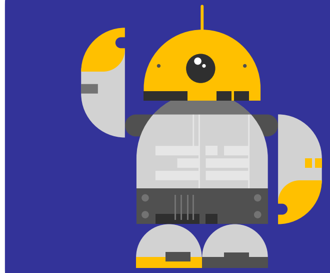


Database Systems

ESOF-3675-WA

Dr. T. Akilan

takilan@lakeheadu.ca

Class Logistics: Schedule and Important Dates

Class Name and Title	Days of Week	Start Time	End Time	Start Date	End Date
ESOF-3675-WA: Database Systems	TTH	04:00PM	05:30PM	Tuesday, January 12, 2021	Tuesday, April 13, 2021
ESOF-3675L-W1	F	08:30AM	10:00AM	Friday, January 22, 2021	Friday, April 9, 2021

- Final Date to Register (Add)–Friday, January 22, 2021
- Reading Week - February 15 to 19, 2021 → **Midterm exam: Feb. 25, 2021 (Thursday)**
- Final Date to Withdraw (Drop)–Friday, March 12, 2021
- Examination Period - Friday April 16 - Sunday April 25, 2021

- All the materials for lectures, labs, assignments, projects, etc. will be posted on the main course website

Class Logistics: Course Outline

- Course outline and evaluation policy
- Project detail
- Graduate assistant:
  - Eduardo R. - edreis@lakeheadu.ca

This Lecture

- Introduction to data mining
- Data
- Database
- Data warehouse
- Data mining challenges
- Applications
- Summary
- Pop quiz

Why Database Systems and Data Mining?

- **Information Age:** there is an explosive growth of data from terabytes to petabytes
  - ○ **Data collection and data availability**
    - ✓ Automated data collection tools, database systems, Web, computerized society
  - ○ **Major sources of abundant data**
    - ✓ Business: Web, e-commerce, transactions, stocks, …
    - ✓ Science: Remote sensing, bioinformatics, scientific simulation, …
    - ✓ Society and everyone: news, digital cameras, YouTube

