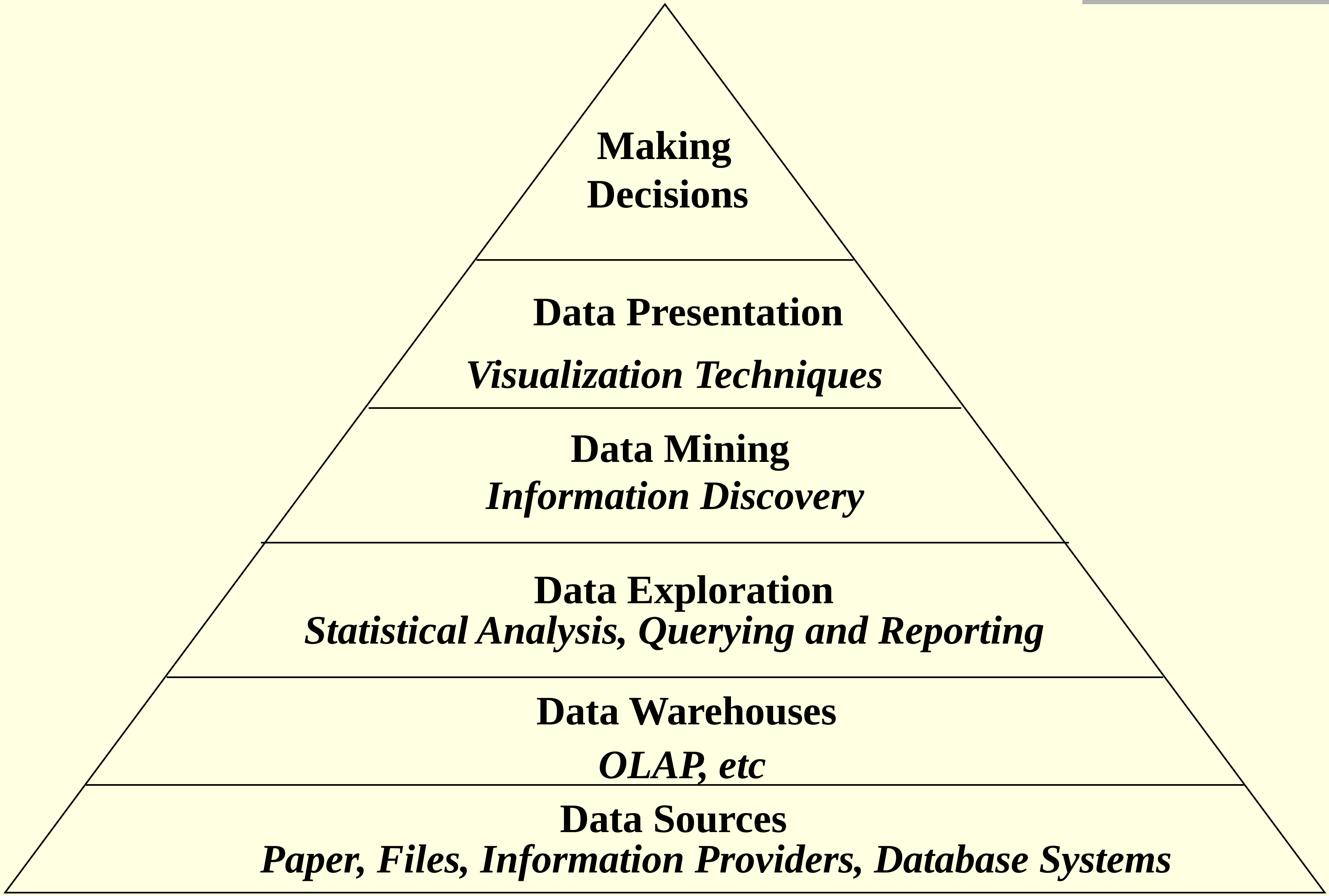
Fundamentals of Data Science

- Rapid advances in data collection and storage technology have enabled the accumulation of vast amounts of data.
- Often, traditional data analysis tools and techniques cannot be used because of the massive size of a data set.
- Solution:
  - Use data science principles to discover interesting knowledge (rules, regularities, patterns, constraints) from data in large databases.

Fundamentals of Data Science



Making Decisions

Data Presentation

Visualization Techniques

Data Mining

Information Discovery

Data Exploration

Statistical Analysis, Querying and Reporting

Data Warehouses

OLAP, etc

Data Sources

Paper, Files, Information Providers, Database Systems

Applications

- The main application areas include
  - Business
  - Science and Engineering
  - Medicine

Business

- Point-of-sale data collection (e.g. bar code scanners, RFID) have allowed retailers to collect up-to-the minute data about customer purchases.

Business

- Retailers can use this information to help them
  - Better understand the needs of their customers
  - Make more informed business decisions.

