18CS54- Data Mining (DM)

Lecture 4 : KDD Process, Motivating challenges

Dr. Vani Vasudevan

OUTLINE

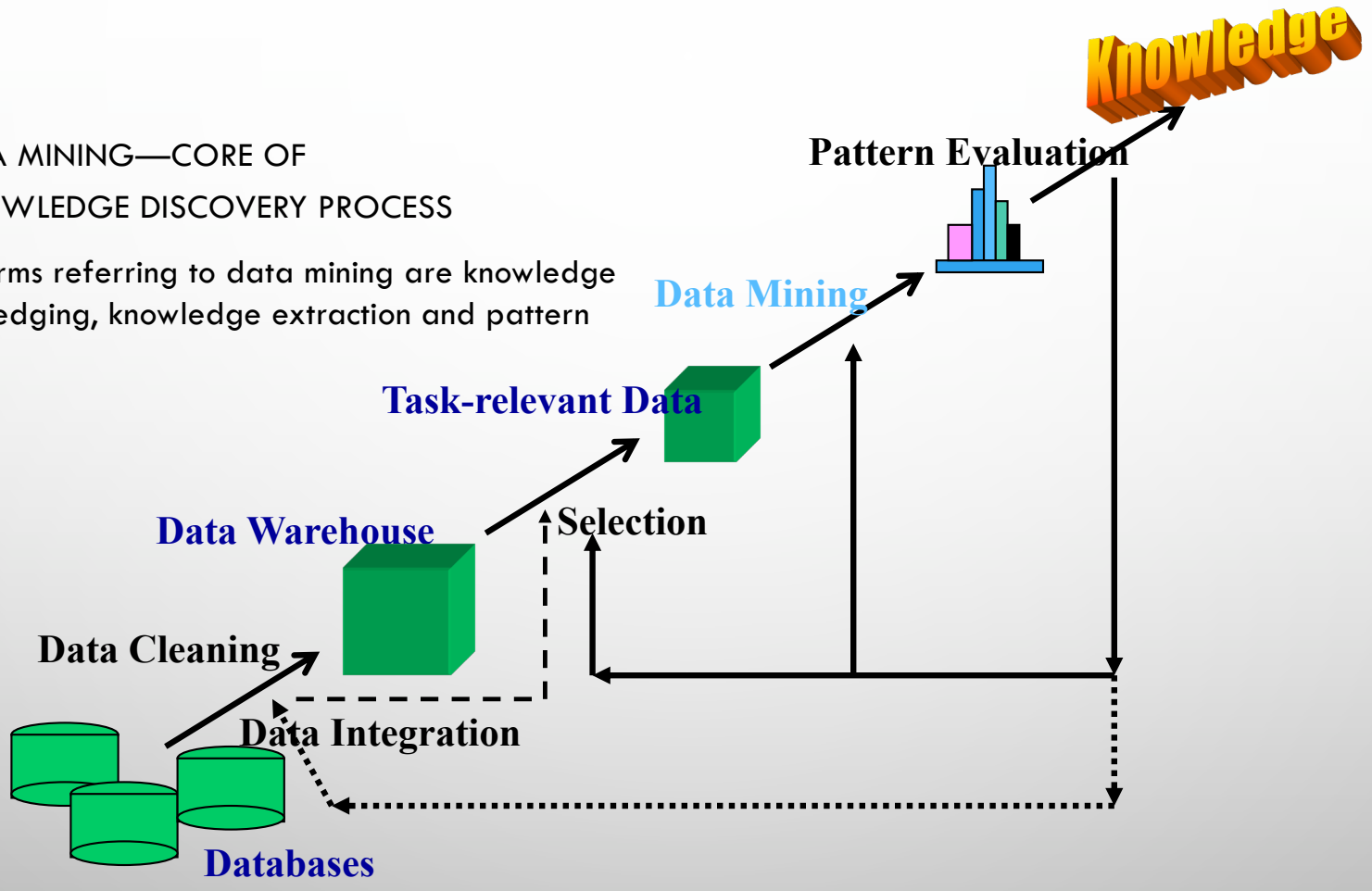
- MOTIVATION: WHY DATA MINING?
- WHAT IS DATA MINING?
- DATA MINING TASKS…
- DATA MINING : A KDD PROCESS
- DATA MINING: ON WHAT KINDS OF DATA?
- CHALLENGES OF DATA MINING

