

Data Mining

CS57300

Purdue University

August 24, 2010

Introduction

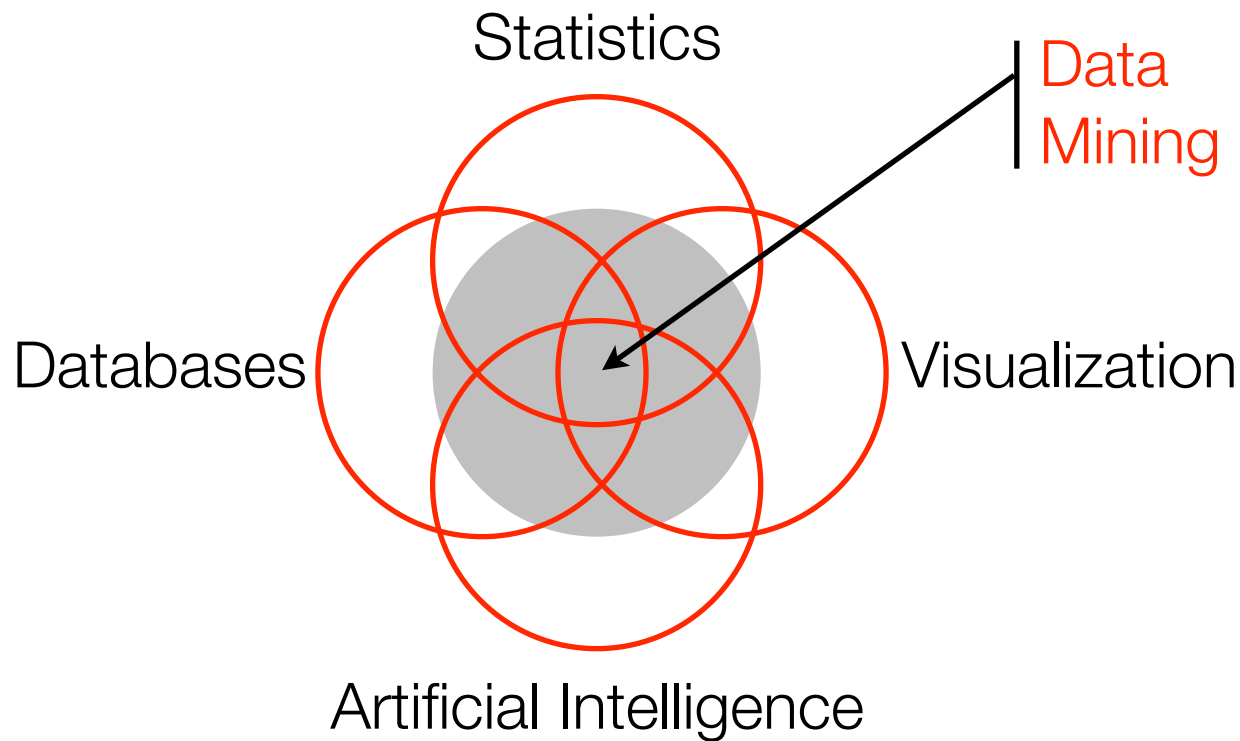
- What is data mining?
- Why now?
- Data mining process
- Example

What is data mining?

“... the non-trivial extraction of implicit, previously unknown, and potentially useful information from data.”
Frawley, Piatetsky-Shapiro, and Matheus (1992)