Business

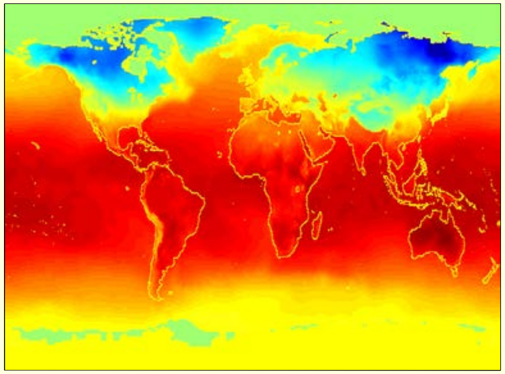
- Data science principles can be used to support a wide range of business intelligence applications:
  - Customer profiling
  - Targeted marketing
  - Work flow management
  - Store layout

Business

- Data science can also help retailers to answer questions such as
  - Who are the most profitable customers ?
  - Which products can be sold together ?
  - What is the revenue outlook of the company for next year?

Science and Engineering

- Researchers in science and engineering are rapidly accumulating data.
- Example: NASA has deployed a series of Earth-orbiting satellites to observe the land surface, oceans and atmosphere.
- Because of the size of the data, traditional methods are often not suitable for analyzing these data sets.

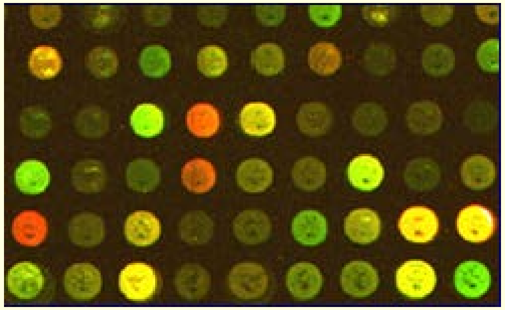


Science and Engineering

- Data science can help earth scientists to answer questions such as
  - How is rainfall affected by ocean surface temperature?
  - How well can we predict the beginning and end of the growing season for a region?

Medicine

- In the past, traditional methods in molecular biology allowed scientists to study only a few genes at a time in a given experiment.
- Recent breakthroughs in gene expression profiling have enabled scientists to compare the behavior of a large number of genes under various situations.



Medicine

- However, the noisy and high dimensional nature of data requires new types of data analysis.
- Data science principles can be applied to
  - Determine the function of each gene
  - Isolate the genes responsible for certain diseases.

What is data mining?

- Data mining is a fundamental approach in data science which allows automatic discovery of useful information in large data repositories.
- These techniques are used to search for novel and useful patterns in the data.
- They also provide capabilities to predict the outcome of a future observation.

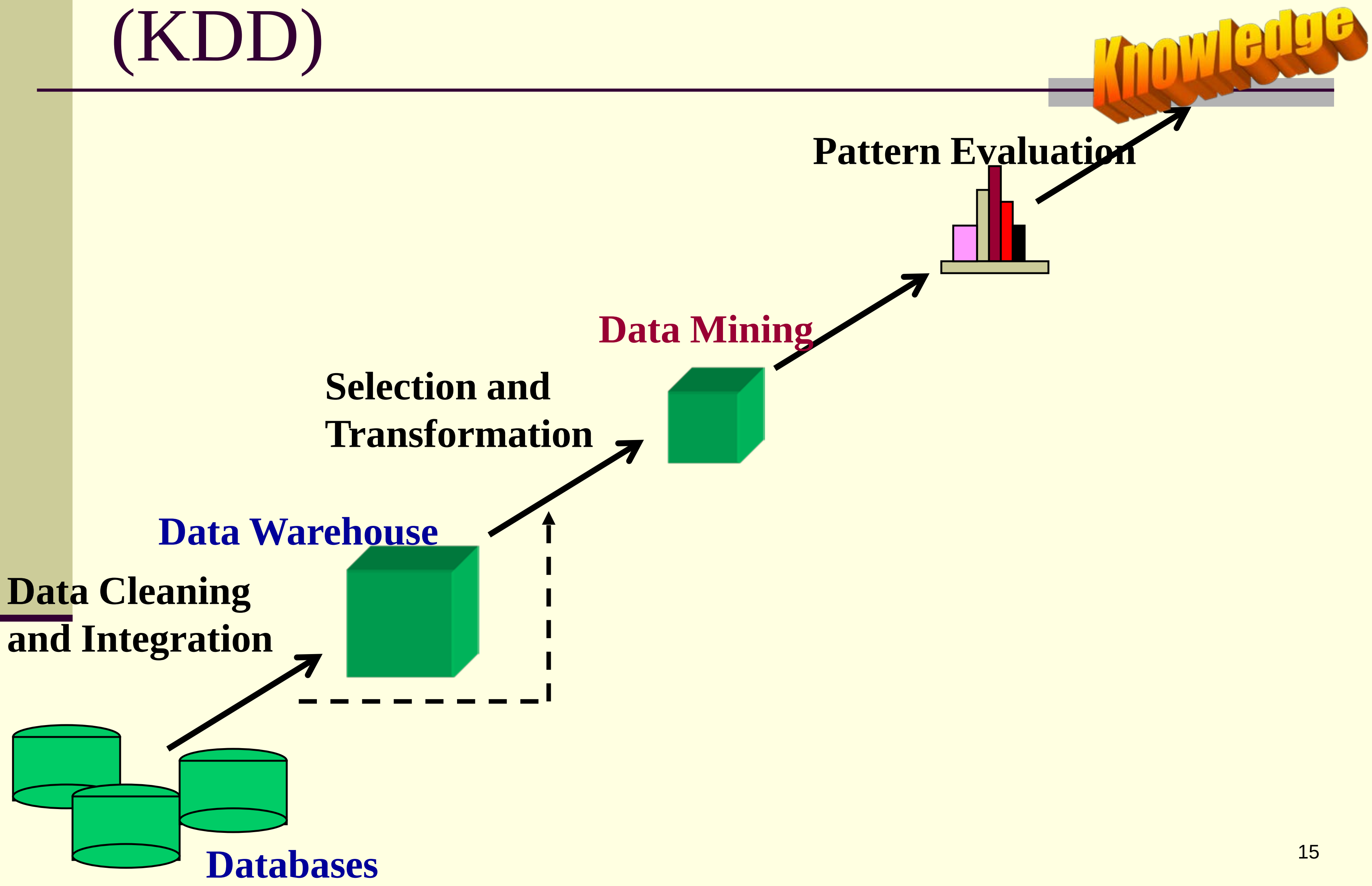
What is data mining?

