Introduction to Data Mining

Lecture 1

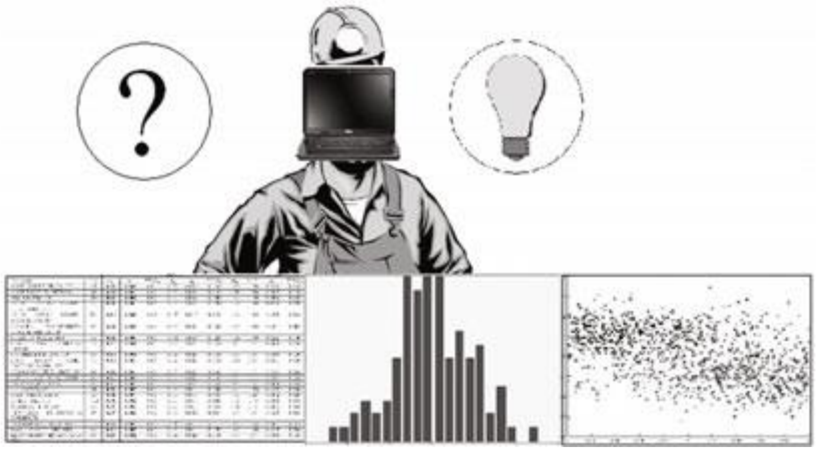
- What is data mining
- Why do we need data mining
- Data mining tasks
- Course requirements

Data and information

- *Data* – recorded facts
- *Information* – set of patterns that underlie the data – data model
- Information is locked up in databases

- What is data mining
- Why do we need data mining
- Data mining tasks
- Course requirements

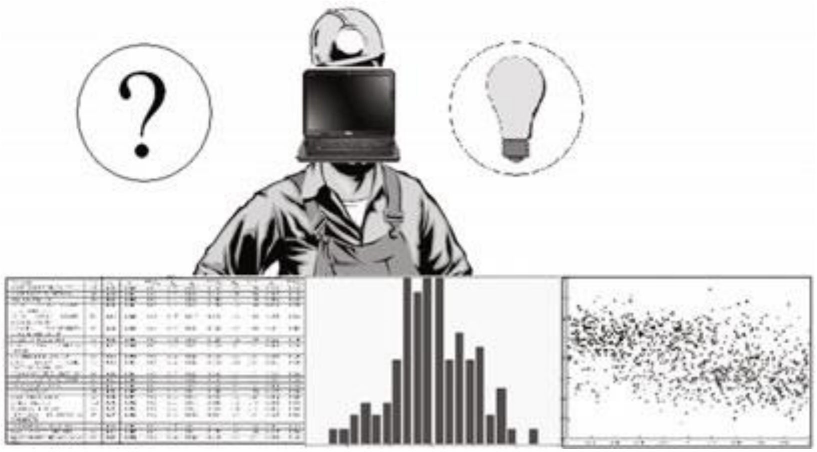
Definition 1



Data mining (Knowledge Discovery in Databases – KDD) – automatic or semi-automatic discovery of *models* and *patterns* from large datasets

- What is data mining
- Why do we need data mining
- Data mining tasks
- Course requirements

Definition 2



Data mining – extraction of implicit, previously unknown and potentially useful information

- What is data mining
- Why do we need data mining
- Data mining tasks
- Course requirements

Inferring models from data

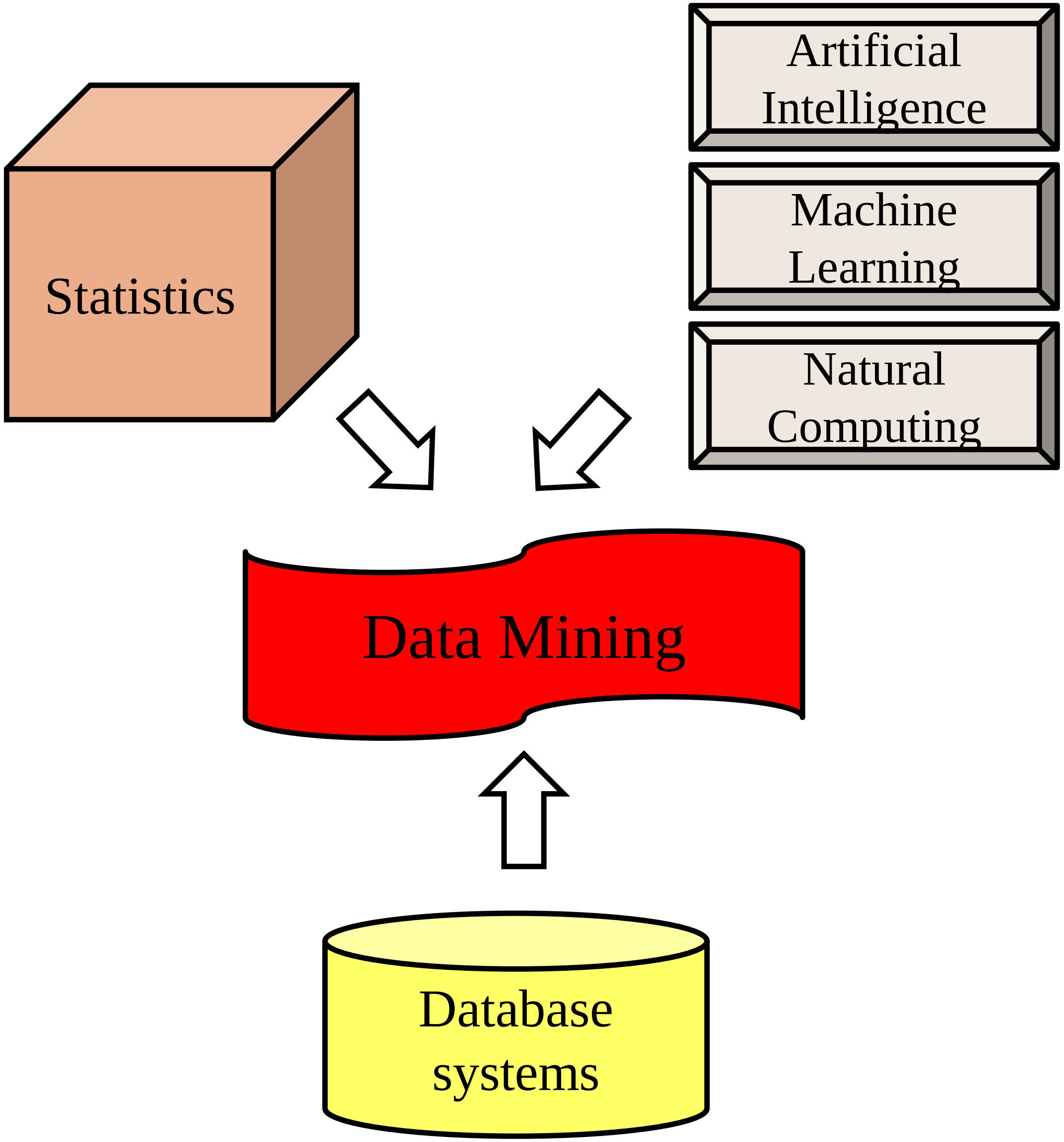
- People learn to associate objects with classes
- People categorize things all the time
- People recognize repeating patterns