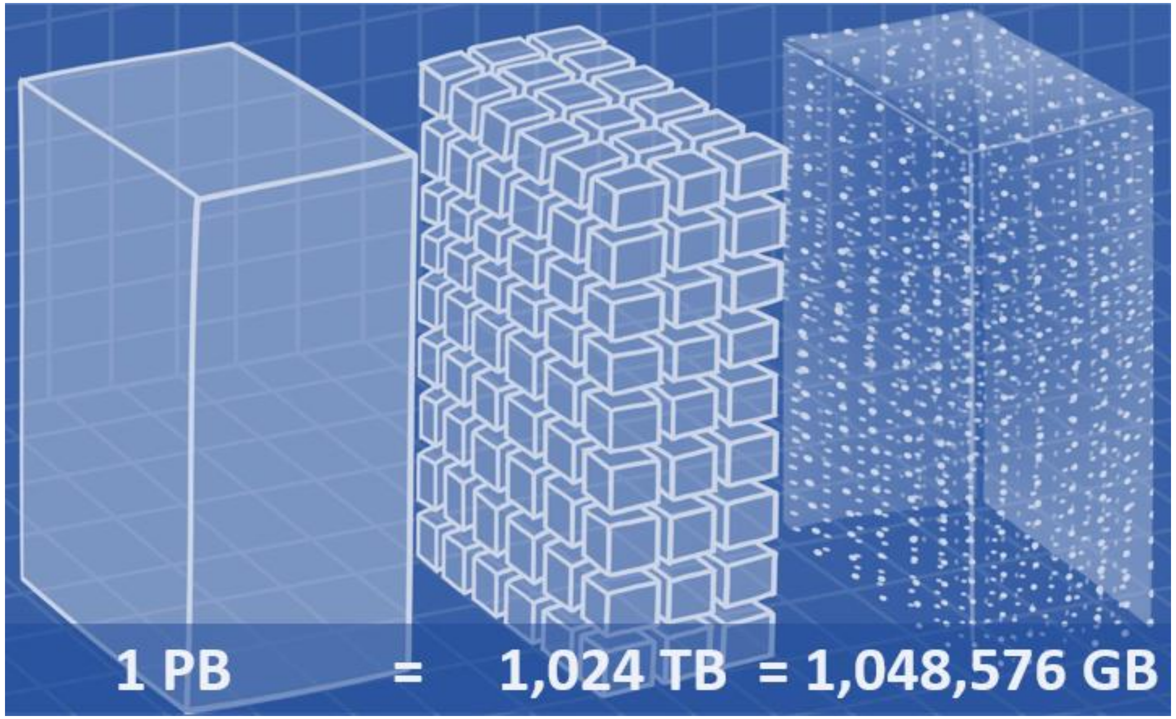


Image Source: https://www.lifewire.com/

Why Database Systems and Data Mining?

- **Issue:** We are **drowning in data** but **starving for knowledge**!

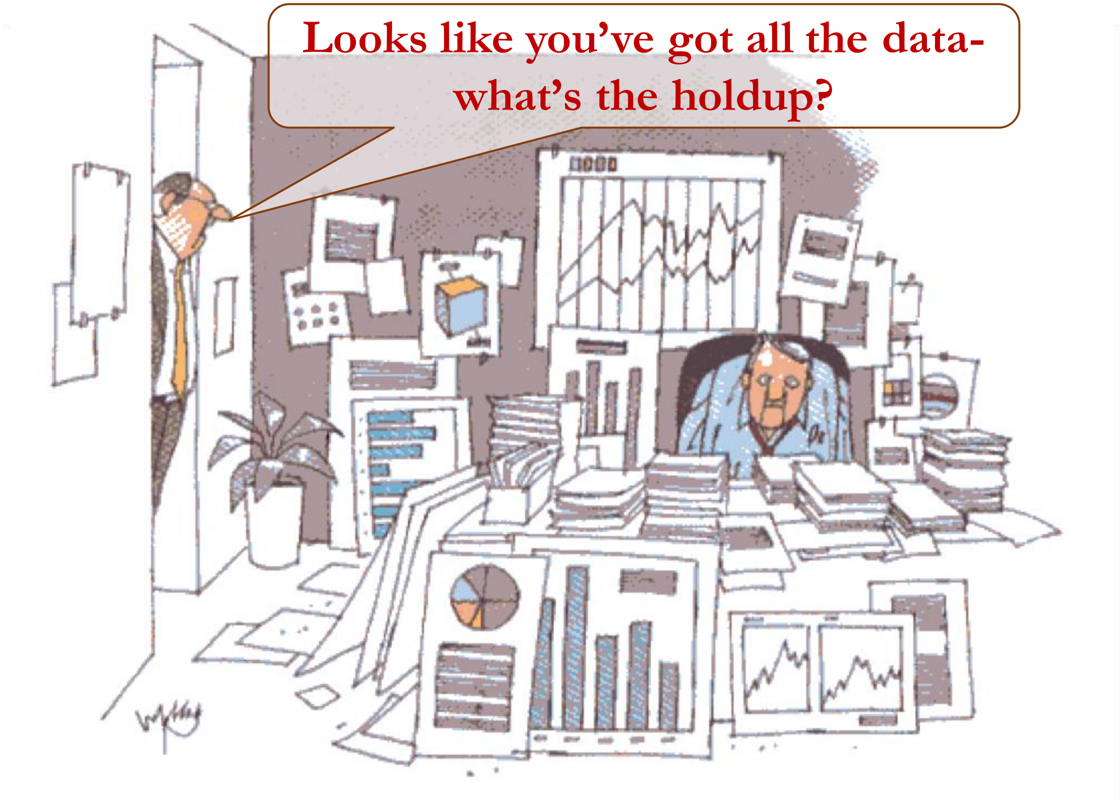


Image Source: David Harbaugh, Harvard Business Review

🗄 VS 💡

- **Data** is just an **accumulation** of **facts**
  - People buy cakes for birthdays
  - People buy cakes for weddings
  - People do not buy cakes for funerals
  - People do not buy cakes for getting fired from job

Are we really like to know this facts?

