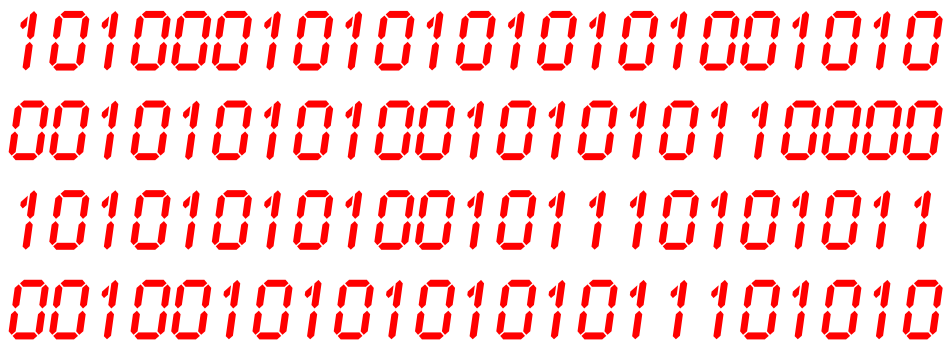
In this video ...

→ a brief history of data

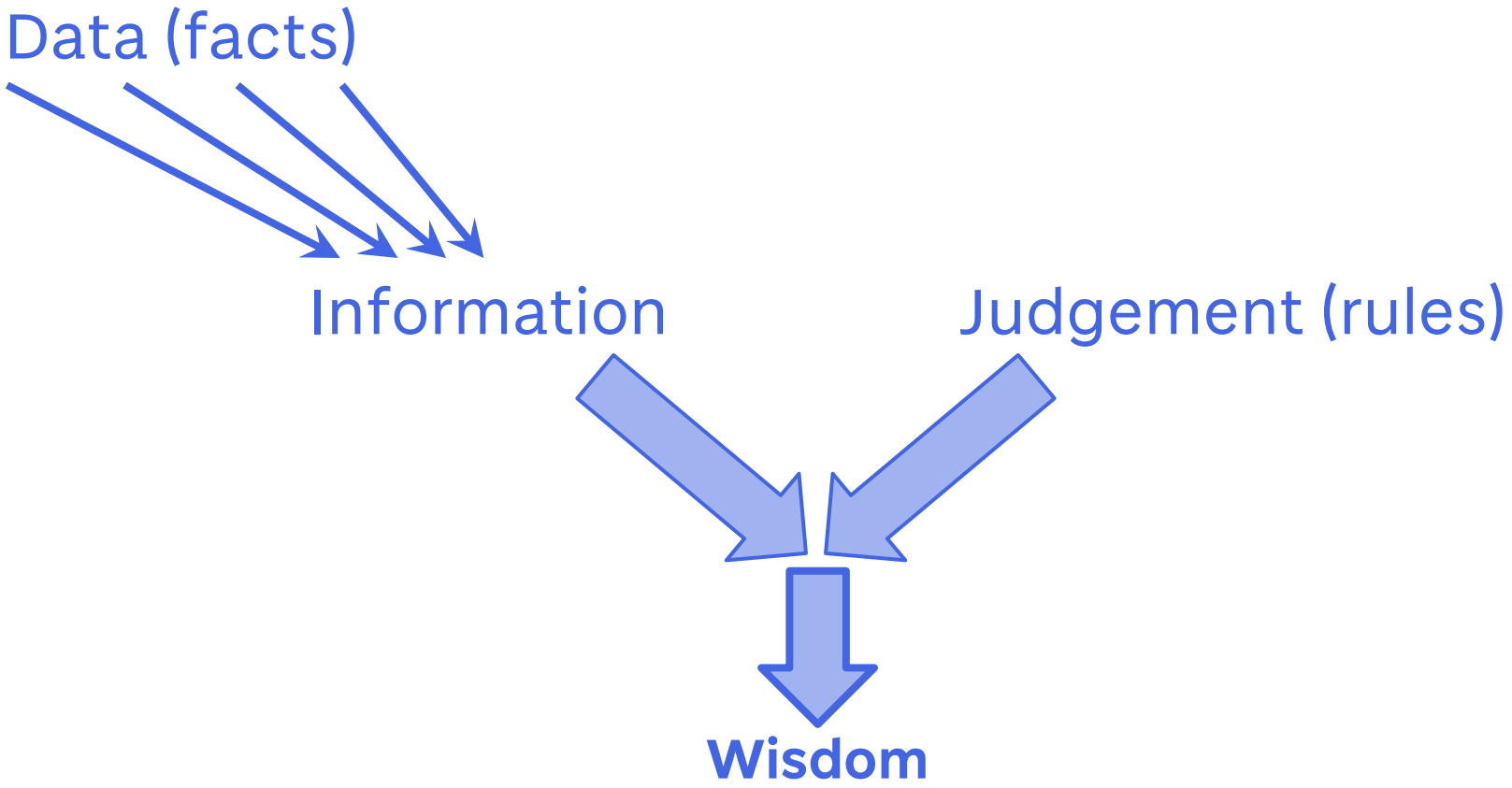
→ the future of data ➔ knowledge

→ introduction to data warehousing

Data:
The Story So Far



“In the Beginning … were the Data”



“Out of the Strong came forth Sweetness”

Data – the story so far

Databases considered so far have been:

- → (comparatively) small and finite
  - → historically, storage in a computer system was limited, so we stored the minimum necessary
- → simple data / facts (numbers, text)
- → designed to hold *specific* data + queries
- → designed to model *the present*
  - → older facts are overwritten by newer facts
- → typically focussed on *transaction processing*
  - → if done "live" – called OnLine Transaction Processing (OLTP)

Data – the future

The trend is towards:

→ collecting data on everything

→ and keeping it all (*nothing* is overwritten)!

→ pulling in data from other external databases

→ use of complex data types (images, video, sounds) which are now practical … and also analysis of their content

The desire is to:

→ find *unknown* patterns (wisdom) in the data

→ use this wisdom to *predict future events*

"Big Data"

Hence we need:

→ massive data warehouses:

  → new data are being added all the time

But … so much data that it's hard to find the knowledge within it:

→ we need tools which can use the data warehouse to extract the wisdom:

  → collectively known as Business Intelligence:

    → processing lots of data

    → guided searching to look in the right places

"Big Data" - caveat

N.B. What is referred to as "big data" is not *all* about the *quantity* of data – the term specifically refers to:

→ large quantities of data, and …

→ data drawn from different sources, and …

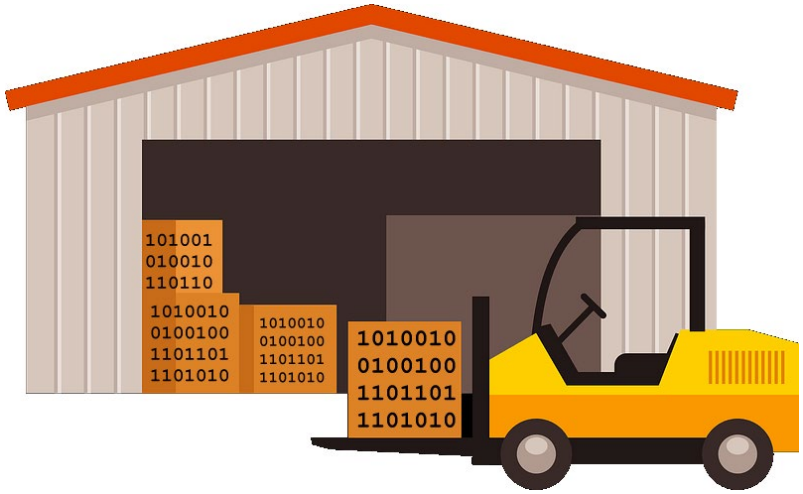
→ integrating the information from various sources, and …

→ extracting information (wisdom) from the integrated sources to get "added value"

Business Intelligence

- → OnLine Transaction Processing (OLTP) giving way to OnLine Analytical Processing (OLAP)
- → New ways to handle "multi-dimensional" data
- → Querying databases when we *don't know what we're looking for*
- → Data mining – looking for *hidden patterns* inside the mass of data
- → Visualising the data to help make sense of it
- → BI is complex … but offers massive dividends
- → Big Data / BI is a new domain … still evolving

Data Warehousing Concepts



Objectives

- How data warehousing has evolved
- The main concepts and benefits associated with data warehousing
- How online transaction processing (OLTP) systems differ from data warehouses
- The problems associated with data warehousing
- The architecture and main components of a data warehouse