Source : Data Mining: Concepts and Techniques, Han & Kamber

DATA MINING: A KDD PROCESS

- DATA MINING—CORE OF KNOWLEDGE DISCOVERY PROCESS

Other similar terms referring to data mining are knowledge mining, data dredging, knowledge extraction and pattern discovery.



STEPS IN KDD…

- THE KNOWLEDGE DISCOVERY IN DATABASES PROCESS COMPRISES OF A FEW STEPS LEADING FROM RAW DATA COLLECTIONS TO SOME FORM OF NEW KNOWLEDGE. THE ITERATIVE PROCESS CONSISTS OF THE FOLLOWING STEPS:

  LEARNING THE APPLICATION DOMAIN: RELEVANT PRIOR KNOWLEDGE AND GOALS OF APPLICATION

  DATA CLEANING: ALSO KNOWN AS DATA CLEANSING, IT IS A PHASE IN WHICH NOISE DATA AND IRRELEVANT DATA ARE REMOVED FROM THE COLLECTION.

  DATA INTEGRATION: AT THIS STAGE, MULTIPLE DATA SOURCES, OFTEN HETEROGENEOUS, MAY BE COMBINED IN A COMMON SOURCE.

  DATA SELECTION: AT THIS STEP, THE DATA RELEVANT TO THE ANALYSIS IS DECIDED ON AND RETRIEVED FROM THE DATA COLLECTION.

  DATA REDUCTION & TRANSFORMATION: ALSO KNOWN AS DATA CONSOLIDATION, IT IS A PHASE IN WHICH USEFUL FEATURES ARE IDENTIFIED, DIMENSIONALITY/VARIABLE REDUCTION IS CARRIED OUT. THE SELECTED DATA IS TRANSFORMED INTO FORMS APPROPRIATE FOR THE MINING PROCEDURE.

STEPS IN KDD…

DATA MINING: IT IS THE CRUCIAL STEP IN WHICH CLEVER TECHNIQUES ARE APPLIED TO EXTRACT PATTERNS POTENTIALLY USEFUL.

Choosing functions of data mining

summarization, classification, regression, association, clustering.

Choosing the mining algorithm(s)

Data mining: search for patterns of interest

1. **PATTERN EVALUATION**: IN THIS STEP, STRICTLY INTERESTING PATTERNS REPRESENTING KNOWLEDGE ARE IDENTIFIED BASED ON GIVEN MEASURES.
2. **KNOWLEDGE REPRESENTATION**: IS THE FINAL PHASE IN WHICH THE DISCOVERED KNOWLEDGE IS VISUALLY REPRESENTED TO THE USER. THIS ESSENTIAL STEP USES VISUALIZATION TECHNIQUES TO HELP USERS UNDERSTAND AND INTERPRET THE DATA MINING RESULTS.