- What we really care about is **knowledge**
  - People buy cakes for joyful events

- Getting knowledge from facts/**knowledge discovery** from data (**KDD**) is a challenge
  - **Extraction of** interesting, i.e., non-trivial, implicit, previously unknown and potentially useful **patterns** from huge amount of data.

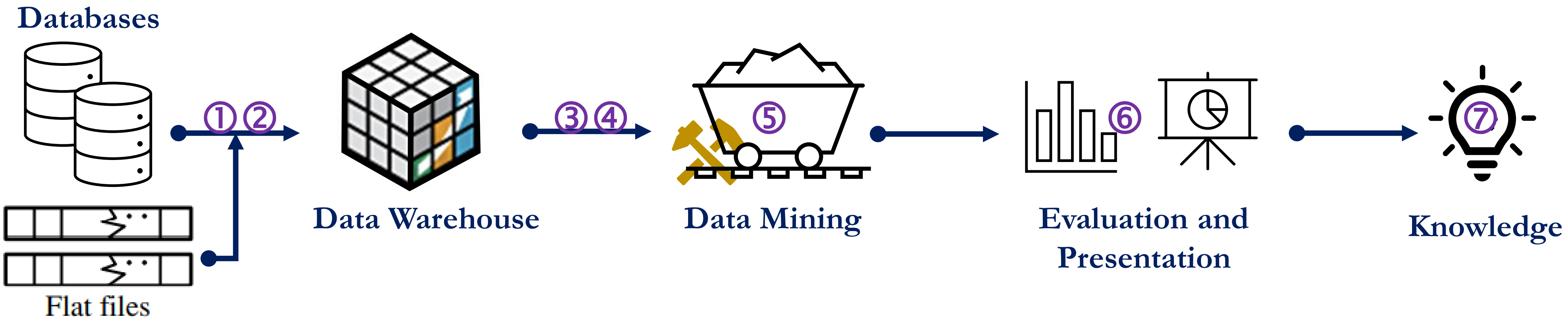
Data vs. Knowledge Cont.

- **Knowledge**: Interesting (non-trivial, implicit, previously unknown and potentially useful) patterns from huge amount of data
  - Non-Trivial
    - ✓ Does not simply restate data
  - Implicit
    - ✓ It is actually supported by the data/ evidence
  - Previously unknown
    - ✓ Don't rediscover things that we already know. However, we can refine the knowledge for more precision.
  - Potentially useful
    - ✓ Allows us to explain something, make predictions, or decisions.

Why Database Systems and Data Mining?

- **Recall:** We are **drowning in data** but **starving for knowledge**!
- **Solution:** "Necessity is the mother of invention"—build automated analysis of massive data sets
- **Important tools:** DBS and data mining are the powerful and robust tools to automatically uncover valuable information from the tremendous amounts of data and to transform such data into organized knowledge.
- **Job demand:** Experts' predict that from 2020 to 2026 the demand for data analytics, data science, data mining, and data related job will grow to 11.5 Million jobs according to U.S. Bureau of Labor Statistics.

Steps in Knowledge Discovery from Data (KDD)



1. Data **cleaning** (to remove noise and inconsistent data),
2. Data **integration** (where multiple data sources may be combined)
3. Data **selection** (where data relevant to the analysis task are retrieved from the database)
4. Data **transformation** (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
5. Data **mining** (an essential process where intelligent methods are applied to extract data patterns)
6. **Pattern evaluation** (to identify the truly interesting patterns representing knowledge based on interestingness measures)
7. **Knowledge presentation** (where visualization and knowledge representation techniques are used to present mined knowledge to users