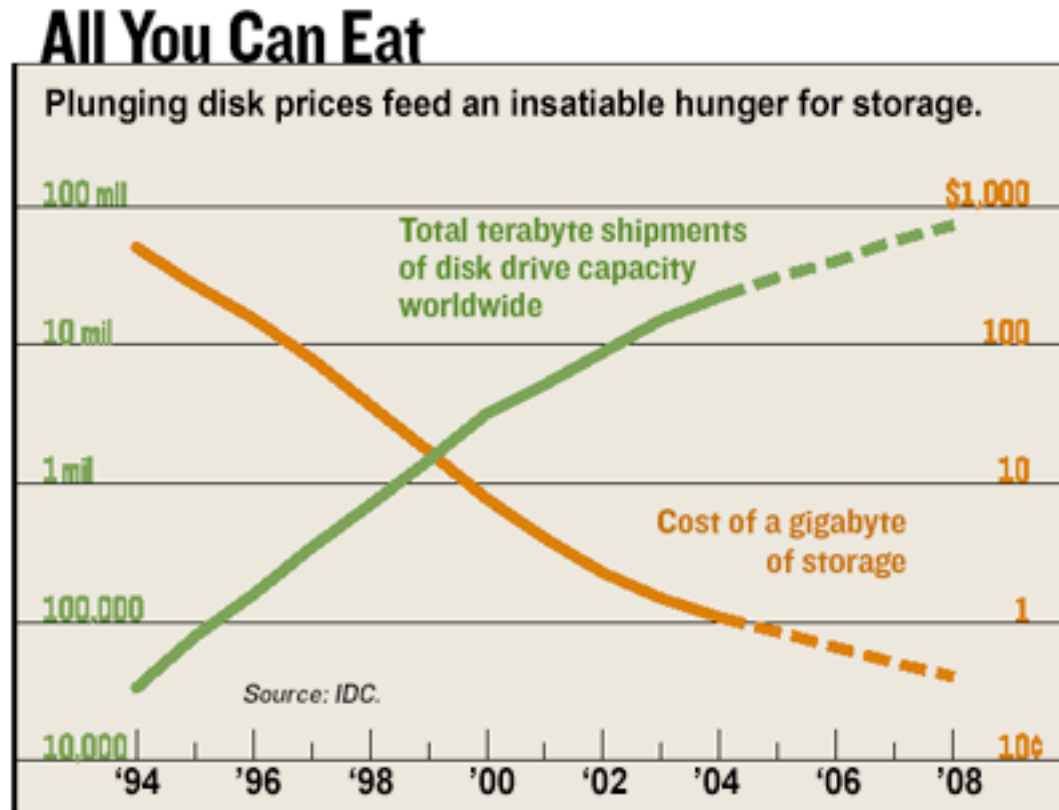
“... a new paradigm that focuses on computerized exploration of large amounts of data and on discovery of relevant and interesting patterns within them.”
Feldman and Dagan (1995)

What is data mining?



Also known as: *knowledge discovery*, exploratory data analysis, applied statistics, machine learning

Why now?