The difference is:

- The data is digital
- The data is massive
- The inference is automatic (or semi-automatic)

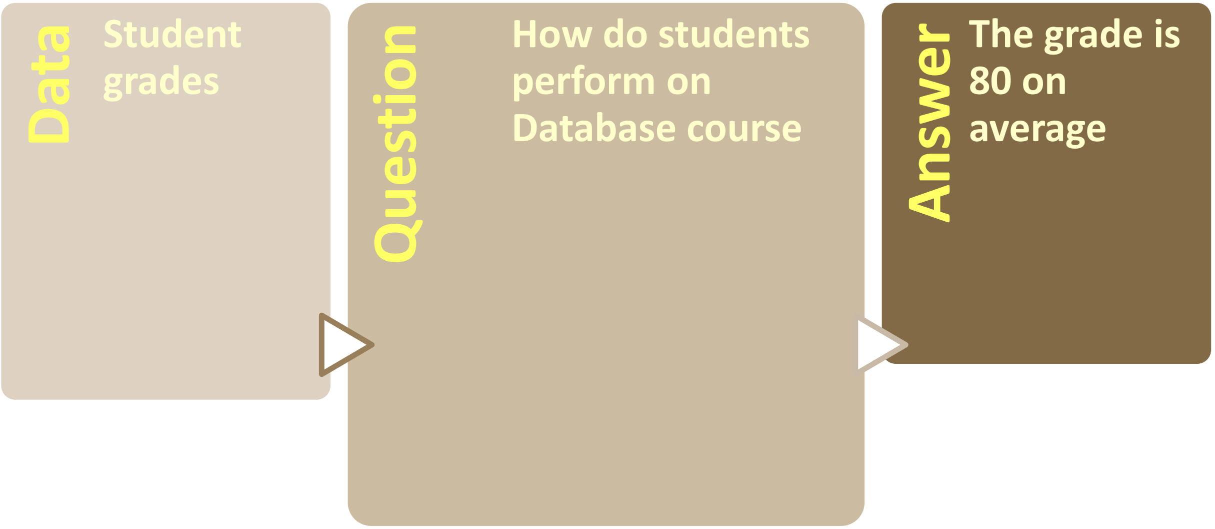
- What is data mining
- Why do we need data mining
- Data mining tasks
- Course requirements

Roots of data mining



- What is data mining
- Why do we need data mining
- Data mining tasks
- Course requirements

What is (not) data mining



Data

Student grades

Question

How do students perform on Database course

Answer

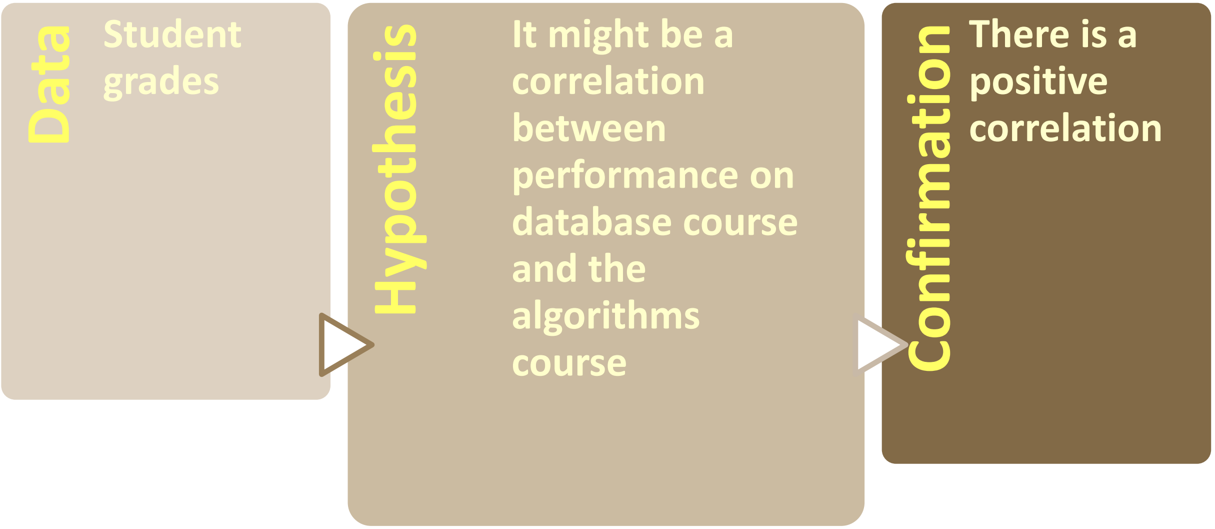
The grade is 80 on average

Not the data mining

- data manipulation (query)

- **What is data mining**
- Why do we need data mining
- Data mining tasks
- Course requirements

What is (not) data mining



Data

Student grades

Hypothesis

It might be a correlation between performance on database course and the algorithms course