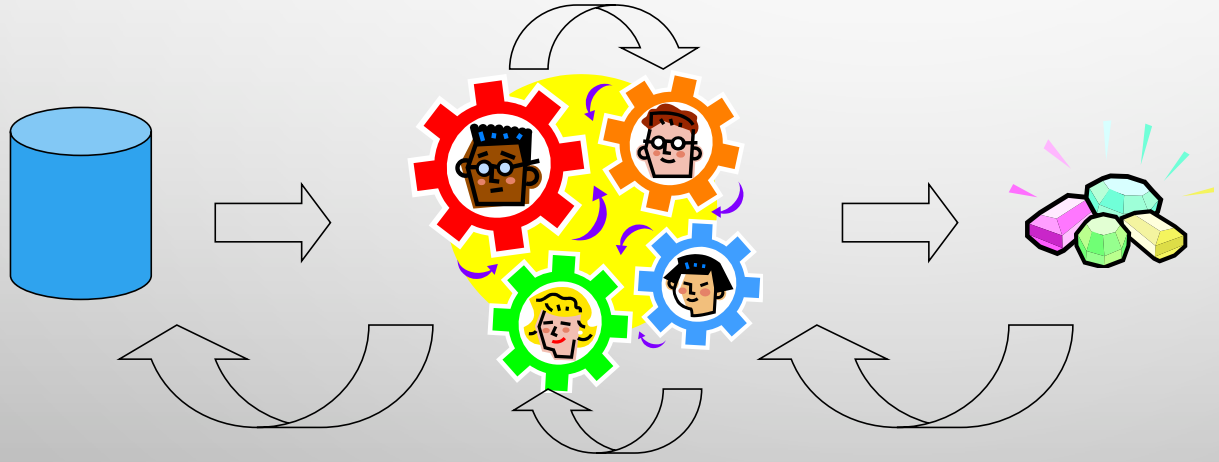
STEPS IN KDD

KDD Steps can be Merged

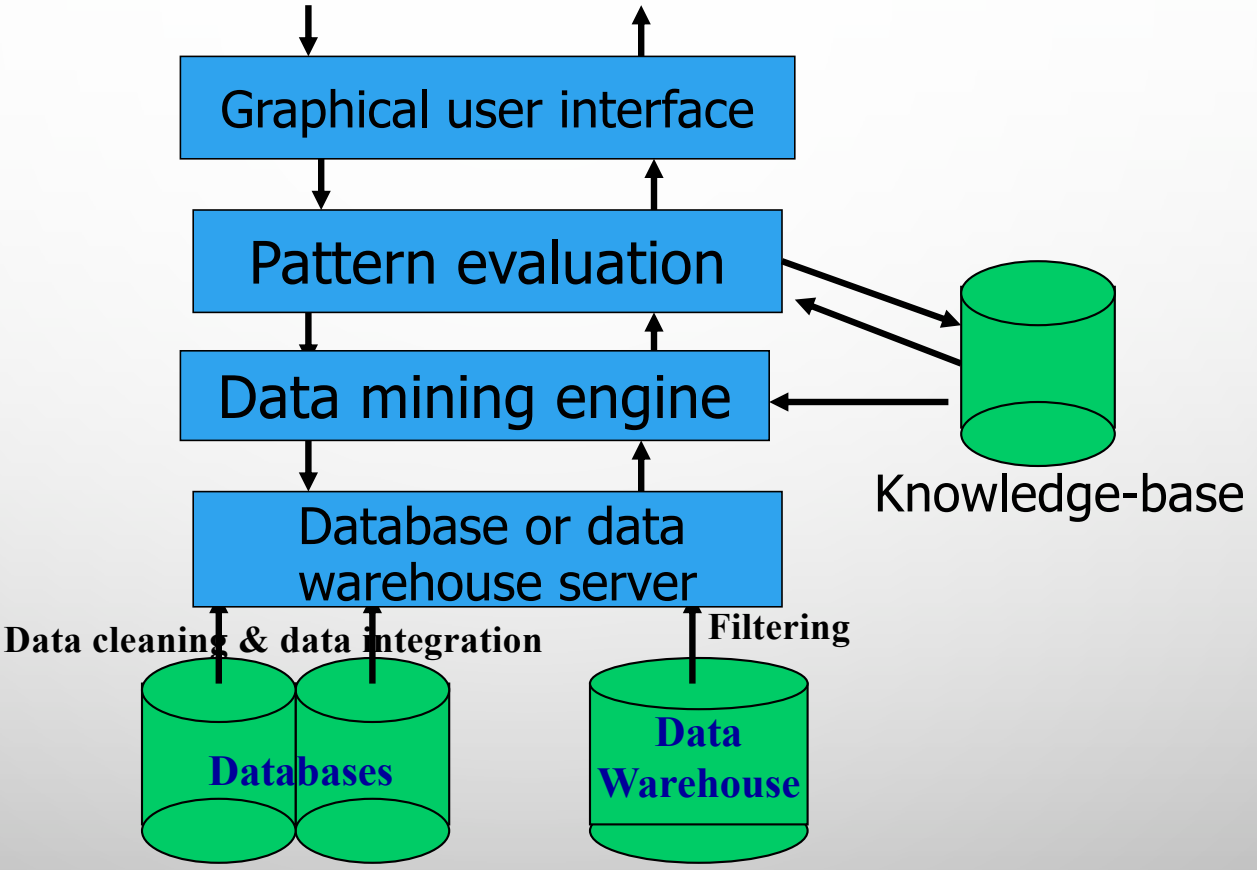
DATA CLEANING + DATA INTEGRATION = DATA PRE-PROCESSING (60% EFFORT)

