Data Mining - Motivation

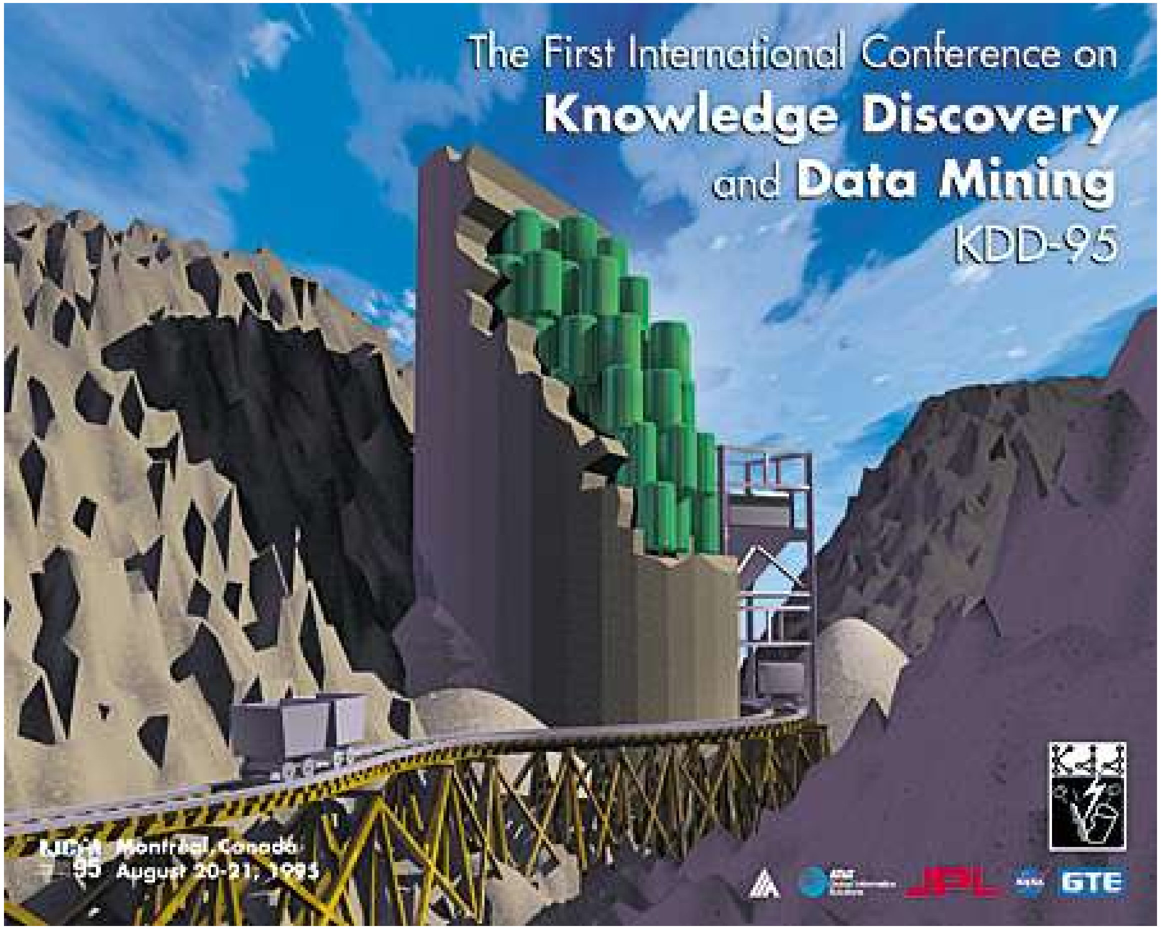
"Computers have promised us a fountain of wisdom but delivered a flood of data."

"It has been estimated that the amount of information in the world doubles every 20 months."

(Frawley, Piatetsky-Shapiro, Matheus, 1992)

Knowledge Discovery in Databases (KDD)

Mining for nuggets of knowledge in mountains of Data.