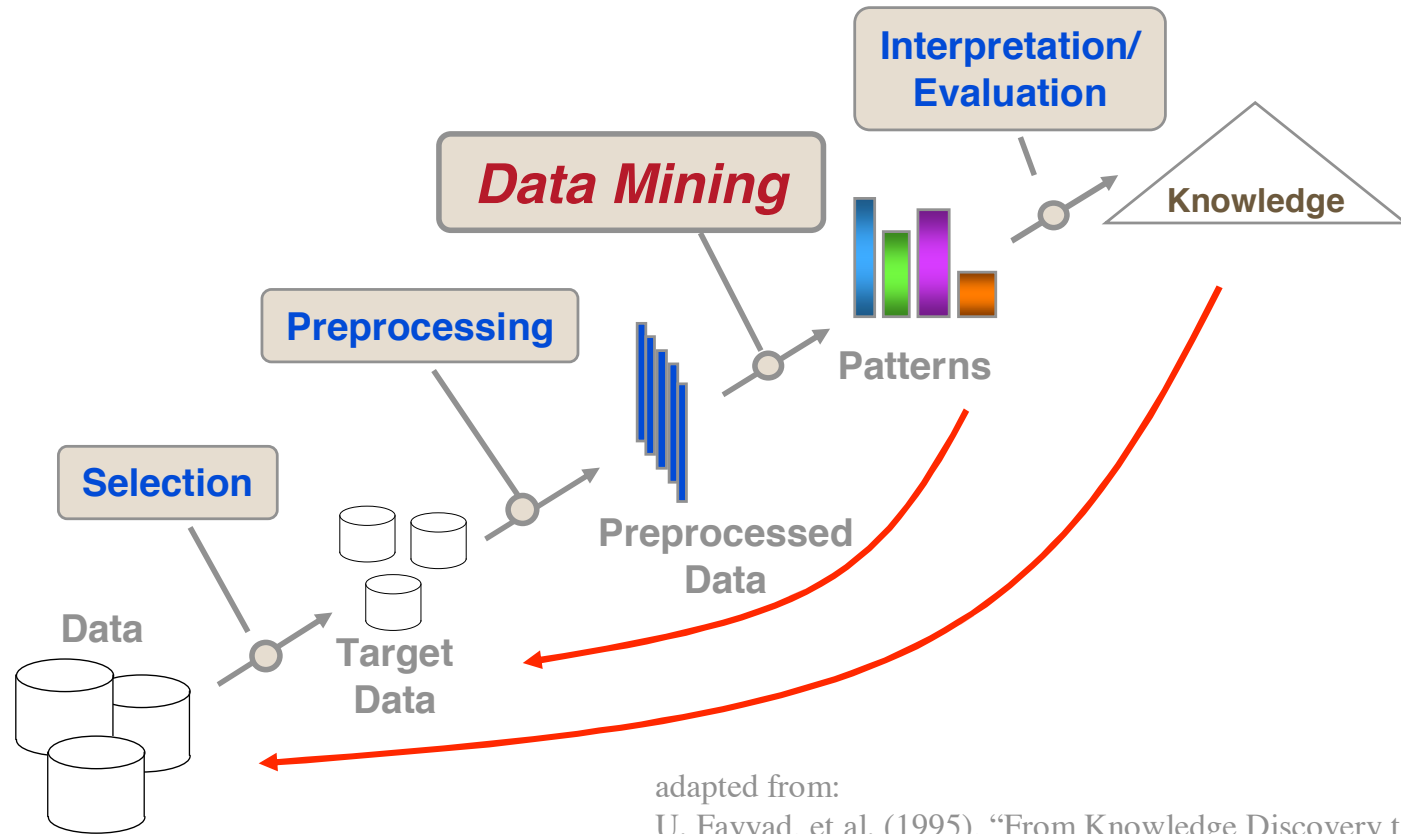


How much information?

Lyman and Varian, UC Berkeley (2003)

- ~5 exabytes of new information stored in 2002
 - 1 Exabyte = 1000 petabytes = 1 mil terabytes = 1 bil gigabytes
- The amount of new information stored has about doubled in the last three years
- Almost 18 exabytes of information flowed through electronic channels in 2002
 - 98% percent of this total is the information sent and received in telephone calls

Data mining process



adapted from:

U. Fayyad, et al. (1995), "From Knowledge Discovery to Data Mining: An Overview," *Advances in Knowledge Discovery and Data Mining*, U. Fayyad et al. (Eds.), AAAI/MIT Press

Data mining process

1. Application setup:

- Acquire relevant domain knowledge
- Assess user goals

2. Data selection

- Choose data sources
- Identify relevant attributes
- Sample data

3. Data preprocessing

- Remove noise or outliers
- Handle missing values
- Account for time or other changes

4. Data transformation

- Find useful features
- Reduce dimensionality

Data mining process

5. Data mining:

- Choose task (e.g., classification, regression, clustering)
- Choose algorithms for learning and inference
- Set parameters
- Apply algorithms to search for patterns of interest

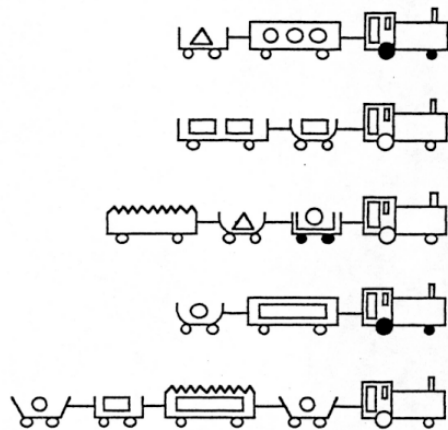
6. Interpretation/evaluation

- Assess accuracy of model/results
- Interpret model for end-users
- Consolidate knowledge

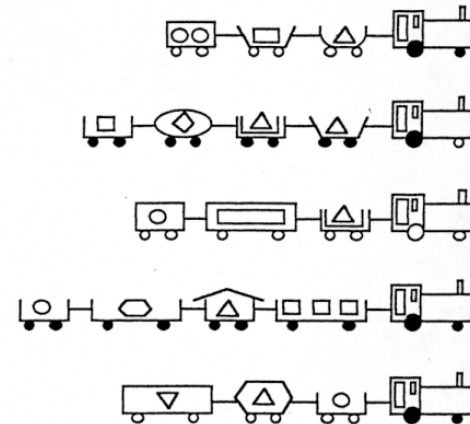
7. Repeat...

Example

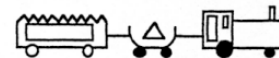
These trains carry toxic chemicals.



These trains do not carry toxic chemicals.



Does this train carry toxic chemicals?



How did you devise rules?

How did you devise rules?

- Look for characteristics of one set but not the other?
- Reject potential rules that didn't cover enough examples?
- Examine several potential rules?
- Consider simple rules first?

This is data mining...

