Discrete Topics in Data Mining Dr. Pauli Miettinen

Discrete Topics in Data Mining
Universität des Saarlandes, Saarbrücken
Winter Semester 2012/13

Introduction

- Advanced course in data mining
  - Assumes IR&DM, Machine learning, or equivalent knowledge
  - Textbooks help with catching up
- Will cover four topics in data mining
  - Emphasis on ideas & intuition, not in implementation
  - Topics are only loosely related (all are data mining)
- Modular structure
  - Fresh restart at the begin of every topic

Course organization

- Lectures: 2 h / week
- No homework meetings
  - No traditional homeworks
- Five essays (one warm-up + one per topic)
  - Require deeper thinking
  - Can require reading material that wasn't covered in the lectures
  - Graded fail/pass/excellent
  - You have 2 weeks for each essay
- Final exam

Requirements

- In order to pass the course, you must
  - get a passing grade from at least four essays (out of five)
  - pass the final exam
    - essays are a prerequisite
- Bonus points:
  - You get 1/3 better grade for each excellent essays
    - From zero to hero with five excellent essays!
  - You still must pass the final exam in order to pass the course

Is this a seminar?

- No!
- You don’t need to present anything
- I will give all the lectures
- We do essays because I want deeper understanding
  - Small, technical questions are not well-suited for this course

Course material

- These slides are **not** comprehensive material
- For each of the specific topics, the related research articles will be made available on the web page
  - Requires a username and password
- For more general introduction and background, textbooks can be used
  - Pang-Ning Tan, Michael Steinbach, Vipin Kumar. *Introduction to Data Mining*, Addison-Wesley, 2006.
  - Jiawei Han, Micheline Kamber, Jian Pei. *Data Mining — Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2011.
  - Mohammed J. Zaki, Wagner Meira Jr. *Fundamentals of Data Mining Algorithms*, manuscript
    - Available on the web page with username and password

About the essays

- The essays must explain the topic in your own words and in your own thoughts
  - I want you to think!
  - Essays don’t have to be 100% technically correct
    - Though they can’t be totally wrong, either
- An excellent essay explains new connections between topics covered in the lectures
  - Shows your own thinking
  - Uses your own words
- A failed essay is plagiarised/doesn’t have your own words/is off-topic/is returned after the DL

More on essays

- Please, use computer to write
  - Bad language will have (indirect) effect
- No strict length limits
  - Content matters, not form
  - Probably about 2–5 A4 pages with 10pt text and 2.5 cm margins
- Normal scientific citation rules apply
- You can use sources outside those covered in lectures
- First essay topics are given today
  - A warm-up essay for calibration

Returning the essays

- You **must** return the essays
  - in PDF format
  - by e-mail (pauli.miettinen@mpi-inf.mpg.de)
  - on time
- Any delay of returning the essay will mean you failed that essay
  - medical conditions might give you an excuse

General schedule

- Each module is three weeks
  - 1st week: introduction to the broad topic
  - 2nd and 3rd week: (typically) two sub-topics from that area
    - A sub-topic: a research article (or few)
    - Sub-topics are related to each other
- Essay topics are given on 3rd week
  - DL two weeks after that