- Not all information discovery tasks are considered to be data mining, e.g. information retrieval tasks such as:
  - Looking up individual records using a database management system.
  - Finding particular Web pages via a query to an Internet search engine.

Knowledge Discovery in Databases (KDD)

- Data mining is an integral part of knowledge discovery in databases (KDD)
- KDD is the overall process of converting raw data into useful information
- This process consists of a series of transformation steps:
  - Data cleaning and integration
  - Data selection and transformation
  - Data mining
  - Pattern evaluation
  - Knowledge presentation

Knowledge Discovery in Databases (KDD)



Knowledge Discovery in Databases (KDD)

- Data cleaning
  - A process that removes noise and inconsistent data.
- Data integration
  - The stage where multiple data sources are combined.
- Data selection
  - The stage where data relevant to the analysis task are retrieved from the database.
- Data transformation
  - The stage where data are transformed into forms suitable for mining.

Knowledge Discovery in Databases (KDD)

- Data mining
  - An important process where intelligent and efficient methods are applied to extract patterns.
- Pattern evaluation
  - A process that identifies the truly interesting patterns representing knowledge based on interestingness measures.
- Knowledge presentation
  - The stage where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

Challenges

- Some of the specific challenges that motivate the development of data science include
  - Scalability
  - High dimensionality
  - Heterogeneous and complex data

Scalability

- Data sets with sizes of gigabytes, terabytes or even petabytes are becoming common.
- Data mining algorithms need to be scalable to handle these massive data sets.
- Scalability may require
  - Employing special search strategies in search problems.
  - Implementing novel data structures to access individual records in an efficient manner.
  - Using sampling.
  - Developing parallel and distributed algorithms.

High dimensionality

- It is now common to encounter data sets with hundreds or thousands of attributes.
- Examples include
  - Gene expression data involving a large number of features.
  - Measurements of temperature at various locations taken repeatedly for an extended period.

High dimensionality

- Traditional data analysis techniques that were developed for low-dimensional data often do not work well for such high-dimensional data.
- In addition, the computational complexity increases rapidly as the dimensionality increases.

Heterogeneous and complex data

- Traditional data analysis methods often deal with data sets containing attributes of the same type.
- Recent years have seen the emergence of more complex data objects.

Heterogeneous and complex data

- Examples include
  - Collections of Web pages containing semi-structured text and hyperlinks
  - DNA data with sequence and 3D structure information.
  - Climate data that consist of time series of different types of measurements at various locations on the earth’s surface.
- Techniques for mining such data should take into consideration the complex relationships in the data.

Areas related to data science

- Data science draws upon ideas from different areas such as
  - Sampling, estimation and hypothesis testing from statistics.
  - Search algorithms, modeling techniques and learning theories from artificial intelligence, pattern recognition, and machine learning.