Definition

- Data Mining is a non-trivial *process* of identifying
  - valid
  - novel
  - potentially useful
  - ultimately understandable

patterns in data.

- *(Fayyad et al. 1996)*

It employs techniques from

- machine learning
- statistics
- databases

Or maybe:

- Data Mining is torturing your database until it confesses.

(Mannila (?))

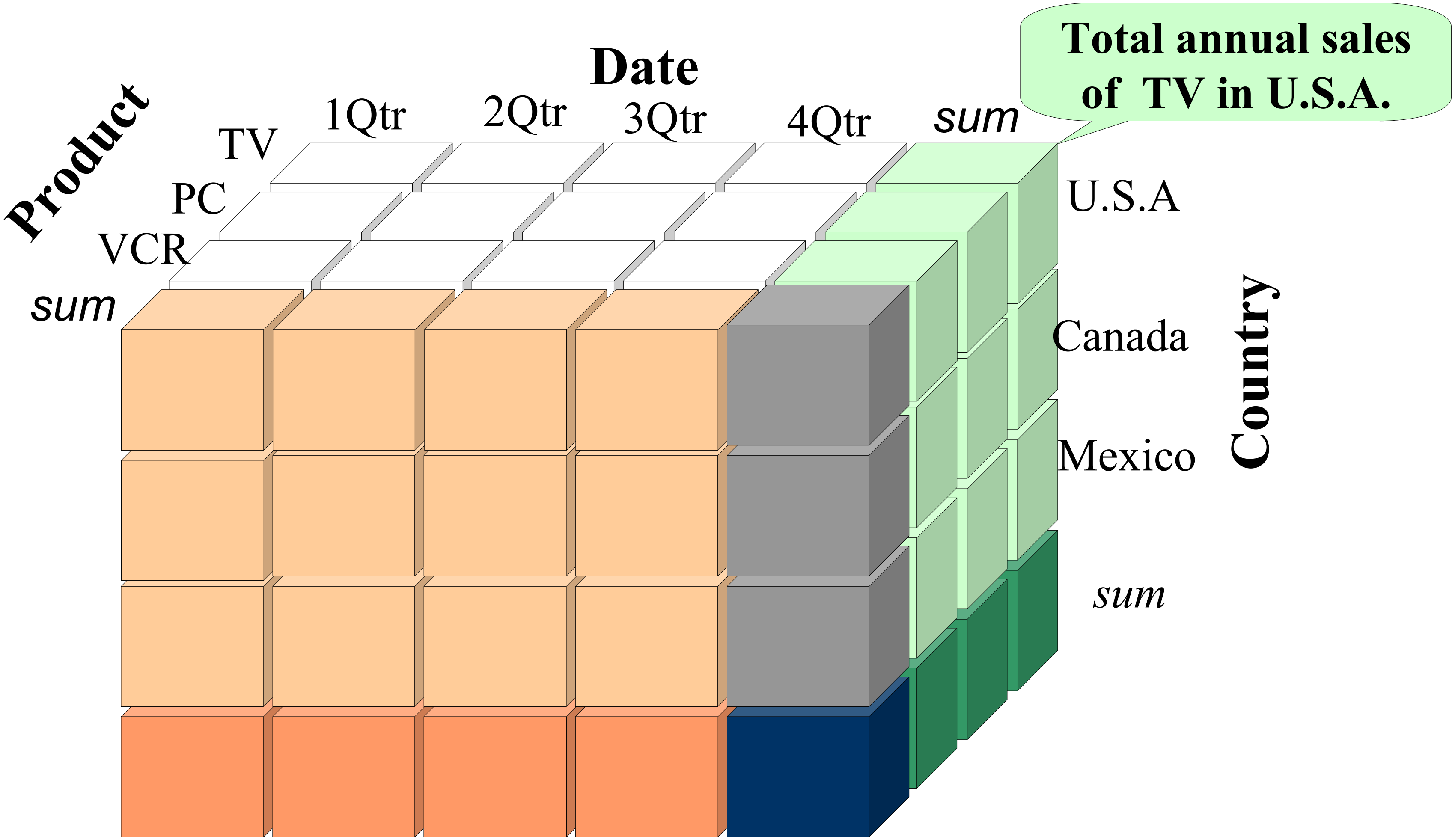
World-Wide Data Growth

- Science
  - satellite monitoring
  - human genome
- Business
  - OLTP (on-line transaction processing)
  - data warehouses
  - e-commerce
- Industry
  - process data
- World-Wide Web

OLTP vs. OLAP

- On-line Transaction Processing (OLTP)
  - Goal: support order processing and billing
  - typically multiple, isolated relational databases
  - even simple queries like "How many items of product X do we sell?" may be hard to answer
- Data Warehouse
  - uniform view to multiple separate data sources
  - fast access across a multitude of dimensions (*data cube*)
  - no need for update in real time
- On-line Analytical Processing (OLAP)
  - Goal: support decision makers
  - allow complex, multi-dimensional queries

Data Cube



Total annual sales of TV in U.S.A.

Taken from Han & Kamber, 2001

OLAP vs. Data Mining

- Verification Model
  - the user needs to verify a hypothesis on the data
  - formulates the query and poses it to the OLAP system
  - Example:
    - Break up sales according to products in the U.S.
    - Break up sales according to products in Canada
    - Any significant difference?
- Discovery Model
  - the system can autonomously propose interesting and novel hypotheses (and verify them on the data)
  - the user formulates the problem
  - Example:
    - Can you find any patterns involving regions, date, products for which the sales differ significantly from the average?

Inductive Databases

(Mannila & Iemielinski, CACM-96)

- make patterns queryable in the database
  - SQL is a query language that generates data sets
  - inductive database should have a language that (also) generates pattern sets
- Examples:
  - "show me all assiciation rules with support > 1% and a minimum confidence > 95% that have *salary* in the head"
  - "show me decision trees with estimated accuracy > 90% that are built on at most 5 of the following 7 attributes"