Confirmation

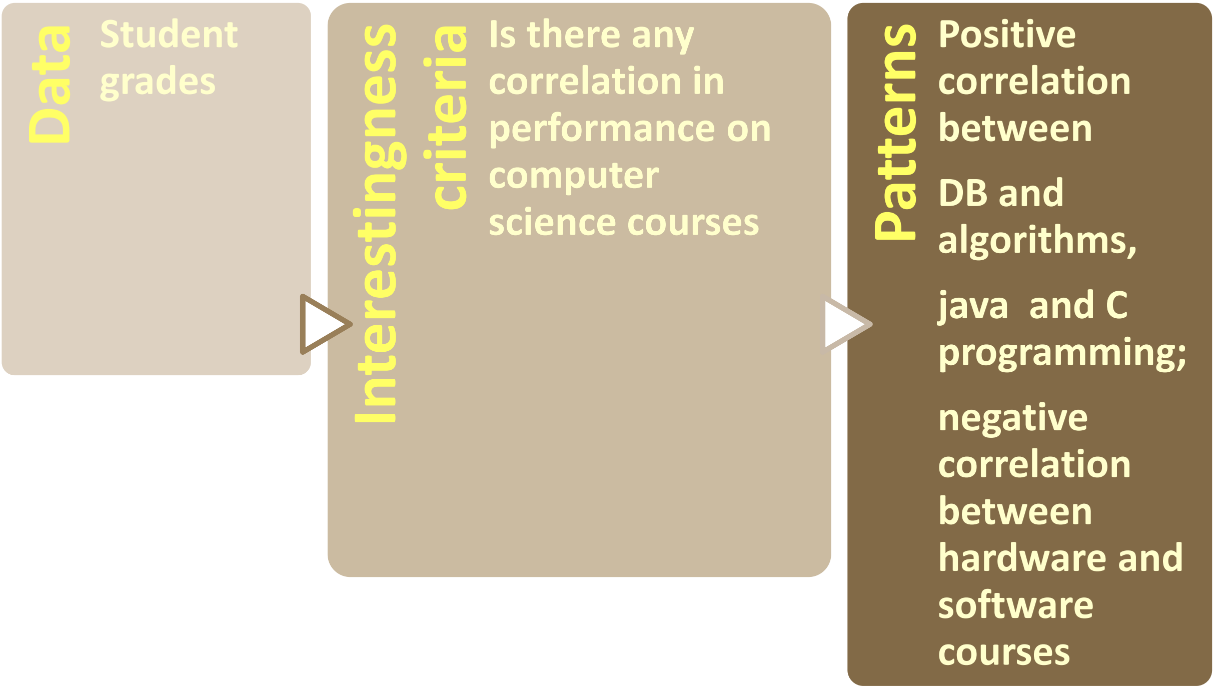
There is a positive correlation

Not the data mining

- statistics (hypothesis testing)

- **What is data mining**
- Why do we need data mining
- Data mining tasks
- Course requirements

What is (not) data mining



Data

Student grades

Interestingness criteria

Is there any correlation in performance on computer science courses

Patterns

Positive correlation between

DB and algorithms,

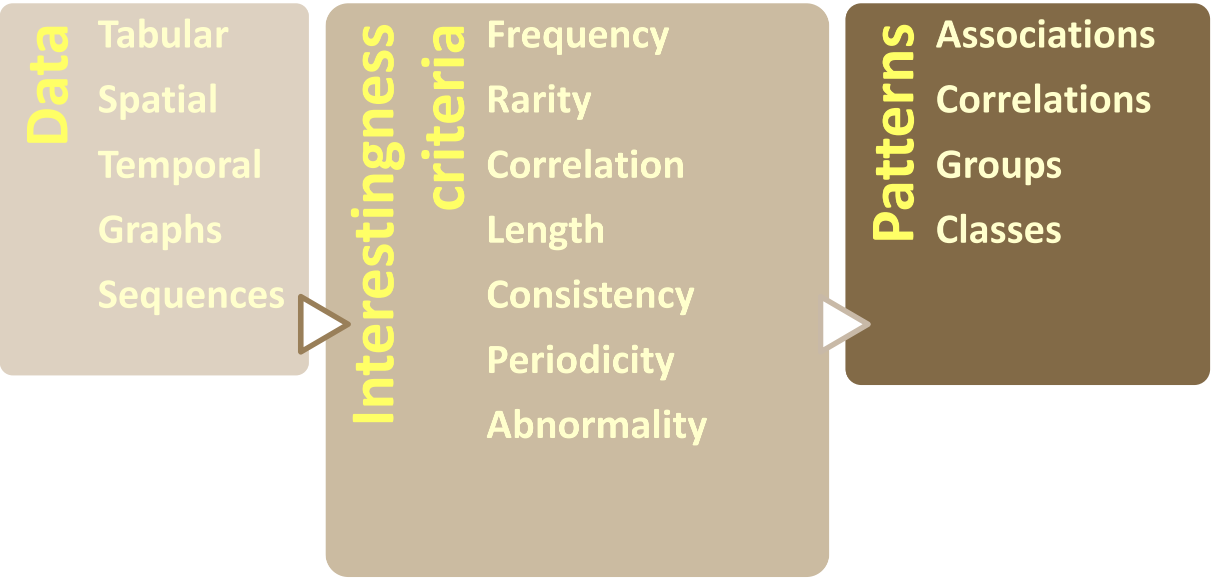
java and C programming;

negative correlation between hardware and software courses

Data mining!

- What is data mining
- Why do we need data mining
- Data mining tasks
- Course requirements

Data mining process



Data
- Tabular
- Spatial
- Temporal
- Graphs
- Sequences

Interestingness criteria
- Frequency
- Rarity
- Correlation
- Length
- Consistency
- Periodicity
- Abnormality

Patterns
- Associations
- Correlations
- Groups
- Classes

- What is data mining
- Why do we need data mining
- Data mining tasks
- Course requirements

Everything is recorded

- We do not discard data – just buy a new disk
- Ubiquitous electronics record our decisions and choices:
  - What do we buy
  - Our financial habits
  - Our comings and goings
- WWW contains tons of data – every choice we make is recorded

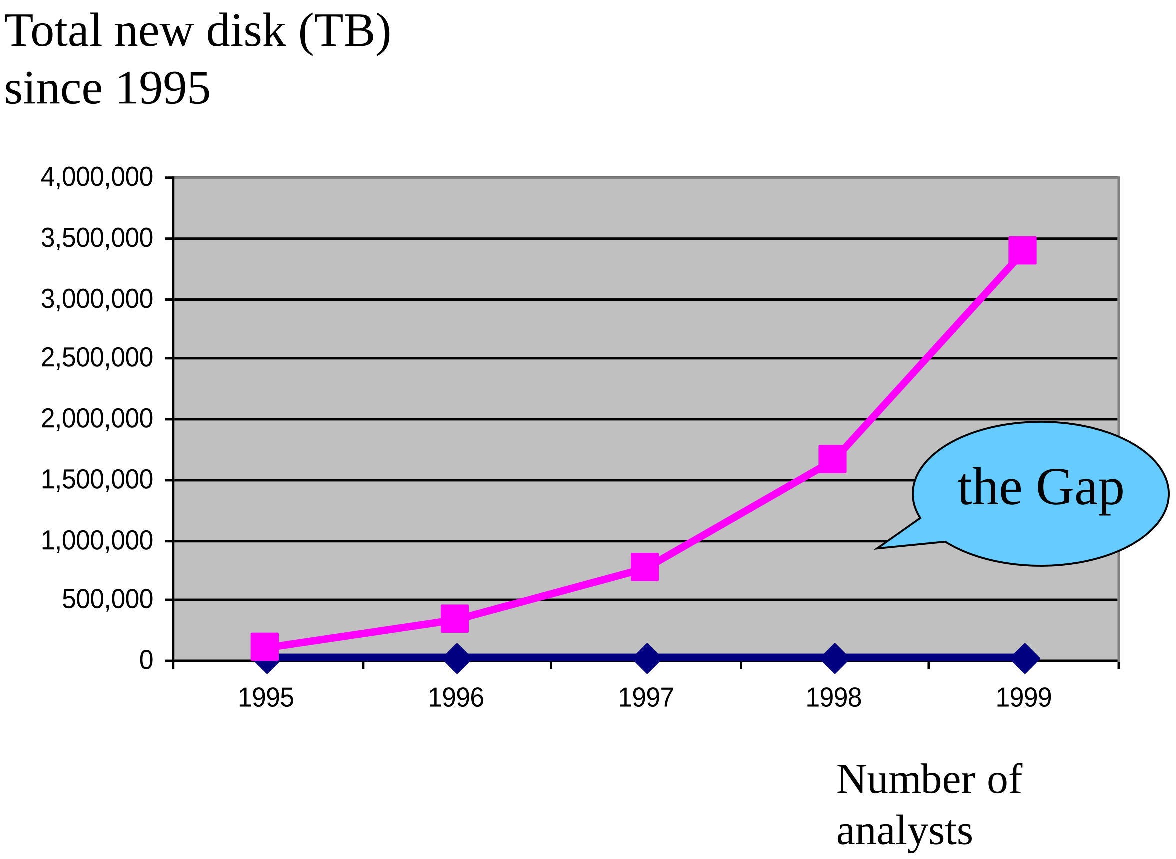
- What is data mining
- Why do we need data mining
- Data mining tasks
- Course requirements