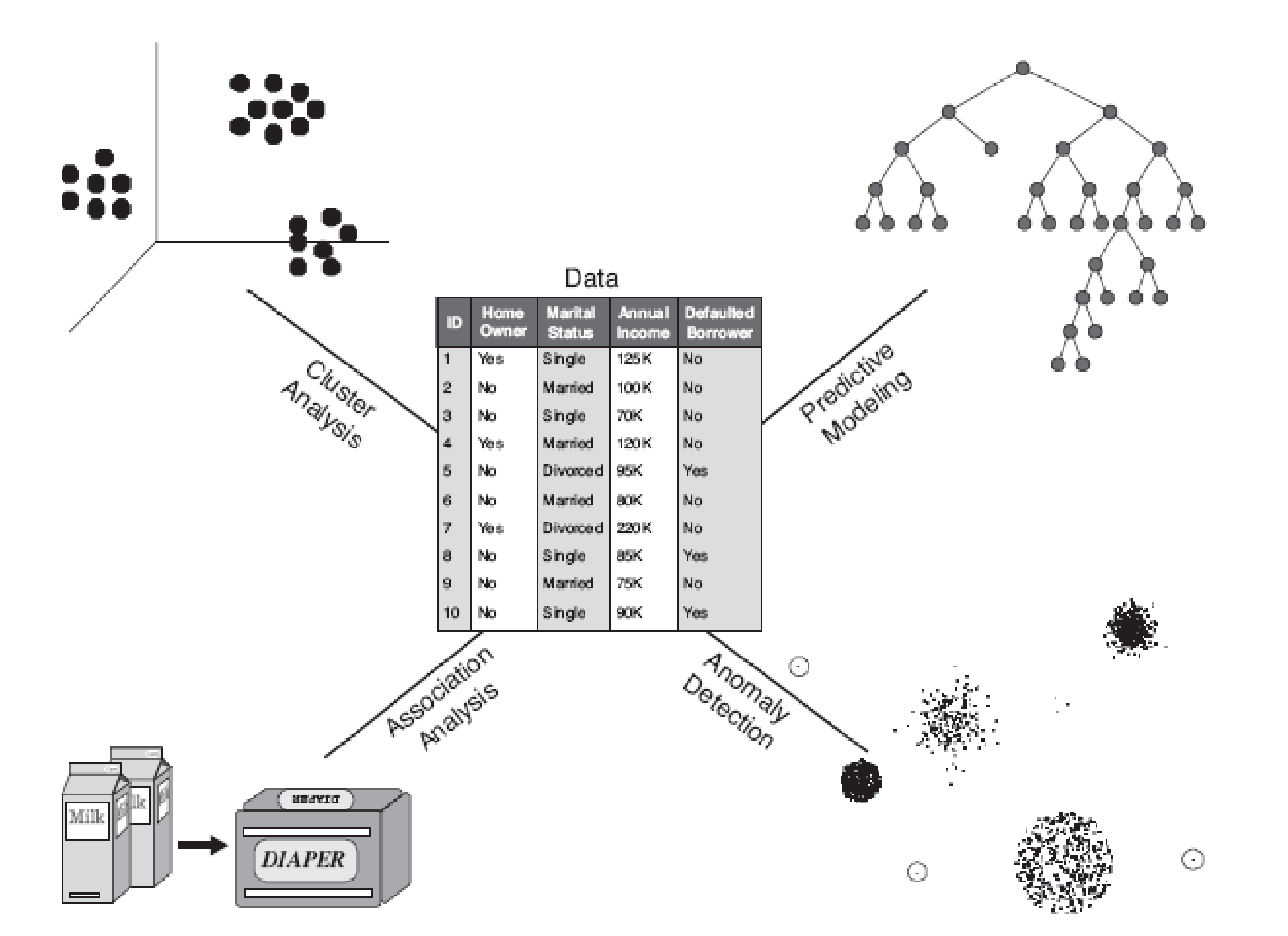
Areas related to data science

- A number of other areas also play key supporting roles
  - Database systems are needed to provide support for efficient storage, indexing and query processing.
  - High performance computation techniques are important in addressing the massive size of the data sets.
  - Distributed techniques are essential when the data cannot be gathered in one location.

Data mining tasks

- We consider the following core data mining tasks:
  - Predictive modeling
  - Association analysis
  - Cluster analysis

Data mining tasks



Data

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
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

Predictive modeling

- The objective of this task is to predict the value of a particular attribute based on the values of other attributes.
- The attribute to be predicted is commonly known as the target or dependent variable.
- The attributes used for making the prediction are known as the explanatory or independent variables.

Predictive modeling

- In particular, we need to build a model for the target variable as a function of the explanatory variables.
- There are two types of predictive modeling tasks
  - Classification, which is used for discrete target variables.
  - Regression, which is used for continuous target variables.
- The goal of both tasks is to learn a model that minimizes the error between the predicted and true values of the target variable.