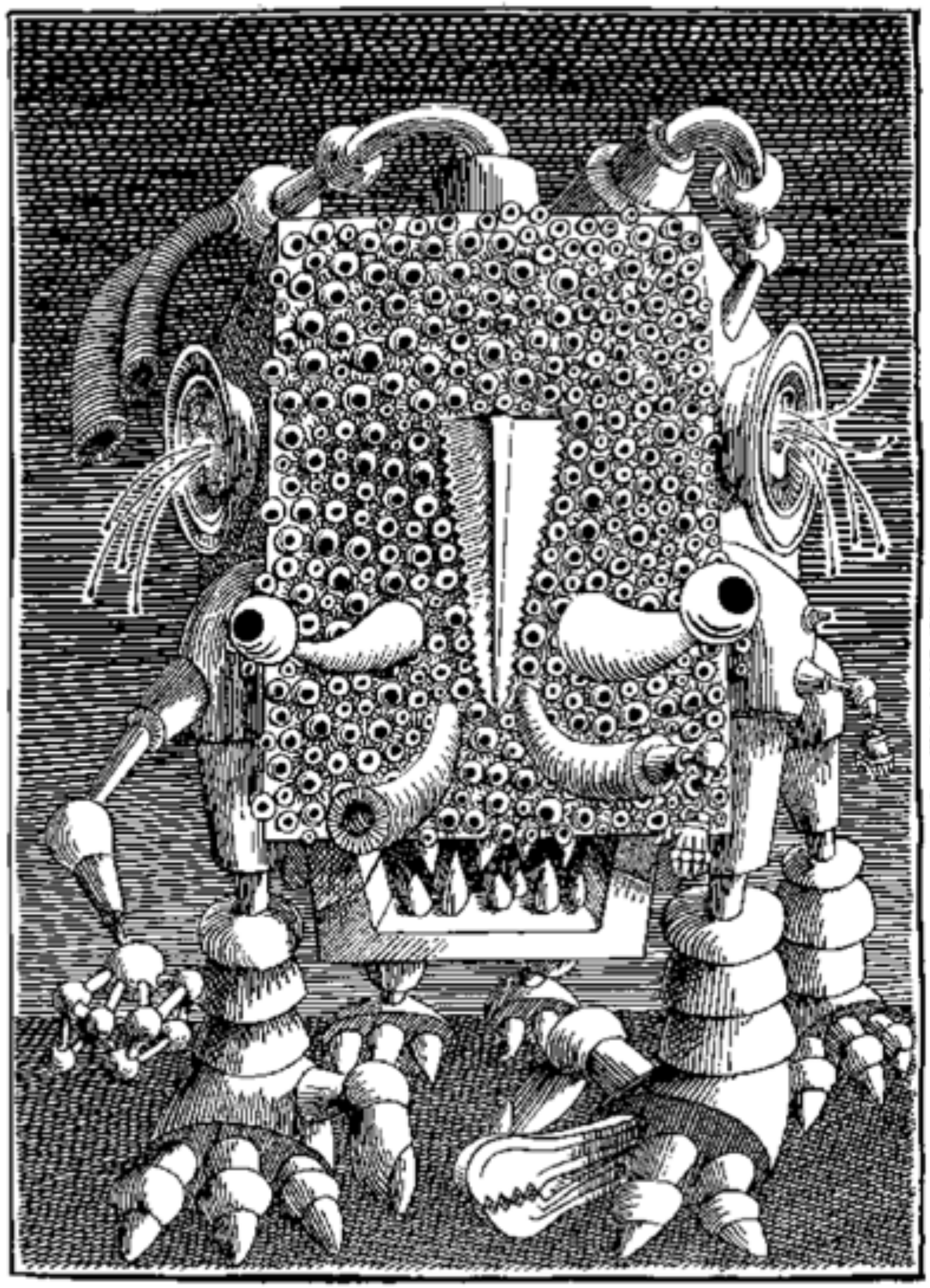
Month	Day	Lecture topic	Essay
October	16	Intro	Warm-up essay
	23	T I intro: Pattern set mining	
	30	T I.1: Tiling	**Warm-up essay DL**
November	6	T I.2	T I essay, w-u feedback
	13	T II intro: Graph mining	
	20	T II.1	**T I essay DL**
	27	T II.2	T II essay, T I feedback
December	4	T III intro: Assessing the significance	
	11	**No lecture**	**T II essay DL**
	18	T III.1	T II essay feedback
	25	**No lecture, Christmas break**	
January	1	**No lecture, Christmas break**	
	8	T III.2	T III essay
	15	T IV intro	
	22	T IV.1	**T III essay DL**
	29	T IV.2	T IV essay, T III feedback
February	5	Wrap-up	
	12		**T IV essay DL**
	??	Exam	

Short Intro to Data Mining

1. What is data mining?
2. Why data mining?
3. Data mining and other sciences
4. Data mining in practice

Data Mining — motivation

What to do with the information you've retrieved?



The ”PHT” Pirate wanted all information of the world. But before he realized most of it was useless, he was already buried under it.

—Stanisław Lem, *The Cyberiad*

Data Mining — definition

Data mining is the process of extracting hidden patterns from data.

—*Wikipedia*

Data mining, in a broad sense, is the set of techniques for analyzing and understanding data.

—Zaki & Meira: *Fundamentals of Data Mining Algorithms*

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.