- Data representation: Describe the data
- Task specification: Outline the goal(s)
- Knowledge representation: Describe the rules
- Learning technique:
 - Search: Identify a rule
 - Evaluation function: Estimate confidence
- Inference technique: Apply the rule
- Data mining system: Do above in combination

Complexities

- Data size: vastly larger or changing rapidly
- Data representation: can affect ability to learn and interpret models
- Knowledge representation: needs to capture more subtle forms of probabilistic dependence
- Search space: vastly larger
- Evaluation functions: difficult to assess confidence in model utility

Course overview

Goals

- Identify key elements of data mining systems and the knowledge discovery process
- Understand how algorithmic elements interact
- Recognize various types of data mining tasks
- Familiarity with standard models/algorithms
- Implement and apply basic algorithms
- Understand how to evaluate performance

Topics

- Elements of data mining algorithms
- Statistical basics and background
- Data preparation and exploration
- Predictive modeling
- Methodology, evaluation, theory
- Descriptive modeling
- Pattern mining and anomaly detection
- Current research topics (as time permits)

Logistics

- Time and location: TTh 1:30-2:45, Haas G066
- Instructor: **Jennifer Neville**
neville@cs.purdue.edu, LWSN 2142D, office hours: By appt
- Teaching assistant: **Hongbin Kuang**
hkuang@cs.purdue.edu, LWSN B116H, office hours: TBA
- Webpage: <http://www.cs.purdue.edu/~neville/courses/CS573.html>
- Email list: fall-2010-cs-57300-001@lists.purdue.edu
- Prerequisites: introductory statistics course (e.g., STAT 516), basic programming skills (e.g., CS381, STAT598G)

Registration

- The course is currently at capacity
- Waiting list procedure:
 - Go to CS department website \Rightarrow Courses \Rightarrow Registration
 - Fill out form to request registration (due to lack of space)
 - Students will be admitted on a first-come, first-serve basis as space becomes available

Workload

- Readings

- *Principles of Data Mining*, Hand, Mannila, and Smyth, MIT Press, 2001.

- Additional readings: TBA

- Homeworks

- Six homeworks including written exercises, programming assignments, analysis in R
- Late policy: 10% off per day late, maximum of 5 days; four *extension* days can be applied anytime during semester



Workload (cont.)

- Project (details to follow)
- Exams
 - Midterm exam: covers first half of material
 - Final exam: covers second half of material
 - CS qualifying exam: additional questions covering full semester of material

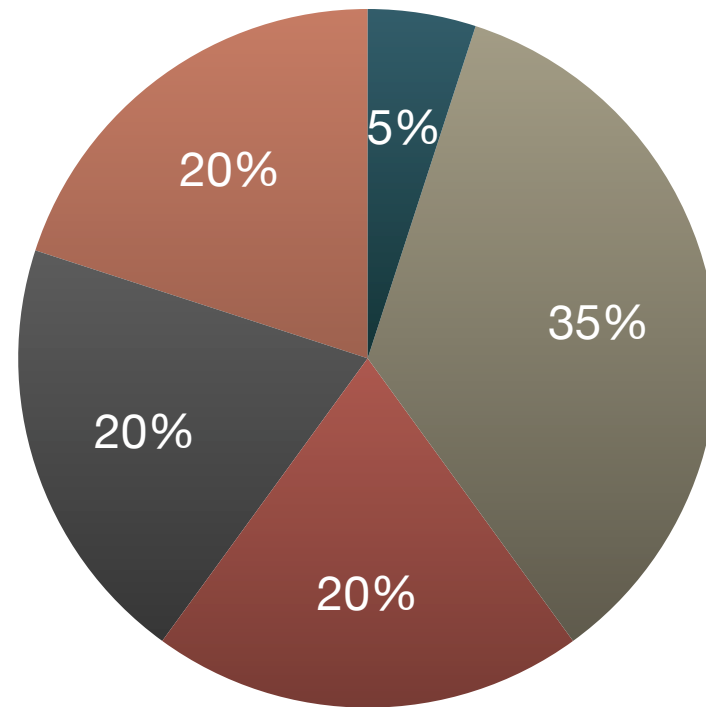
Project

- Goal: Experience the **process** of data mining
 - Data preparation
 - Task definition
 - Feature identification/construction
 - Algorithm selection, parameter tuning
 - Experimental application and evaluation
 - Iterative refinement

Project (cont.)

- Focus of project
 - Choose data, apply ≥ 2 existing methods and compare performance, explore reasons for underlying performance
 - Choose data, formulate novel task, design new method or extend existing methods, evaluate on data
 - *Hypothesis testing*: You must formulate and test at least two specific hypotheses in your project.
- Project proposal: due Sept 30
- Final report: due Dec 9

Grading



● Participation ● Homework ● Project ● Midterm ● Final

Acknowledgements

- Some of the lecture material is drawn from other data mining courses, which use PDM:
 - Professor David Jensen at UMass Amherst (CS591Y)
 - Professor Lise Getoor at UMaryland College Park (CS828G)

Next class

- Reading: Chapter 1 PDM
- Topic: Elements of data mining algorithms