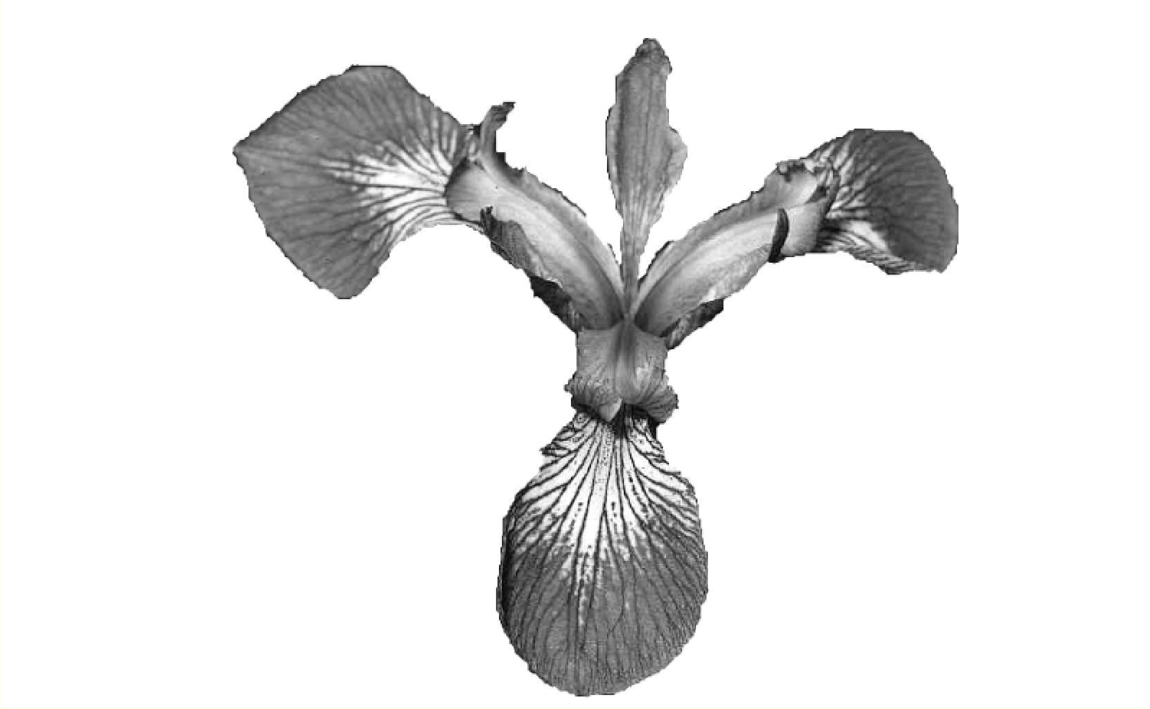
Predictive modeling

- Example of classification
  - Predicting whether a Web user will make a purchase at an online bookstore.
  - The target variable is binary-valued.
- Example of regression
  - Forecasting the future price of a stock.
  - Price is a continuous-valued variable.

Example: Flower species prediction

- We consider the task of predicting the species of a flower based on its characteristics.
- Specifically, we consider the classification of an Iris flower as to whether it belongs to one of the following species:
  - Setosa
  - Versicolour
  - Virginica

Example: Flower species prediction

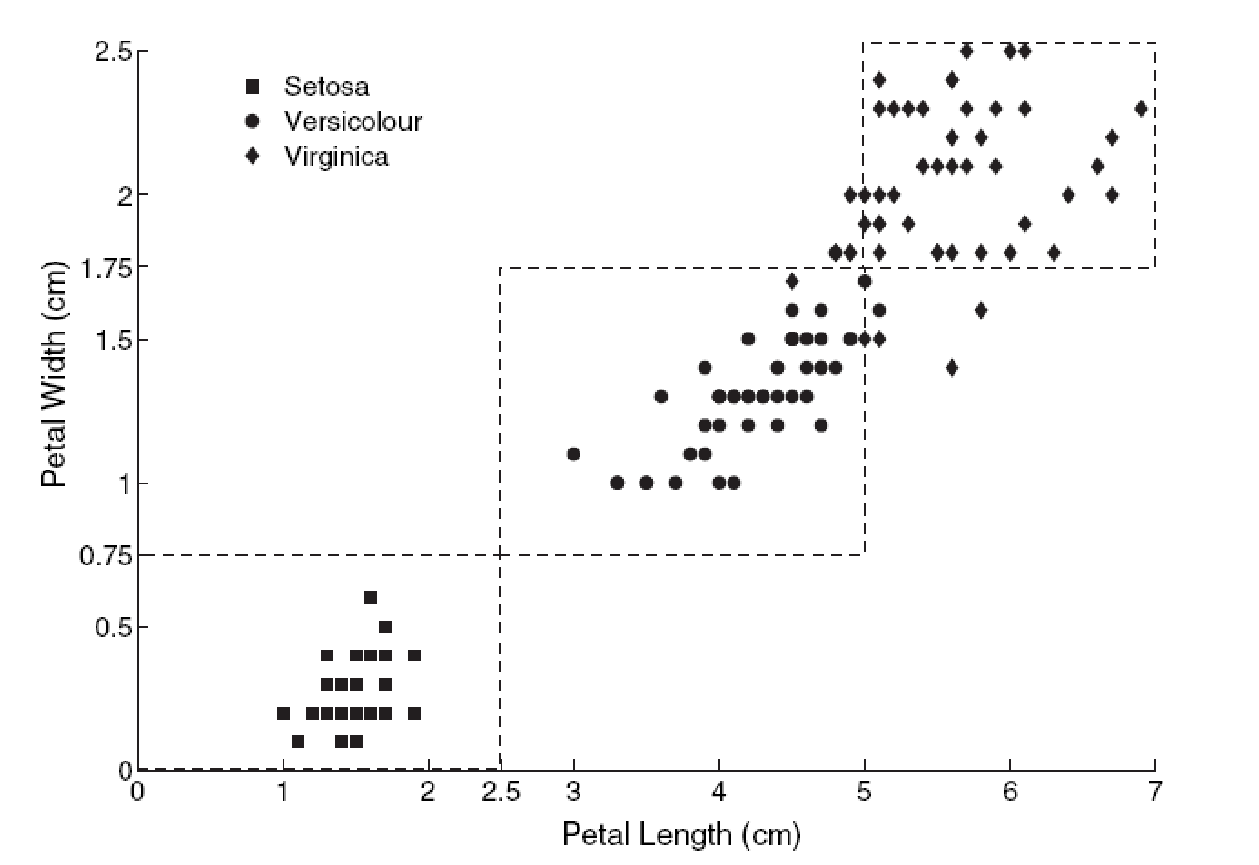


Picture of Iris Virginica

Predictive modeling

- We need a data set which contains the characteristics of various flowers of these three species.
- We can obtain these information from the well-known Iris data set from the UCI Machine Learning Repository.
- In addition to the species of a flower, this data set contains four other attributes:
  - Sepal length
  - Sepal width
  - Petal length
  - Petal width