—Hand, Mannila & Smyth: *Principles of Data Mining*

Data Mining — definition

Data mining, in a broad sense, is **the set of techniques for analyzing and understanding data**.

—Zaki & Meira: *Fundamentals of Data Mining Algorithms*

Data Mining Applications

- Business intelligence
  - What customers buy together?
  - What are the seasonal trends?
  - How to make more money?

- Scientific data analysis
  - What genes cause diseases?
  - What species co-inhabit areas?
  - What happens if average temperature raises?
- And anything else where you have data…
  - Who Barack Obama should persuade to vote him?
  - Is there a problem in International Space Station?

What do You need to do Data Mining

- Data
- Domain knowledge
- Data mining techniques

This course

Data mining's position in sciences

- Data mining uses statistical tools and methods to infer from data
  - Is data mining just fancy name for statistics?
- Data mining uses methods to learn unseen
  - Is data mining just boring name for machine learning?

Is data mining a voodoo science?

Data mining vs. the scientific method

- The scientific method:
  - Form a hypothesis
  - Collect the data
  - Test the hypothesis
- The data mining method:
  - Get the data
  - Select a data mining method that makes sense in your data
  - Apply the method to the data

Selects a "family" of hypotheses

Finds the "valid" hypotheses for the data

The voodoo science

The response from several social scientists has been rather unappreciative along the following lines: “Where is your hypothesis? What you’re doing isn’t science! You’re doing DATA MINING !”

http://andrewgelman.com/2007/08/a_rant_on_the_v/

Data mining vs. statistics

- Statistics provides tools to validate the hypotheses
- Data mining generates the hypotheses
- But data mining uses tools from statistics
  - Toolbox of mathematical methods
  - Validation of results
  - Formalization of methods