Data flood

- Largest database in the world: World Data Centre for Climate (WDCC)
  - *220 terabytes of data on climate research and climatic trends,*
  - *110 terabytes worth of climate simulation data.*
  - *6 petabytes worth of additional information stored on tapes.*
- AT&T
  - *323 terabytes of information*
  - *1.9 trillion phone call records*
- Google
  - *91 million searches per day,*
    - *After a year more than 33 trillion database entries.*

- What is data mining
- Why do we need data mining
- Data mining tasks
- Course requirements

Gap between data and information



From: R. Grossman, C. Kamath, V. Kumar, “Data Mining for Scientific and Engineering Applications”

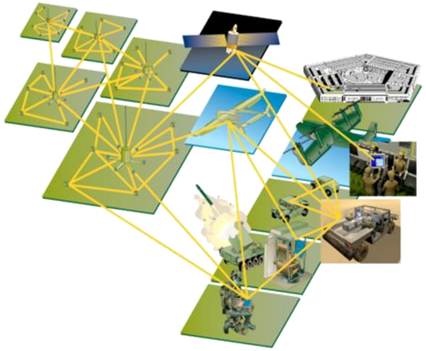
- What is data mining
- Why do we need data mining
- Data mining tasks
- Course requirements

Commercial viewpoint

- Twice as much information was created in 2002 as in 1999 (~30% growth rate)
  - E-commerce
  - Chain transactions
  - Bank transactions
  - Customer profiles

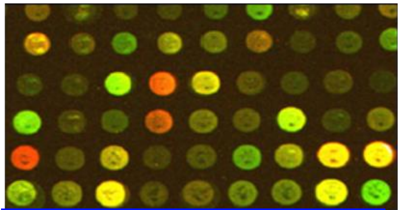
- We can find
  - Purchase patterns
  - Credit Card frauds
  - Border crossing alerts
  - Customer retention

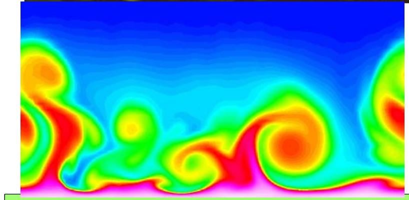
- What is data mining
- Why do we need data mining
- Data mining tasks
- Course requirements

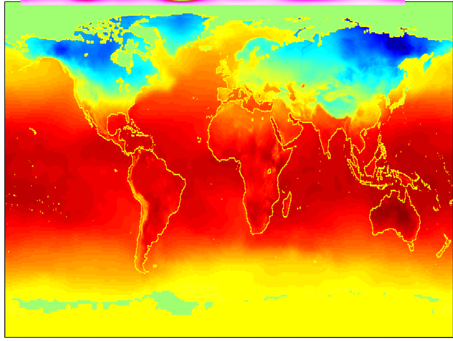


Scientific viewpoint

- Data is collected and stored at enormous speeds (GB/hour).
  - remote sensors on a satellite
  - telescopes scanning the skies
  - scientific simulations generating terabytes of data
  - gene expression profiles
- We can:
  - Classify faint galaxies
  - Find similar gene expressions for different drug treatments
  - Predict structure of a chemical from magnetic resonance data







- What is data mining
- Why do we need data mining
- Data mining tasks
- Course requirements

Data mining helps to discover *knowledge*

"*Scientia potentia est*"

("Knowledge is power")

F. Bacon, 1597

Remark:

Like in the original mining, it is possible for data mining to dig the 'mine' of data without eventually discovering the lode containing the "gold nugget" of knowledge.

- What is data mining
- Why do we need data mining
- Data mining tasks
- Course requirements

Data mining and privacy

- Can we include sexual and racial attributes?
  - in medicine?
  - in loan application?
- Implicit privacy violations: zip code

- What is data mining
- Why do we need data mining
- **Data mining tasks**
- Course requirements