Example: Flower species prediction



Petal width versus petal length for 150 Iris flowers

Example: Flower species prediction

- The figure shows a plot of petal width versus petal length for the 150 flowers in the data set.
- Petal width is broken into the following categories
  - Low: [0, 0.75)
  - Medium: [0.75, 1.75)
  - High: [1.75, ∞)
- Petal length is also broken into the following categories:
  - Low: [0, 2.5)
  - Medium: [2.5, 5)
  - High: [5, ∞)

Example: Flower species prediction

- Based on these categories, the following rules can be derived:
  - Petal width low and petal length low implies Setosa.
  - Petal width medium and petal length medium implies Versicolour.
  - Petal width high and petal length high implies Virginica.

Example: Flower species prediction

- These rules do not classify all the flowers correctly.
- Nevertheless, they are capable of classifying most of the flowers.
- Flowers from the Setosa species are well separated from the other two species with respect to petal width and length.
- However, the Versicolour and Virginica species overlap somewhat with respect to these two attributes.

Example: Web robot detection

- The main objective of Web usage mining is the discovery of useful patterns from Web access logs.
- These patterns can reveal interesting characteristics of site visitors.

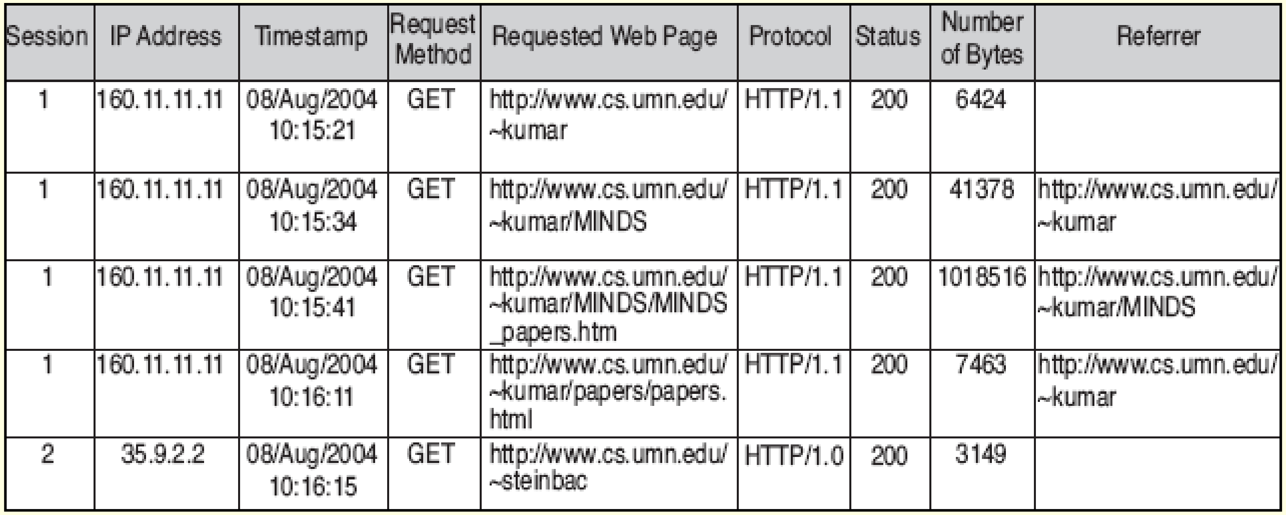
Example: Web robot detection

- A Web robot is a software program that automatically retrieves information by following the hyperlinks in Web pages.
- These programs are deployed by search engines to gather the documents necessary for indexing the Web
- In Web usage mining, it is important to distinguish accesses made by human users from those due to Web robots.

Example: Web robot detection

- Predictive modeling can be applied to distinguish between accesses by human users and those by Web robots.
- The input data was obtained from a Web server log.
- Each line corresponds to a single page request made by a Web client.
- A Web session is a sequence of requests made by a client during a single visit to a Web site.