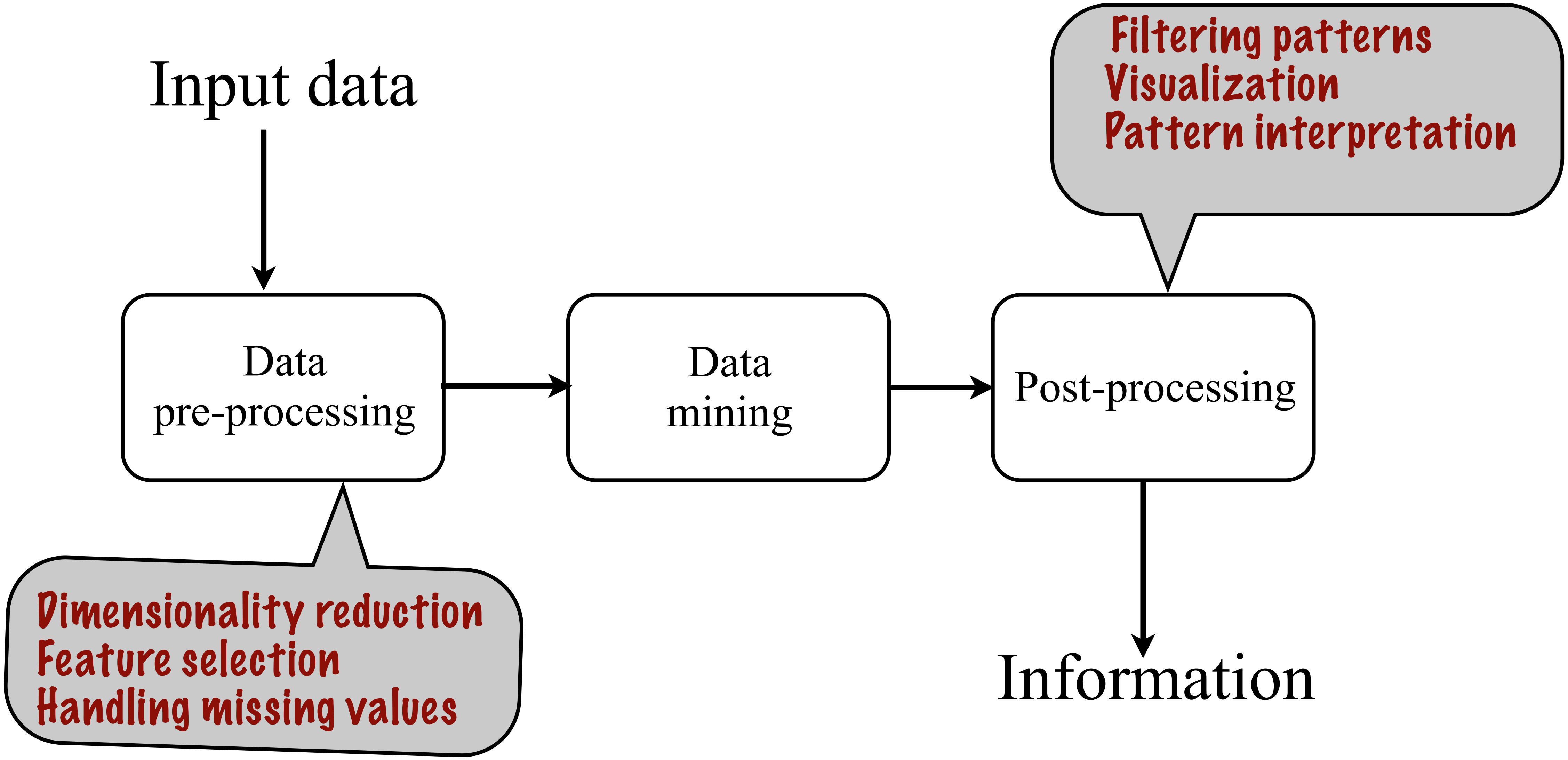
Data mining vs. machine learning

- Data mining uses machine learning methods
  - An application
  - More practical issues
    - Missing values
    - Odd correlations
    - Scalability
    - Domain knowledge
- No clear difference between developing data mining methods vs. developing machine learning methods

Data mining in practice

- Real world is a messy place
  - Real-world data is even messier
  - Data needs pre-processing
- Applications have (hopefully) domain experts
  - Domain knowledge should be incorporated
  - Domain experts should be able to interpret the results
    - Not too many results
    - Post-processing

The KDD process



Input data

Data pre-processing

Data mining

Post-processing

Information

Dimensionality reduction
Feature selection
Handling missing values

Filtering patterns
Visualization
Pattern interpretation

Data pre-processing

- Garbage in, garbage out
- Many issues
  - What to do with missing values
    - Are missing values clearly marked?
  - What's the dimensionality vs. sample size
    - Anyway, which way the observations are?
  - Do some features correlate with each other in an uninteresting way
    - Record ID and class label
  - Is data type suitable for our algorithm
    - Binary, categorical, numerical
  - And many, many more…

Post-processing

- Humans can only interpret so many results
  - Computers are a different thing
- Select top-*k* results
  - What criteria?
- Are the results significant?
  - Statistics
- Are the results meaningful?
  - Domain expert
- Visualization
- Humans are great at finding patterns (even when they don't exist)
  - Computers are a different thing

Leakage

- *Leakage in data mining* refers to the case when prediction algorithm learns from data is should not have access to
  - Problem as the quality is assessed using already-historical test data
  - E.g. INFORMS'10 challenge: predict the value of a stock
    - Exact stock was not revealed
    - But ”future” general stock data was available!
      ⇒ 99% AUC (almost perfect prediction!)
  - More subtle one's exist
    - E.g. removing a crucial feature creates a new type of correlation

Data mining applications

- Data mining is quite commonplace
  - Often not called like that
  - Sometimes something other is meant by data mining: ”An Unethical Econometric practice of massaging and manipulating the data to obtain the desired results” —W.S. Brown “Introducing Econometrics”
- Many sciences are turning into *data-driven sciences*
  - How to deal with all the data obtained?