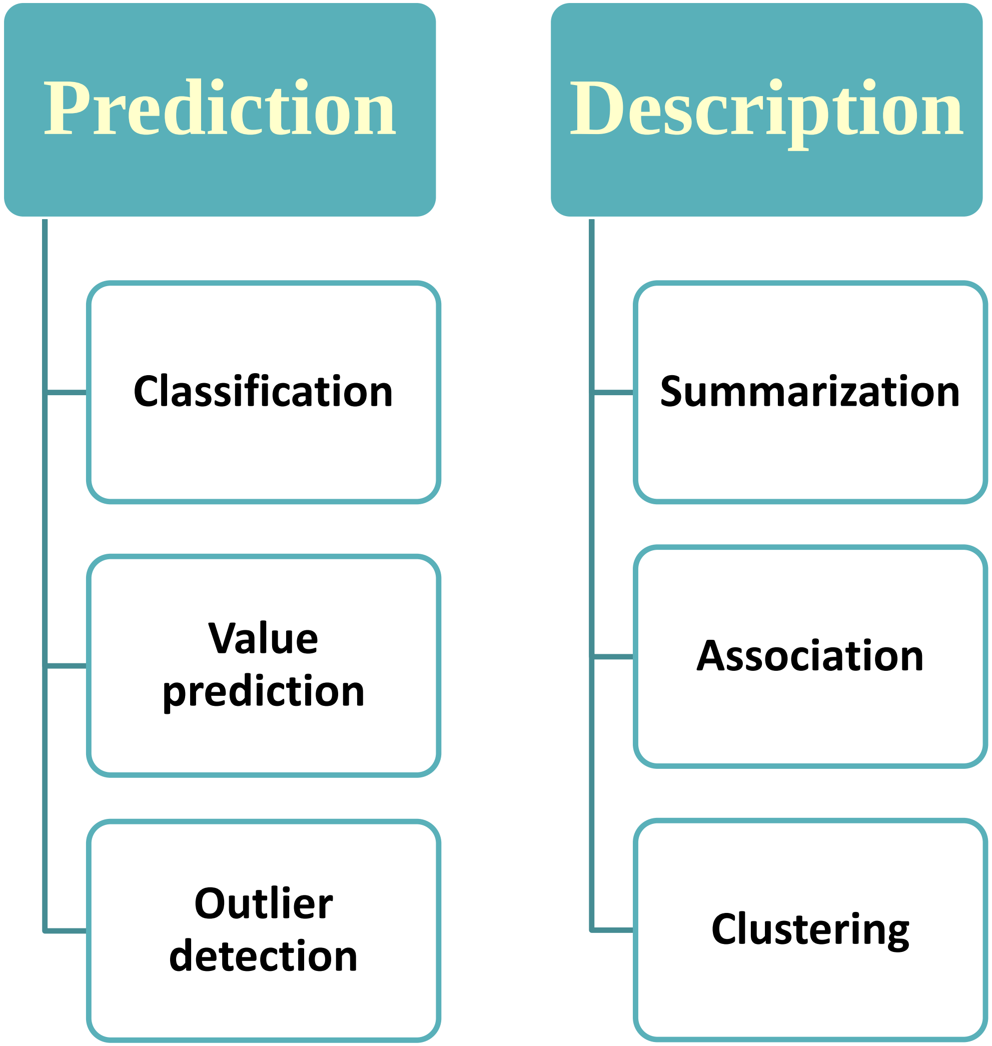
Interestingness criteria

Interestingness criteria

- Frequency
- Rarity
- Correlation
- Periodicity
- Consistency
- Length

- What is data mining
- Why do we need data mining
- Data mining tasks
- Course requirements

Task types



Prediction

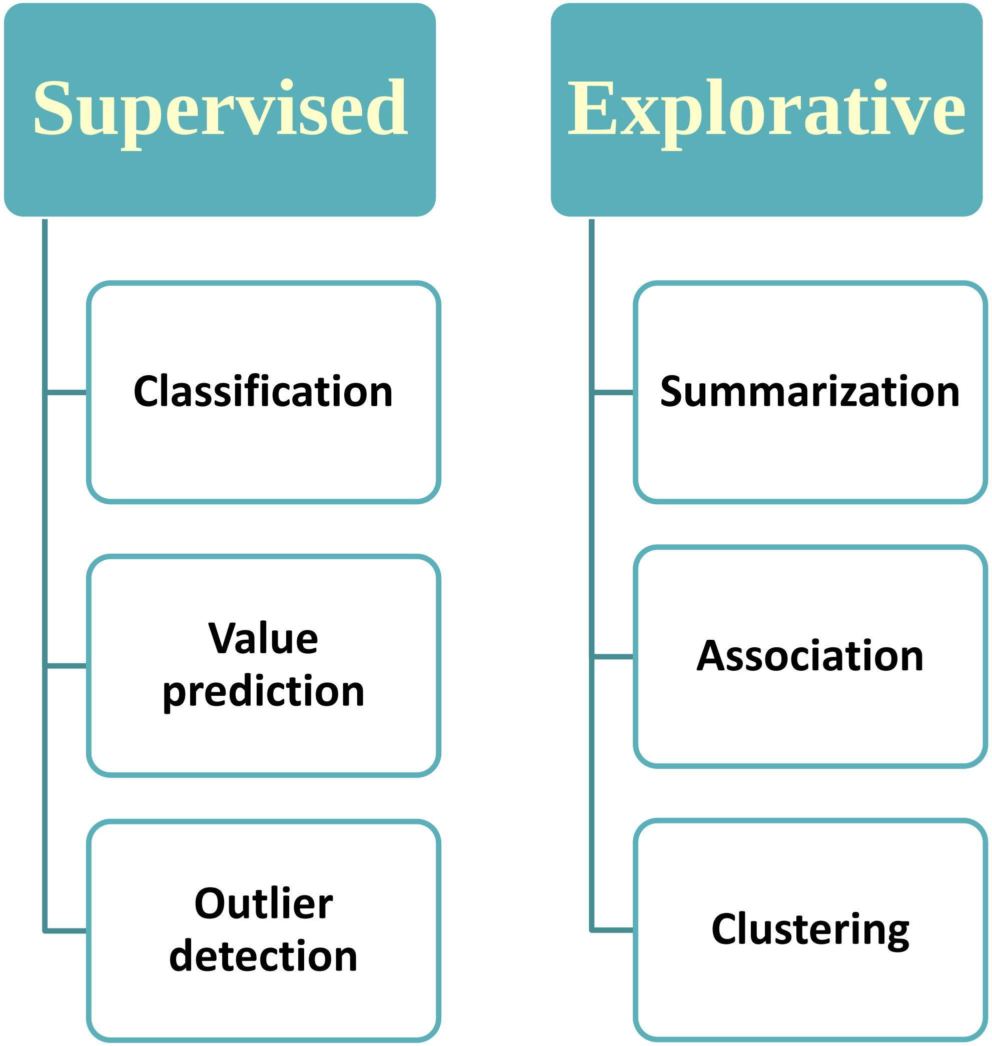
- Classification
- Value prediction
- Outlier detection

Description

- Summarization
- Association
- Clustering

- What is data mining
- Why do we need data mining
- Data mining tasks
- Course requirements

Task types



Supervised

- Classification
- Value prediction
- Outlier detection

Explorative

- Summarization
- Association
- Clustering

- What is data mining
- Why do we need data mining
- Data mining tasks
- Course requirements

Tabular input

attributes

class

data record

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

Task of type 1: Classification

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.

- Find ("learn") a *model* for the class attribute as a function of the values of the other attributes.

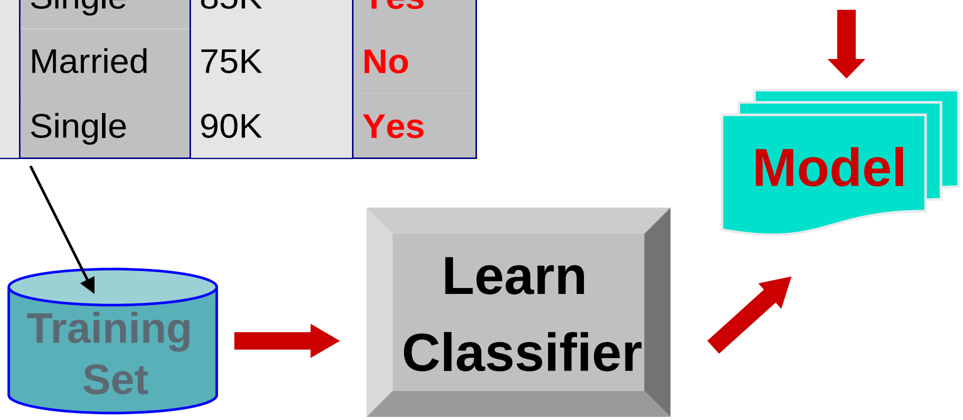
- Goal: **previously unseen** records should be assigned a class as accurately as possible.