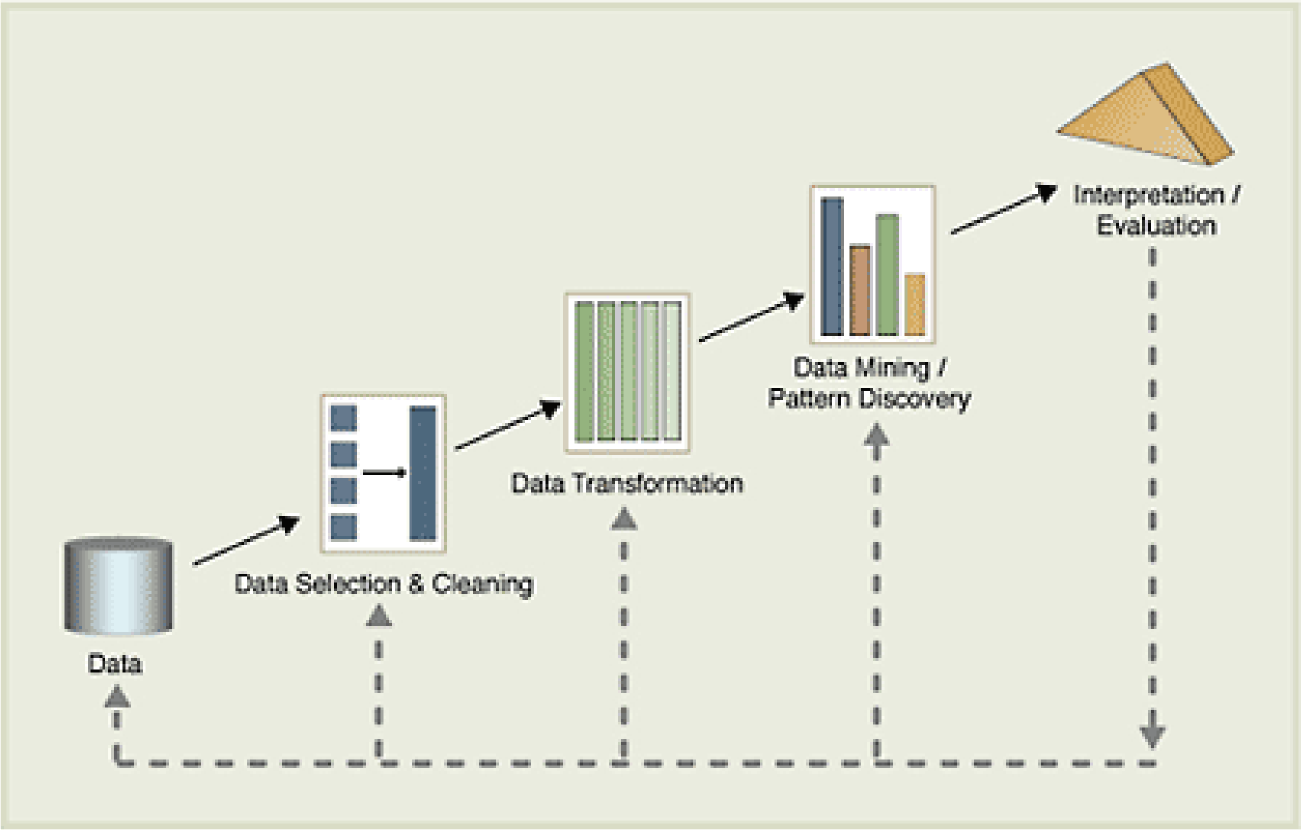
Knowledge Discovery in Databases: Key Steps

Key steps in the Knowledge Discovery cycle:

1. Data Cleaning: remove noise and incosistent data
2. Data Integration: combine multiple data sources
3. Data Selection: select the part of the data that are relevant for the problem
4. Data Transformation: transform the data into a suitable format (e.g., a single table, by summary or aggregation operations)
5. Data Mining: apply machine learning and machine discovery techniques
6. Pattern Evaluation: evaluate whether the found patterns meet the requirements (e.g., interestingness)
7. Knowledge Presentation: present the mined knowledge to the user (e.g., visualization)