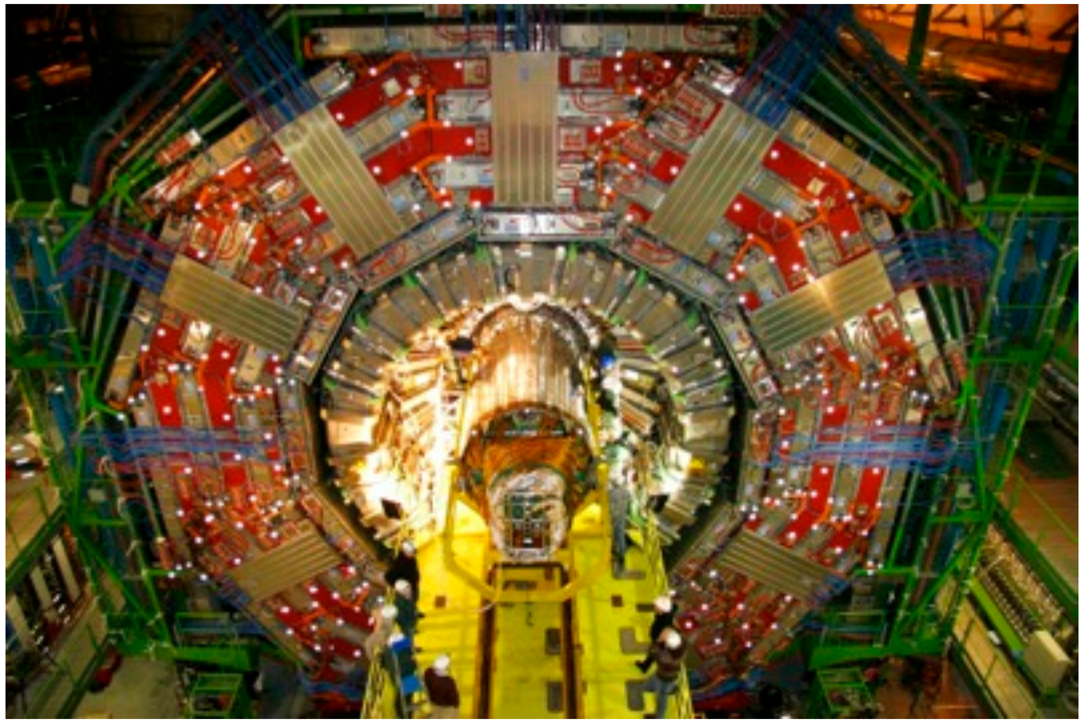


Image: CERN

Data mining and biology

- Genome databanks
  - Identify genes (pattern mining)
  - Identify groups of related genes (graph mining)
- Protein activity
  - Developmental biology (Episode mining)
- Protein-protein interactions
  - Thousands of different proteins
  - Proteins have different roles in different situations in different compartments

Data mining and medicine

- Hospitals
  - Patient diagnoses (decision trees)
  - House, M.D., in a computer (probability estimation)
- Pharmaceutical companies
  - Drug development (much like bioinformatics)
  - Not all drug prototypes can be tested
    - Too many
    - Potentially lethal
    - Learning

Data mining and economy

- Recommender systems
  - Netflix, Amazon, well, anybody
- User segments
  - Clustering
- Machine learning for stock prices
- Part of algorithmic trading
  - Faster than humans
  - Not prone to human errors?
    - But…

Data mining and the Internet

- Social networks (FB, LinkedIn, MySpace, …)
  - Social scientists love!
  - Link prediction
  - Supply recommender systems
- Searching
  - Ad targeting
  - User profiling

Data mining and secret services

- Terrorist profiling/detection

Data mining and privacy

- Very important!
  - Everybody wants to do data mining, nobody wants to be data mined
- Often imposed by laws
  - Medical data, personal information records, …
- *Privacy-preserving data mining*
  - Data provider anonymizes data
  - Data miner does not know (and can't learn) the identities of the data entities
  - Hard to guarantee
    - Leakage!

Preserving privacy in data mining

- Remove sensitive features
  - Can be re-mapped using publicly available data
- *k*-anonymity
  - All released records are similar to at least *k* other records in the released features
    - Homogenous sensitive data can still be learned
- Differential privacy
  - *Differentially private algorithm* will behave (approximately) the same in two data sets that differ only on a tiny subset
    - Presence or absence of single individual don’t matter

~~Summary~~ Picture of data mining

Image: Wikipedia

Essay topics

- Choose one of the following
  - *What is data mining?*
    - DM vs. ML, statistics; DM as a process; DM as a CS discipline, DM and other sciences, …
    - DM textbook introductions are a good source
  - *Is data mining a science?*
    - The scientific method vs. DM; DM as a methodological science; data-driven sciences; philosophy of science
- Remember: have more than one source; use your own thinking
- DL 30 October, 14:00 hours
  - I suggest mailing be before the lecture to get a reply
  - Submission guidelines will be in the web page soon