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

Classification example

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

Solving classification problem

My neighbour dataset

class label

Temp	Precip	Day	Shop	Clothes	
25	None	Sat	No	Casual	**Walk**
-5	Snow	Mon	Yes	Casual	**Drive**
15	Snow	Mon	Yes	Casual	**Walk**

(Adapted from Leslie Kaelbling's example in the MIT courseware)

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

Classification problem

class label

Temp	Precip	Day	Shop	Clothes	
25	None	Sat	No	Casual	**Walk**
-5	Snow	Mon	Yes	Casual	**Drive**
15	Snow	Mon	Yes	Casual	**Walk**
-5	Snow	Mon	Yes	Casual	**?**

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

Classification problem: memory

class label

Temp	Precip	Day	Shop	Clothes	
25	None	Sat	No	Casual	**Walk**
-5	Snow	Mon	Yes	Casual	**Drive**
15	Snow	Mon	Yes	Casual	**Walk**
-5	Snow	Mon	Yes	Casual	**Drive**

(Adapted from Leslie Kaelbling's example in the MIT courseware)

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

Classification problem: noise

Temp	**Precip**	**Day**	**Clothes**	
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Drive**
25	None	Sat	Casual	**Drive**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**?**

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

Classification problem: averaging

Temp	Precip	Day	Clothes	
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Drive**
25	None	Sat	Casual	**Drive**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

Classification problem: generalization

Temp	**Precip**	**Day**	**Clothes**	
22	None	Fri	Casual	**Walk**
3	None	Sun	Casual	**Walk**
10	Rain	Wed	Casual	**Walk**
30	None	Mon	Casual	**Drive**
20	None	Sat	Formal	**Drive**
25	None	Sat	Casual	**Drive**
-5	Snow	Mon	Casual	**Drive**
27	None	Tue	Casual	**Drive**
24	Rain	Mon	Casual	**?**

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

Learning to predict class label

Three different problems involved in learning:

- memory
- averaging
- generalization.

Type 2. Explorations

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

Tid	Refund	Marital Status	Taxable Income
1	Yes	Single	125K
2	No	Married	100K
3	No	Single	70K
4	Yes	Married	120K
5	No	Divorced	95K
6	No	Married	60K
7	Yes	Divorced	220K
8	No	Single	85K
9	No	Married	75K
10	No	Single	90K

Discover groups, no class labels

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

Task of type 2. Associations

The Market-Basket Model

- A large set of *items*, e.g., things sold in a supermarket.
- A large set of *baskets*, each of which is a small set of the items, e.g., the things one customer buys in one transaction.

Fundamental problem

- What sets of items are often bought together?

Application

- If a large number of baskets contain both hot dogs and mustard, we can use this information. How?

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

Solving association problem: market basket

	Itemsets
1	{bread, milk, peanut butter}
2	{bread, milk}
3	{beer, potato chips}
4	{beer, diapers}
5	{beer, milk, diapers}
6	{bread, milk, yogurt}
7	{beer, bread, diapers}
8	{bread, milk, jelly}
9	{beer, cigarettes, diapers}
10	{bread, milk}

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

Association problem

	Itemsets
1	{**bread**, **milk**, peanut butter}
2	{**bread**, **milk**}
3	{**beer**, potato chips}
4	{**beer**, **diapers**}
5	{**beer**, **milk**, **diapers**}
6	{**bread**, **milk**, yogurt}
7	{**beer**, **bread**, **diapers**}
8	{**bread**, **milk**, jelly}
9	{**beer**, cigarettes, **diapers**}
10	{**bread**, **milk**}

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

Beer and diapers?

	Itemsets
1	{bread, milk, peanut butter}
2	{bread, milk}
3	{**beer**, potato chips}
4	{**beer**, **diapers**}
5	{**beer**, milk, **diapers**}
6	{bread, milk, yogurt}
7	{**beer**, bread, **diapers**}
8	{bread, milk, jelly}
9	{**beer**, cigarettes, **diapers**}
10	{bread, milk}

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

On-Line Purchases: potentially useful patterns

Log file

Date	Customer	Product
Dec 20	John	iPod
Dec 23	John	Video camera
Jan 4	Mary	Dumbbells
Jan 4	John	Kindle
Jan 20	Tim	Laptop
Jan 23	Mary	Kindle
Feb 1	Tim	iPod
Feb 3	Tim	Video camera

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

On-Line Purchases: group by customer

Transaction: customer, item: product

Date	Customer	Product
Dec 20	John	iPod
Dec 23	John	Video camera
Jan 4	John	Kindle
Jan 4	Mary	Dumbbells
Jan 23	Mary	Kindle
Jan 20	Tim	Laptop
Feb 1	Tim	iPod
Feb 3	Tim	Video camera

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

On-Line Purchases: group by product

Transaction: product, item: customer

Date	Customer	Product
Dec 20	John	iPod
Feb 1	Tim	iPod
Jan 4	Mary	Dumbbells
Dec 23	John	Video camera
Feb 3	Tim	Video camera
Jan 20	Tim	Laptop
Jan 4	John	Kindle
Jan 23	Mary	Kindle