Objectives

- → The tools associated with data warehousing
- → The main requirements for a data warehouse DBMS and the importance of managing meta data
- → The concept of the data mart and the main reasons for implementing a data mart

The Evolution of Data Warehousing

- Since the 1970s, organisations have gained competitive advantage through systems that automate business processes to offer more efficient and cost-effective services to the customer
- This resulted in accumulation of growing amounts of data in operational databases

The Evolution of Data Warehousing

- → Organisations now focus on ways to use operational data to support decision-making, as a means of gaining competitive advantage … however, operational systems were *never designed* to support such business activities
- → Businesses typically have numerous operational systems with *overlapping* and sometimes *contradictory* definitions – these are often legacy systems that *evolved* with the business (they were not part of a "master plan")

The Evolution of Data Warehousing

- → Organisations need to turn their archives of data into a source of knowledge, so that a single integrated / consolidated view of the organisation's data is presented to the user
- → The data warehouse (DW)* is a widely-adopted solution to meet the requirements of a system capable of supporting decision-making, receiving data from multiple operational data sources

* you may also see Enterprise Data Warehouse (EDW) – *same* as DW

Data Warehousing Concepts

A data warehouse is:

→"a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process"

Inmon (1993)

Subject-oriented Data

- The warehouse is organised around the major subjects of the enterprise (e.g. customers, products, and sales) rather than the major application areas (e.g. customer invoicing, stock control, and product sales)

- This is reflected in the need to store decision-support data rather than application-oriented data

Integrated Data

- The data warehouse integrates corporate application-oriented data from *different* source systems
- These sources often contain data that is *inconsistent*
- The integrated data sources must be *made consistent* to present a *unified* view of the data to the users

Time-variant Data

→ Data in the warehouse is only accurate and valid *at some point in time* or over some time interval (which needs stored too)

→ Time-variance is also shown in the extended time that the data is held, the implicit or explicit association of *time* with all data, and the fact that the data represents a *series of snapshots*

Non-volatile Data

- New data is always added as a supplement to the database, rather than a replacement (no overwriting)
- Data in the warehouse is not normally updated in real-time (RT), but is refreshed from operational systems on a *regular basis*:
  - although there is an emerging trend *towards* RT or near-RT DWs
  - often, the *most recent* data is the *most valuable*

Oh no!

Data warehouses are typically **based on** relational databases, and much of the data will typically be **sourced from** existing relational databases, so all of our RDB experience **remains valid**!

PHEW!

Data Warehouse Queries

- The type of queries that a data warehouse is expected to answer ranges from the *relatively simple* to the *highly complex* and is dependent on the type of end-user access tools that are used

- End-user access tools include:
  - traditional reporting and query
  - OnLine Analytical Processing (OLAP)
  - data mining

Sample Data Warehouse Queries

- What was the total revenue for Scottish branches in Q3 of 2018?
- What was the total revenue for property sales for each type of property in Great Britain in 2018?
- What are the three most popular areas in each city for the renting of property in 2018 and how does this compare with the figures for the previous two years?
- What is the monthly revenue for property sales at each branch office, compared with rolling 12-monthly prior figures?
- Which type of property sells for prices above the average selling price for properties in the main cities of Great Britain and how does this correlate to demographic data?
- What is the relationship between the total annual revenue generated by each branch office and the total number of sales staff assigned to each branch office?

Sample Data Warehouse Queries

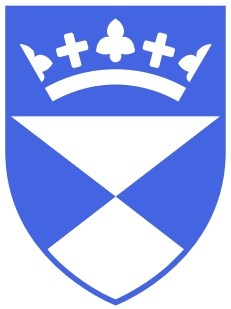
- How do we maximise conversion of applicants to students?
- When is the best time to put the Christmas stock on display?
- What adverts should we show to a specific customer to maximise the chance of a new purchase?
- Can we predict a heart attack based on current medical data?
- Which people are likely to buy our product if we mail them a coupon?
- Which people are likely to buy our product if we cold call them?
- When is the best time of day to call customers?
- Is my car using more fuel than it should?
- Which of these images contains a ship?
- In the ten minutes before the bomb went off, which mobile phone transmissions included the word “bomb”?

Summary

We have seen:

→the reasons for data warehousing

→the main features of a data warehouse



University of Dundee

irmurray@dundee.ac.uk