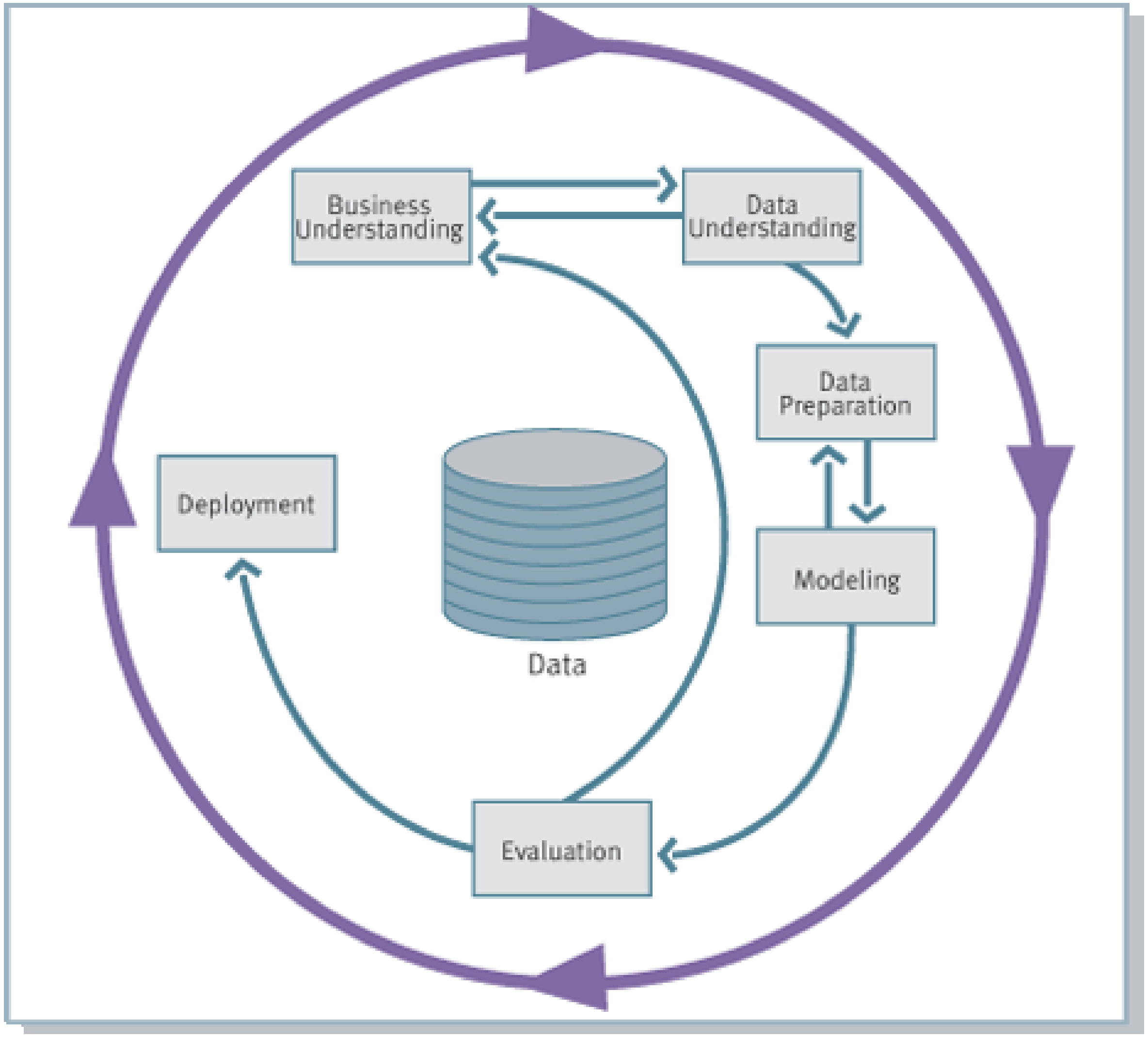
Data Mining is a Process !

The steps are not followed linearly, but in an iterative process.


Interpretation / Evaluation
Data Mining / Pattern Discovery
Data Transformation
Data Selection & Cleaning
Data

Source: http://alg.ncsa.uiuc.edu/tools/docs/d2k/manual/dataMining.html, after Fayyad, Piatetsky-Shapiro, Smyth, 1996

Another Process Model



Source: http://www.crisp-dm.org/

Research Issues

- Techniques for mining different types of knowledge
  - Predictions, Associations, Clusters, Outliers, ...
- Interactive Data Mining Techniques
  - A Human/Computer Team may be more efficient
- Incorporation of Background Knowledge
  - Knowledge about the task helps.
- Data Mining Query Languages
  - Querying patterns instead of querying database entries
- Presentation and Visualization of Results
  - How to explain the results to the CEO?
- Handling Noisy or Incomplete Data
  - Data are typically not neat and tidy, but noisy and messy.
- Pattern Evaluation
  - How can we define interestingness?

(A few) Data Mining Applications

- Business
  - predict credit rating
  - identify customer groups
  - direct marketing
  - market basket analysis
  - recommender systems
  - fraud detection
- Web Mining
  - categorize Web pages (web catalogues)
  - classify E-mail (spam filters)
  - identify Web usage patterns
- Quality control
  - learn to assess quality of products
- Biological/Chemical
  - discover toxicological properties of chemicals
- Game Playing
  - identify common (winning) patterns in game databases
- ....