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

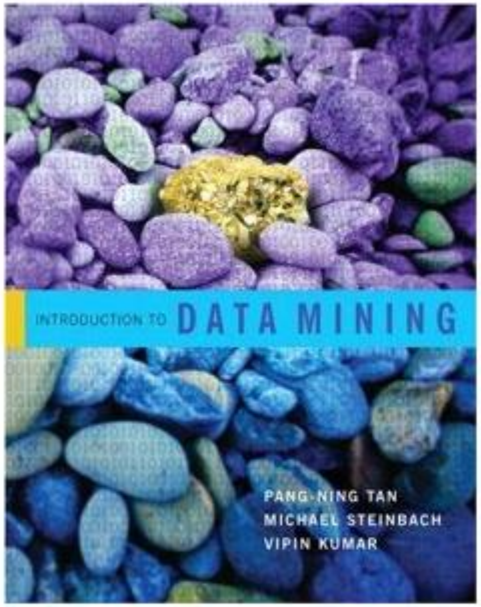
On-Line Purchases: group by month

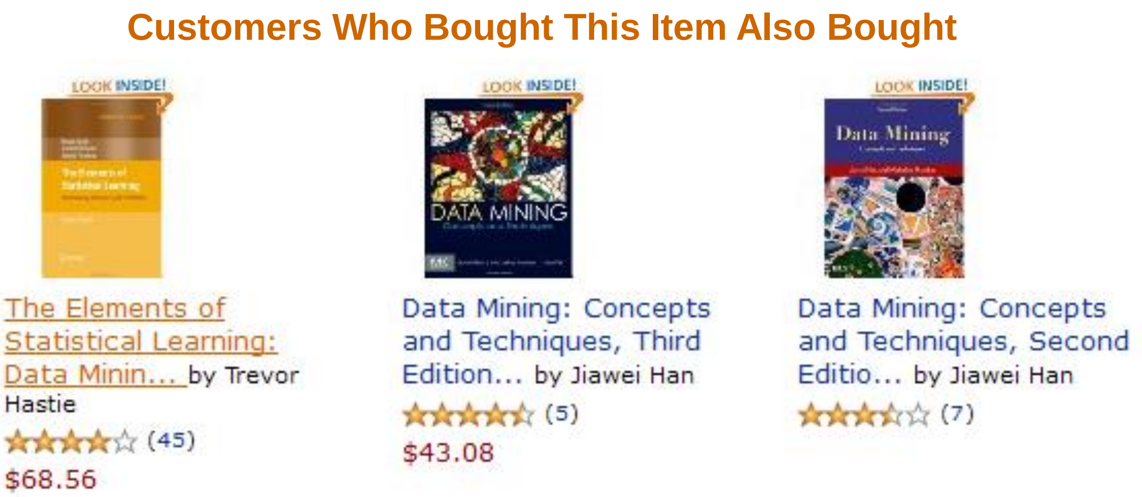
Transaction: month, item: product

Date	Customer	Product
Dec 20	John	iPod
Dec 23	John	Video camera
Jan 4	Mary	Dumbbells
Jan 4	John	Kindle
Jan 20	Tim	Laptop
Jan 23	Mary	Kindle
Feb 1	Tim	iPod
Feb 3	Tim	Video camera

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

Amazon example





Customers Who Bought This Item Also Bought

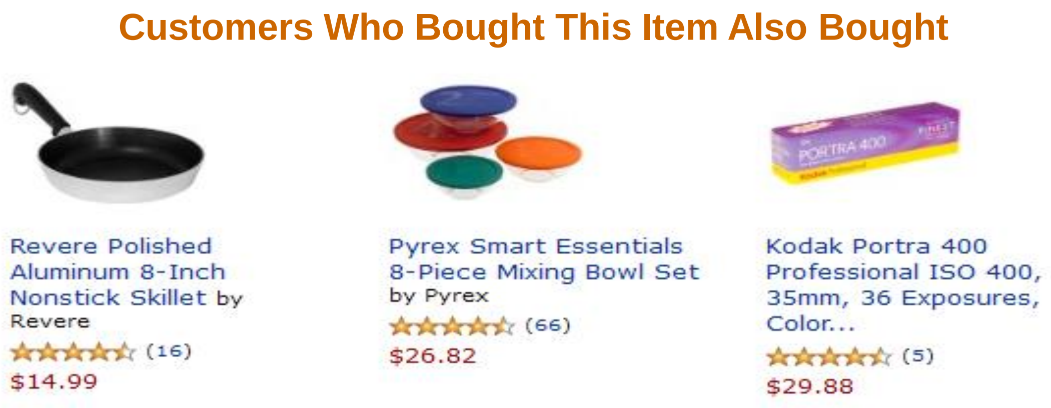
The Elements of Statistical Learning: Data Minin... by Trevor Hastie
★★★★☆ (45)
$68.56

Data Mining: Concepts and Techniques, Third Edition... by Jiawei Han
★★★★★ (5)
$43.08

Data Mining: Concepts and Techniques, Second Editio... by Jiawei Han
★★★★☆ (7)

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

Amazon example ?



Customers Who Bought This Item Also Bought

Revere Polished Aluminum 8-Inch Nonstick Skillet by Revere
★★★★☆ (16)
$14.99

Pyrex Smart Essentials 8-Piece Mixing Bowl Set by Pyrex
★★★★☆ (66)
$26.82

Kodak Portra 400 Professional ISO 400, 35mm, 36 Exposures, Color...
★★★★☆ (5)
$29.88

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

Topics: algorithms

- Classification:
  - Decision trees and rule-based classifiers
  - Bayesian inference
  - Support vector machines
  - Natural computing: genetic algorithm and neural networks
- Correlation
  - Frequent itemsets
  - Association rules
  - Frequent sequential and graph patterns
- Clustering
- Feature selection (Principal component analysis)
- Link analysis (PageRank algorithm)

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

Labs: learning by doing

- Learning by example: on toy datasets which exhibit features of real-life datasets
- WEKA*) – Waikato Environment for Knowledge Analysis
- JAVA implementations and extensions
- Real-life datasets analysis

*)Weka- unique New Zealand flightless bird with inquisitive nature

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

Prerequisites

- Basic knowledge of probabilities
- Linear algebra basics
- Reasoning about the data

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

Expected outcomes

- Understanding of basic algorithms
- Ability to select the right algorithm for a problem at hand
- Ability to perform data mining task (coding is optional)
- Validation of results (coding is optional)
- Presentation of results (coding is optional)

- What is data mining
- Why do we need data mining
- Data mining tasks
  - Predictive
  - Descriptive
- Course requirements

Grading

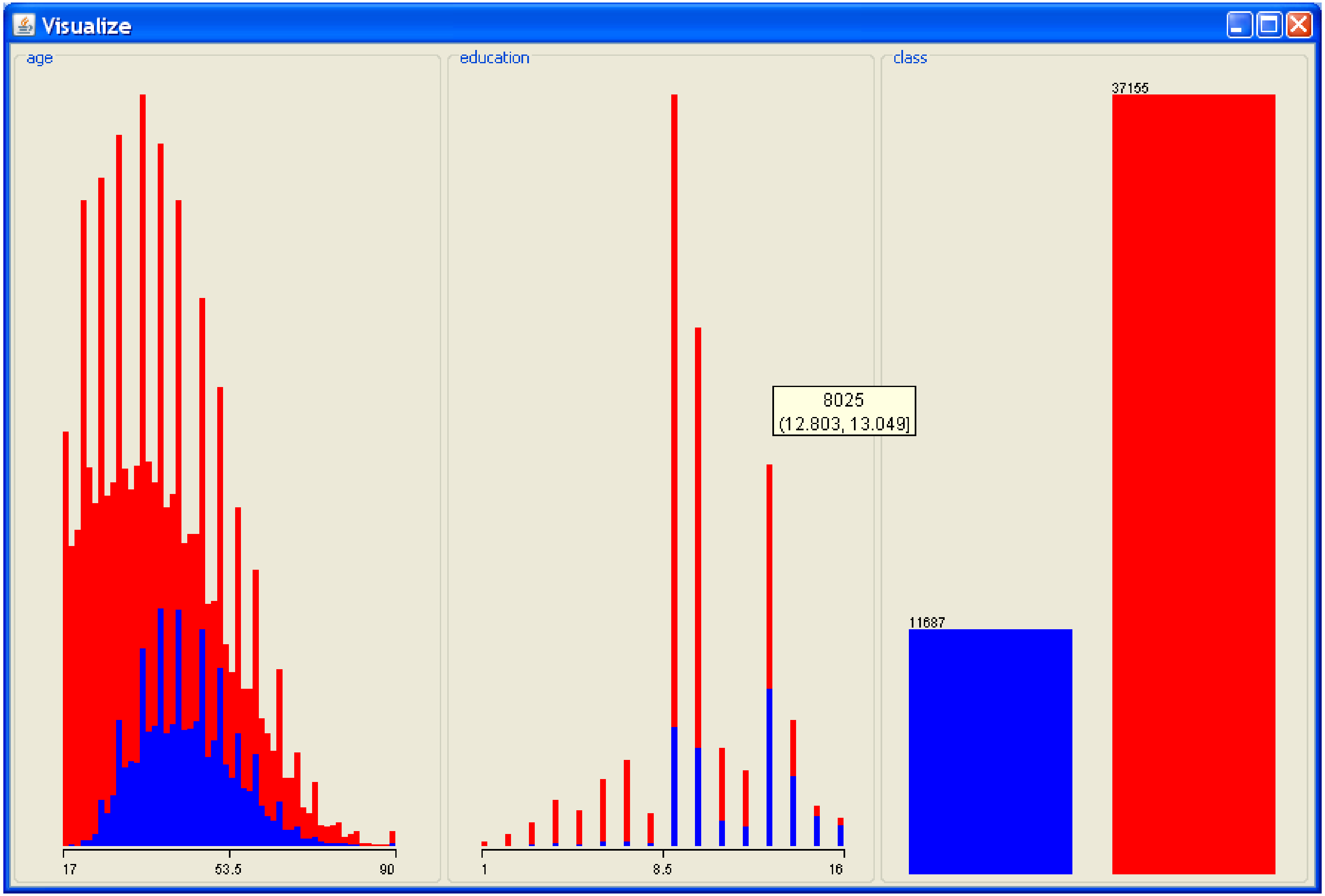
- Quizzes: to monitor understanding. Each correct quiz + 0.5 bonus
- 3 assignments (10% each):
  - Part 1. Solve a toy problem by hand (understanding)
  - Part 2. Perform data mining task on a real dataset (doing)
- Projects (20%) – two types
  - Type 1. Take a real dataset, suggest data mining task, perform task, evaluate and present results
  - Type 2. Introduce a novel data mining approach based on recent publications, show connections to the learned concepts and ability to do independent data mining research
- Exams: (20% and 30%) – test understanding (open book exams)