Example: Web robot detection



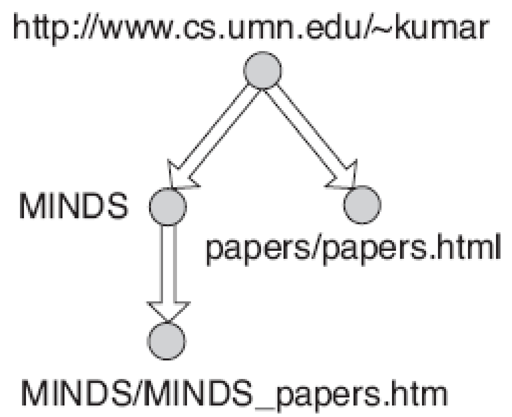
Session	IP Address	Timestamp	Request Method	Requested Web Page	Protocol	Status	Number of Bytes	Referrer
1	160.11.11.11	08/Aug/2004 10:15:21	GET	http://www.cs.umn.edu/ ~kumar	HTTP/1.1	200	6424	
1	160.11.11.11	08/Aug/2004 10:15:34	GET	http://www.cs.umn.edu/ ~kumar/MINDS	HTTP/1.1	200	41378	http://www.cs.umn.edu/ ~kumar
1	160.11.11.11	08/Aug/2004 10:15:41	GET	http://www.cs.umn.edu/ ~kumar/MINDS/MINDS _papers.htm	HTTP/1.1	200	1018516	http://www.cs.umn.edu/ ~kumar/MINDS
1	160.11.11.11	08/Aug/2004 10:16:11	GET	http://www.cs.umn.edu/ ~kumar/papers/papers. html	HTTP/1.1	200	7463	http://www.cs.umn.edu/ ~kumar
2	35.9.2.2	08/Aug/2004 10:16:15	GET	http://www.cs.umn.edu/ ~steinbac	HTTP/1.0	200	3149	

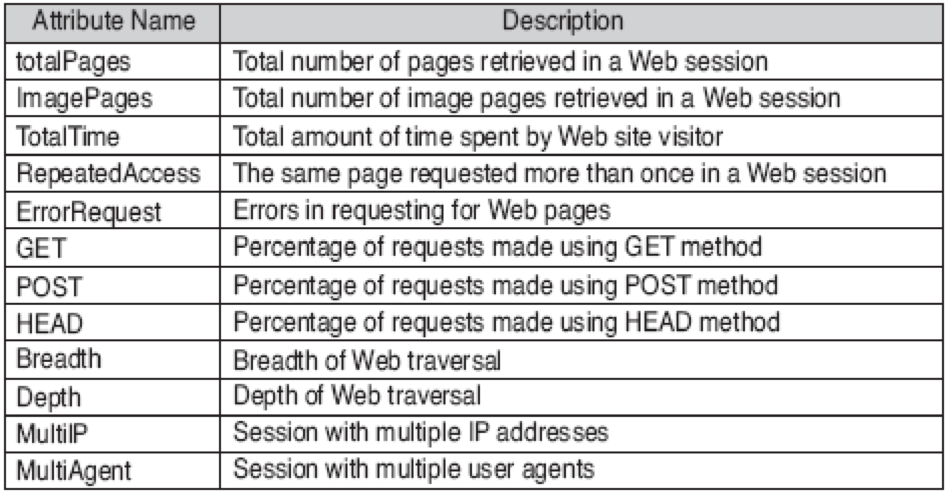
Example of a Web server log

Example: Web robot detection

- Each web session can be modeled as a directed graph, in which
  - The nodes correspond to Web pages and
  - The edges correspond to hyperlinks connecting one Web page to another.
- To classify the Web sessions, features are constructed to describe the characteristics of each session.
- A decision tree classifier is constructed to perform the classification.

Example: Web robot detection



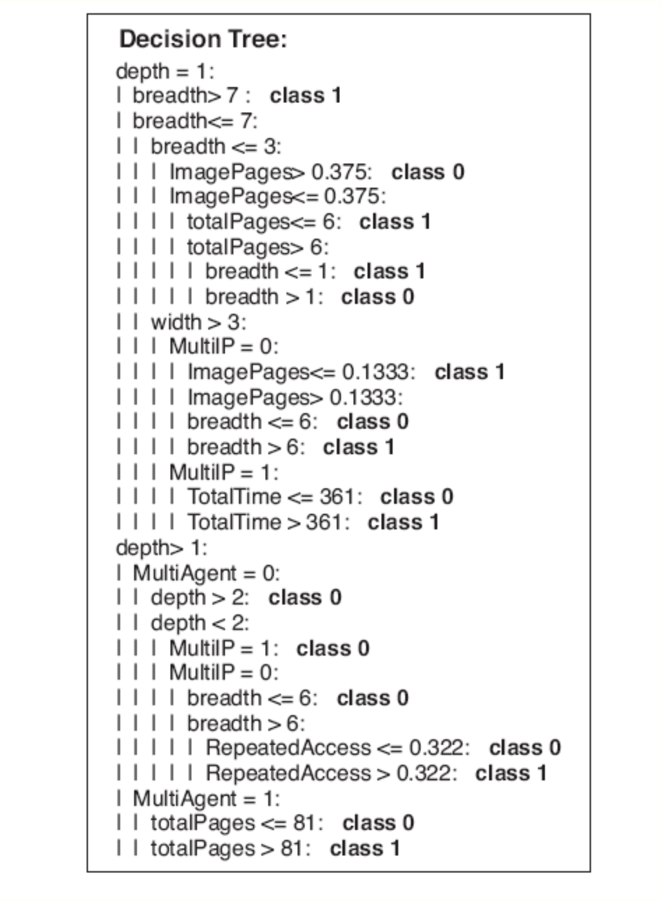


Attribute Name	Description
totalPages	Total number of pages retrieved in a Web session
ImagePages	Total number of image pages retrieved in a Web session
TotalTime	Total amount of time spent by Web site visitor
RepeatedAccess	The same page requested more than once in a Web session
ErrorRequest	Errors in requesting for Web pages
GET	Percentage of requests made using GET method
POST	Percentage of requests made using POST method
HEAD	Percentage of requests made using HEAD method
Breadth	Breadth of Web traversal
Depth	Depth of Web traversal
MultiIP	Session with multiple IP addresses
MultiAgent	Session with multiple user agents