DATA SELECTION + DATA TRANSFORMATION = DATA CONSOLIDATION

KDD Is an Iterative Process



ARCHITECTURE: TYPICAL DATA MINING SYSTEM



DATA MINING: ON WHAT KINDS OF DATA?

- RELATIONAL DATABASE
- DATA WAREHOUSE
- TRANSACTIONAL DATABASE
- ADVANCED DATABASE AND INFORMATION REPOSITORY
  - SPATIAL AND TEMPORAL DATA
  - TIME-SERIES DATA
  - STREAM DATA
  - MULTIMEDIA DATABASE
  - TEXT DATABASES & WWW

EVALUATION OF KNOWLEDGE

- ARE ALL MINED KNOWLEDGE INTERESTING?
  - ONE CAN MINE TREMENDOUS NUMBER OF "PATTERNS" AND KNOWLEDGE
  - SOME MAY FIT ONLY CERTAIN DIMENSION SPACE (TIME, LOCATION, …)
  - SOME MAY NOT BE REPRESENTATIVE, MAY BE TRANSIENT, …
- EVALUATION OF MINED KNOWLEDGE → DIRECTLY MINE ONLY INTERESTING KNOWLEDGE?
  - DESCRIPTIVE VS. PREDICTIVE
  - COVERAGE
  - TYPICALITY VS. NOVELTY
  - ACCURACY
  - TIMELINESS
  - …

DATA MINING: CLASSIFICATION SCHEMES

- DIFFERENT VIEWS, DIFFERENT CLASSIFICATIONS
  - KINDS OF DATA TO BE MINED
  - KINDS OF KNOWLEDGE TO BE DISCOVERED
  - KINDS OF TECHNIQUES UTILIZED
  - KINDS OF APPLICATIONS ADAPTED

MOTIVATING CHALLENGES

- SCALABILITY
- DIMENSIONALITY
- COMPLEX AND HETEROGENEOUS DATA
- DATA OWNERSHIP AND DISTRIBUTION
- NON-TRADITIONAL ANALYSIS

WHERE TO FIND REFERENCES? DBLP, CITESEER, GOOGLE

- DATA MINING AND KDD (SIGKDD: CDROM)
  - CONFERENCES: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, ETC.
  - JOURNAL: DATA MINING AND KNOWLEDGE DISCOVERY, KDD EXPLORATIONS, ACM TKDD
- DATABASE SYSTEMS (SIGMOD: ACM SIGMOD ANTHOLOGY—CD ROM)
  - CONFERENCES: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
  - JOURNALS: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., INFO. SYS., ETC.
- AI & MACHINE LEARNING
  - CONFERENCES: MACHINE LEARNING (ML), AAAI, IJCAI, COLT (LEARNING THEORY), CVPR, NIPS, ETC.
  - JOURNALS: MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, KNOWLEDGE AND INFORMATION SYSTEMS, IEEE-PAMI, ETC.
- WEB AND IR
  - CONFERENCES: SIGIR, WWW, CIKM, ETC.
  - JOURNALS: WWW: INTERNET AND WEB INFORMATION SYSTEMS,
- STATISTICS
  - CONFERENCES: JOINT STAT. MEETING, ETC.
  - JOURNALS: ANNALS OF STATISTICS, ETC.
- VISUALIZATION
  - CONFERENCE PROCEEDINGS: CHI, ACM-SIGGRAPH, ETC.
  - JOURNALS: IEEE TRANS. VISUALIZATION AND COMPUTER GRAPHICS, ETC.