Lab example: what determines high salary

Adult income dataset (US census 1994)

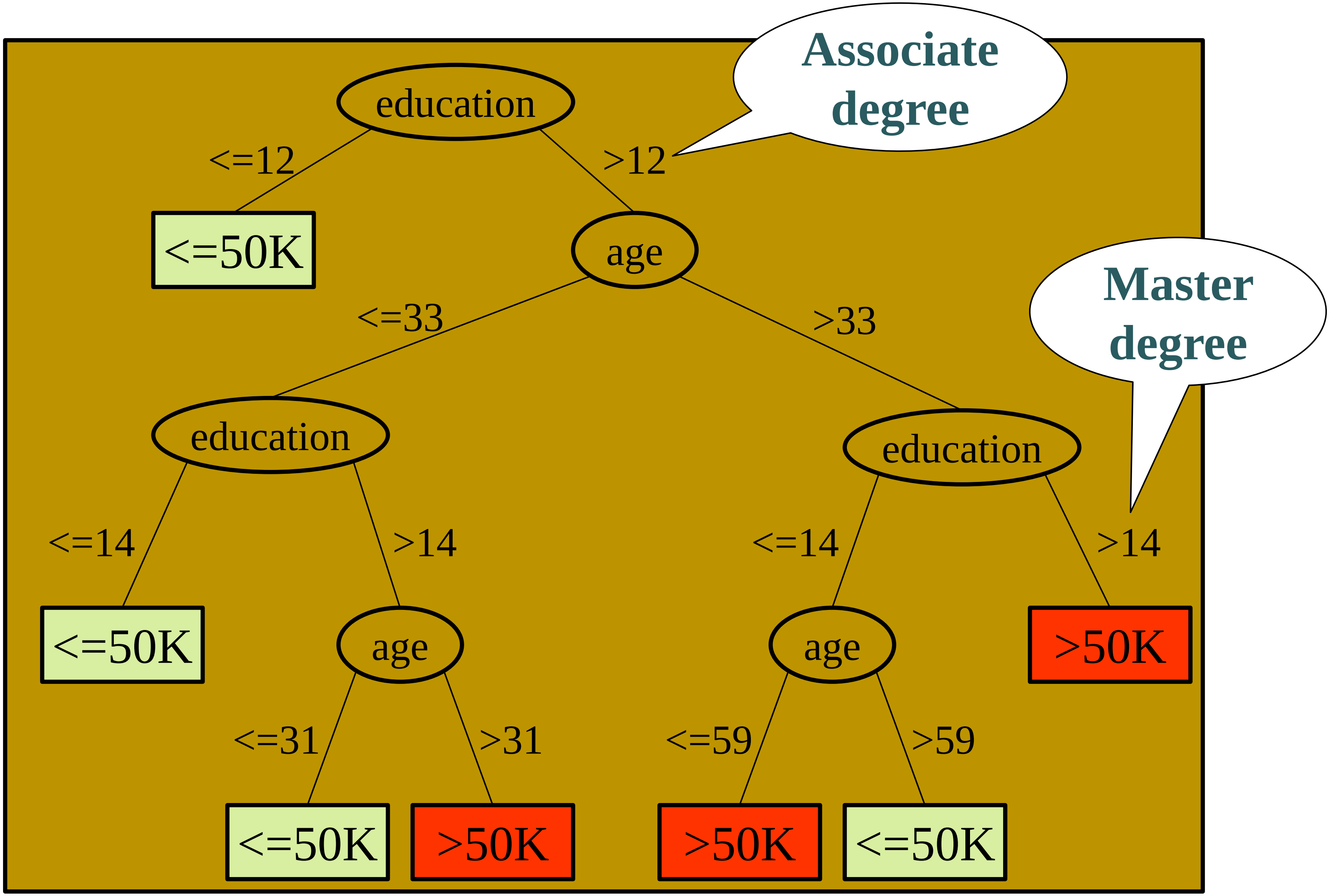
Age	Education	Mar. status	Occupation	Race	Sex	Born in	Yearly income
39	Bachelors	Never-married	Adm-clerical	White	M	US	**<=50 K**
50	Bachelors	Married-civ-spouse	Exec-managerial	White	M	US	**<=50 K**
54	7th-8th	Married-civ-spouse	Machine-op-inspct	White	M	US	**>50K**
37	Bachelors	Never-married	Exec-managerial	Black	M	US	**>50K**
28	Bachelors	Married-civ-spouse	Prof-specialty	Black	F	Cuba	**<=50 K**
37	Masters	Married-civ-spouse	Exec-managerial	White	F	US	**<=50 K**

Visualization of attributes age and education (not data mining)



The results of data mining:

decision tree on age and education attributes


education
<=12
>12
Associate degree
<=50K
age
<=33
>33
Master degree
education
education
<=14
>14
<=14
>14
<=50K
age
age
>50K
<=31
>31
<=59
>59
<=50K
>50K
>50K
<=50K