

Anna Östlin Pagh and Rasmus Pagh  
IT University of Copenhagen

# Advanced Database Technology

April 29, 2004

# DATA MINING

Lecture based on [GUW 20.6], [BrinPage98, sec. 1, 2, 4.2],  
[AgrawalSrikant94], and [Ullman00]

# Today

- Advertisements.
- Data mining
  - Introduction.
  - Data mining in Google.
  - Association rules: A priori algorithm and improvements.
- Course summary and exam information.
- Thesis info.

# Advertisement

- No regular advanced algorithms course F2004.
- Instead: Possibility to do a 12-week project in **randomized algorithms**.
- The project should result in a report surveying (part of) the material, plus an "own part".
- There will be a PhD course in randomized algorithms, with lectures you may choose to follow.
- Spread the word if you know someone who might be interested!

## Advertisement 2

- Teaching assistants needed for "Introduction to databases", fall 2004.
- Please consider applying!

# Data mining: Many flavors

- Statistics.
- Artificial intelligence.
- OLAP/dimensional modeling: Complex aggregation queries over possibly huge data sets ("decision support").
- Here: "Discovery of useful summaries of data".

# Examples of data mining queries

- **Clustering.** Group objects together in clusters of "similar" objects, e.g., customer groups that need different treatment.
- **Association rules.** Find "interesting" correlations in data. Amazon.com: "Customers who bought this book also bought ..."

# Data mining in this lecture

- Google's method for ranking web pages according to how authoritative they are.
- Finding association rules
  - The A-Priori algorithm...
  - and improvements