SUMMARY

- DATA MINING: DISCOVERING INTERESTING PATTERNS FROM LARGE AMOUNTS OF DATA
- A NATURAL EVOLUTION OF DATABASE TECHNOLOGY, IN GREAT DEMAND, WITH WIDE APPLICATIONS
- A KDD PROCESS INCLUDES DATA CLEANING, DATA INTEGRATION, DATA SELECTION, TRANSFORMATION, DATA MINING, PATTERN EVALUATION, AND KNOWLEDGE PRESENTATION
- MINING CAN BE PERFORMED IN A VARIETY OF INFORMATION REPOSITORIES
- DATA MINING TASKS: CHARACTERIZATION, DISCRIMINATION, ASSOCIATION, CLASSIFICATION, CLUSTERING, OUTLIER AND TREND ANALYSIS, ETC.
- DATA MINING SYSTEMS AND ARCHITECTURES
- MOTIVATING CHALLENGES

ADDITIONAL SLIDES…

WHAT IS DATA WAREHOUSE?

A LARGE STORE OF DATA ACCUMULATED FROM A WIDE RANGE OF SOURCES WITHIN A COMPANY AND USED TO GUIDE MANAGEMENT DECISIONS.

WHAT IS DATA SCIENCE?

- **DATA SCIENCE** IS A MULTI-DISCIPLINARY FIELD THAT USES SCIENTIFIC METHODS, PROCESSES, ALGORITHMS AND SYSTEMS TO EXTRACT KNOWLEDGE AND INSIGHTS FROM DATA IN VARIOUS FORMS, BOTH STRUCTURED AND UNSTRUCTURED, SIMILAR TO DATA MINING.

WHAT IS BUSINESS ANALYTICS? …

- **BUSINESS ANALYTICS** IS THE PROCESS OF COLLATING, SORTING, PROCESSING, AND STUDYING **BUSINESS** DATA, AND USING STATISTICAL MODELS AND ITERATIVE METHODOLOGIES TO TRANSFORM DATA INTO **BUSINESS** INSIGHTS.

WHAT IS BUSINESS ANALYTICS?...

- DESCRIPTIVE ANALYTICS – INVOLVE GATHERING AND DESCRIBING DATA
- PREDICTIVE ANALYTICS – USE THE PAST TO PREDICT THE FUTURE
- PRESCRIPTIVE ANALYTICS – SUGGEST A COURSE OF ACTION

WHAT IS BUSINESS ANALYTICS?

WHAT IS BIG DATA?

BIG DATA IS A FIELD THAT TREATS WAYS TO ANALYZE, SYSTEMATICALLY EXTRACT INFORMATION FROM, OR OTHERWISE DEAL WITH DATA SETS THAT ARE TOO LARGE OR COMPLEX TO BE DEALT WITH BY TRADITIONAL DATA-PROCESSING APPLICATION SOFTWARE."

EXAMPLES

- AMAZON
  - 59 MILLION ACTIVE CUSTOMERS
  - MORE THAN 42 TERABYTES OF DATA
- YOUTUBE
  - 100 MILLION VIDEOS WATCHED PER DAY
  - 65,000 VIDEOS ADDED EACH DAY
  - 60% OF ALL VIDEOS WATCHED ONLINE
  - AT LEAST 45 TERABYTES OF VIDEOS
- GOOGLE
  - 91 million searches per day
  - accounts for 50% of all internet searches
  - Virtual profiles of countless number of users
- AT&T
  - 323 terabytes of information
  - 1.9 trillion phone call records