Graph of a Web session

Attributes for Web robot detection

Example: Web robot detection



Decision Tree:

```
depth = 1:
| breadth> 7 :  class 1
| breadth<= 7:
| | breadth <= 3:
| | | ImagePages> 0.375:  class 0
| | | ImagePages<= 0.375:
| | | | totalPages<= 6:  class 1
| | | | totalPages> 6:
| | | | | breadth <= 1:  class 1
| | | | | breadth > 1:  class 0
| | width > 3:
| | | MultiIP = 0:
| | | | ImagePages<= 0.1333:  class 1
| | | | ImagePages> 0.1333:
| | | | breadth <= 6:  class 0
| | | | breadth > 6:  class 1
| | | MultiIP = 1:
| | | | TotalTime <= 361:  class 0
| | | | TotalTime > 361:  class 1
depth> 1:
| MultiAgent = 0:
| | depth > 2:  class 0
| | depth < 2:
| | | MultiIP = 1:  class 0
| | | MultiIP = 0:
| | | | breadth <= 6:  class 0
| | | | breadth > 6:
| | | | | RepeatedAccess <= 0.322:  class 0
| | | | | RepeatedAccess > 0.322:  class 1
| MultiAgent = 1:
| | totalPages <= 81:  class 0
| | totalPages > 81:  class 1
```

Decision tree for Web robot detection

Example: Web robot detection

- The model suggests that Web robots can be distinguished from human users in the following way
  - Accesses by Web robots tend to be broad but shallow, whereas accesses by human users tend to be more focused (narrow but deep).
  - Unlike human users, Web robots seldom retrieve the image pages associated with a Web document.
  - Sessions due to Web robots tend to be long and contain a large number of requested pages.

Association analysis

- Association analysis is used to discover patterns that describe strongly associated items in the data.
- The discovered patterns are typically represented in the form of implication rules or item subsets.
- The goal of association analysis is to extract the most interesting patterns in an efficient manner.

Association analysis

- Applications of association analysis include
  - Finding groups of genes that have related functionality.
  - Identifying Web pages that are accessed together.

Association analysis

- The transactions shown in the following table illustrate point-of-sale data collected at the checkout counter of a grocery store.
- Association analysis can be applied to find items that are frequently bought together by customers.

Association analysis

Transaction ID	Items
1	{Bread, Butter, Diapers, Milk}
2	{Coffee, Sugar, Cookies, Salmon}
3	{Bread, Butter, Coffee, Diapers, Milk, Eggs}
4	{Bread, Butter, Salmon, Chicken}
5	{Eggs, Bread, Butter}
6	{Salmon, Diapers, Milk}
7	{Bread, Tea, Sugar, Eggs}
8	{Coffee, Sugar, Chicken, Eggs}
9	{Bread, Diapers, Milk, Salt}
10	{Tea, Eggs, Cookies, Diapers, Milk}

Association analysis

- Example: We may discover a rule like {Diapers}→{Milk}
- This rule suggests that customers who buy diapers also tend to buy milk.
- This type of rule can be used to identify potential cross-selling opportunities among related items.

Cluster analysis

- Cluster analysis seeks to find groups of closely related observations.
- The observations that belong to the same cluster are more similar to each other than observations that belong to other clusters.
- Clustering can be applied to
  - Group sets of related customers.
  - Find groups of genes that have similar functions.
  - Compress data.

Cluster analysis

- The collection of news articles in the following table can be grouped based on their respective topics.
- Each article is represented as a set of word-frequency pairs (w,c)
  - w is a word
  - c is the number of times the word appears in the article.

Cluster analysis

Article	Words
1	dollar:1, industry:4, country: 2, load:3, deal: 2, government:2
2	machinery:2, labor:3, market:4, industry: 2, work:3, country:1
3	job:5, inflation:3, rise:2, jobless:2, market:3, country:2, index:3
4	domestic:3, forecast:2, gain:1, market:2, sale:3, price:2
5	patient:4, symptom:2, drug:3, health:2, clinic:2, doctor:2
6	pharmaceutical:2, company:3, drug:2, vaccine:1, flu:3
7	death:2, cancer:4, drug:3, public:4, health:3, director:2
8	medical:2, cost:3, increase:2, patient:2, health:3, care:1

Cluster analysis

- There are two natural clusters in the data set.
- The first cluster consists of the first four articles, which correspond to news about the economy.
- The second cluster contains the last four articles, which correspond to news about health care.
- A good clustering algorithm should be able to identify these two clusters.