COMP 482: Data Mining

—

Department of Computer Science and Engineering
Kathmandu University

Prerequisites

- Basic knowledge of algorithms
- Reasonable programming skill
  - C++, Java, Python, R will be helpful
- Basic linear algebra
- Basic probability and statistics
- Familiarity with databases

What will you learn?

- Basic concepts and techniques of data mining
- Common approaches for data preprocessing, feature generation/selection, and data wrangling
- Basic tools to apply basic data mining / machine learning algorithms
- Effective visualization techniques to describe data

Contents

Communication

Email: rajani.chulyadyo@ku.edu.np

Canvas: https://canvas.instructure.com/courses/4927354

Evaluation

- Internal evaluation
  - Assignments (Theoretical and Practical) - 20%
  - Quizzes - 10%
  - Exams - 30%
  - Mini Project - 25%
  - Viva - 15%
- End-semester exam

Books

1. Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques* (3rd ed.). Elsevier.
2. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann.
3. Tan, P.-N., Steinbach, M., Karpatne, A.,Kumar, V. (2019). *Introduction to Data Mining* (2nd ed.). Pearson.

Chapter 1: Introduction to Data Mining

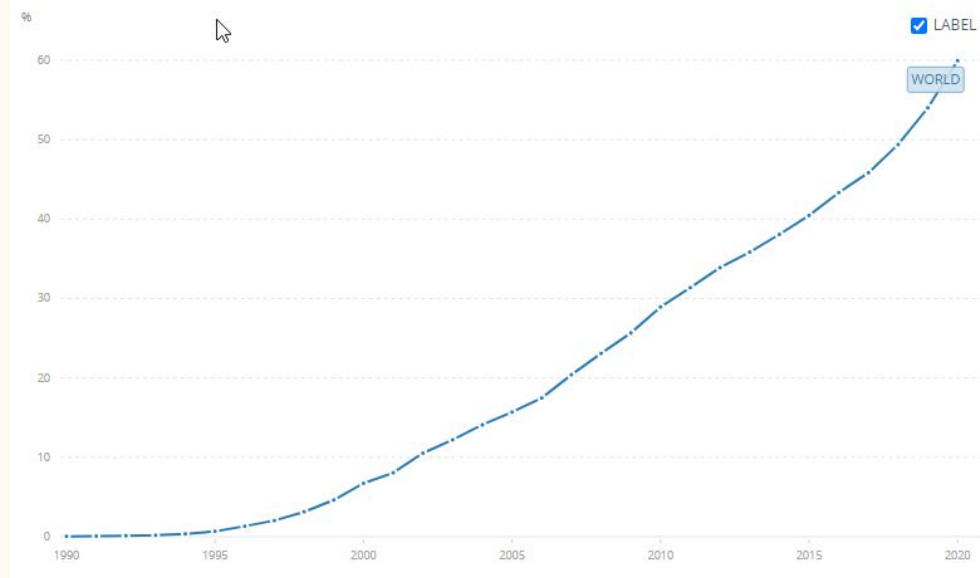# Increase in Internet users a global phenomenon

Fig: Individuals using the Internet (% of population)
(Source: https://data.worldbank.org/indicator/IT.NET.USER.ZS)

Increase in websites

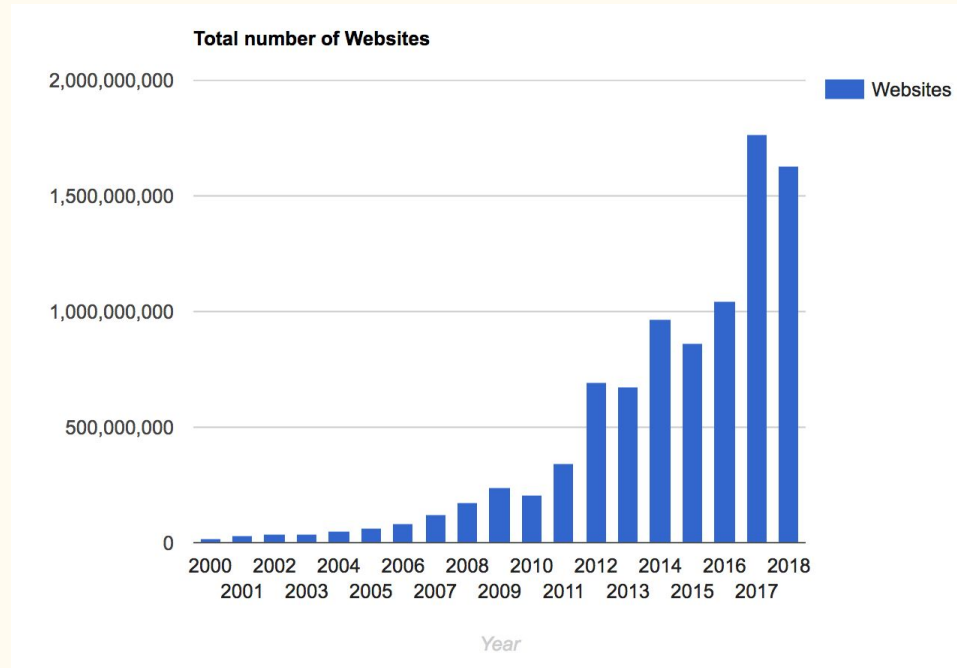Fig: Total number of websites by year

Abundance of data

Computerization of our society, and fast development of technologies for collecting, disseminating, processing, and analyzing data resulted in a tremendous volume of data day by day.

Sources of data:

- Online data: Internet users' activities (browsing, searching, blogging, sharing photos/videos, online volunteering/collaboration, surveying etc.), e-commerce, social networks, etc.

- Offline data: environmental, biological, and economical studies, business, medical/health data (e.g., medical records, patient monitoring, medical imaging etc.), various applications etc.

What is Data Mining?

Data mining is defined as the process of **discovering patterns** in data, automatically or semi-automatically, in **large quantities of data**. The patterns discovered must be **meaningful** in that they lead to some advantage, usually an economic one. [Witten et al. 2011]

Data mining is the process of discovering insightful, interesting, and novel **patterns**, as well as descriptive, understandable, and predictive models from **large-scale data**. [Zaki et al. 2014]

Data mining is the process of discovering interesting **patterns** and **knowledge** from **large amounts of data**. [Han et al. 2012]

What is pattern/knowledge?

A kind of summary of the input data, which may be used in further analysis

Examples of patterns

1. Groups of data records,
2. Unusual records
3. Dependencies etc.

An interesting pattern represents **knowledge**

Interestingness of the extracted patterns

A pattern is interesting if it is

1. **Understandable**: humans should be able to interpret the pattern
2. **Valid**: hold on new or test data with some degree of certainty
3. **Potentially useful**: should be possible to act on the item
4. **Novel/unexpected**: non-obvious to the system

What is not data mining?

- Finding a certain person in an employee database
- Computing the minimum, maximum, sum, count or average values based on table/tables columns
- Using a search engine to find your name occurrences on the web etc.

Interestingness of the extracted patterns

...the curse of big data is the fact that when you search for patterns in very, very large data sets with billions or trillions of data points and thousands of metrics, you are bound to identify coincidences that have no predictive power."

[The curse of big data: Vincent Granville](#)

Bonferroni's principle

Roughly, if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap [Leskovec et al. 2014]

Data Mining

Other names [Han et al. 2012]:

1. Knowledge discovery from data (KDD)
2. Knowledge mining from data
3. Knowledge extraction
4. Data / pattern analysis
5. Data archaeology
6. Data dredging etc.
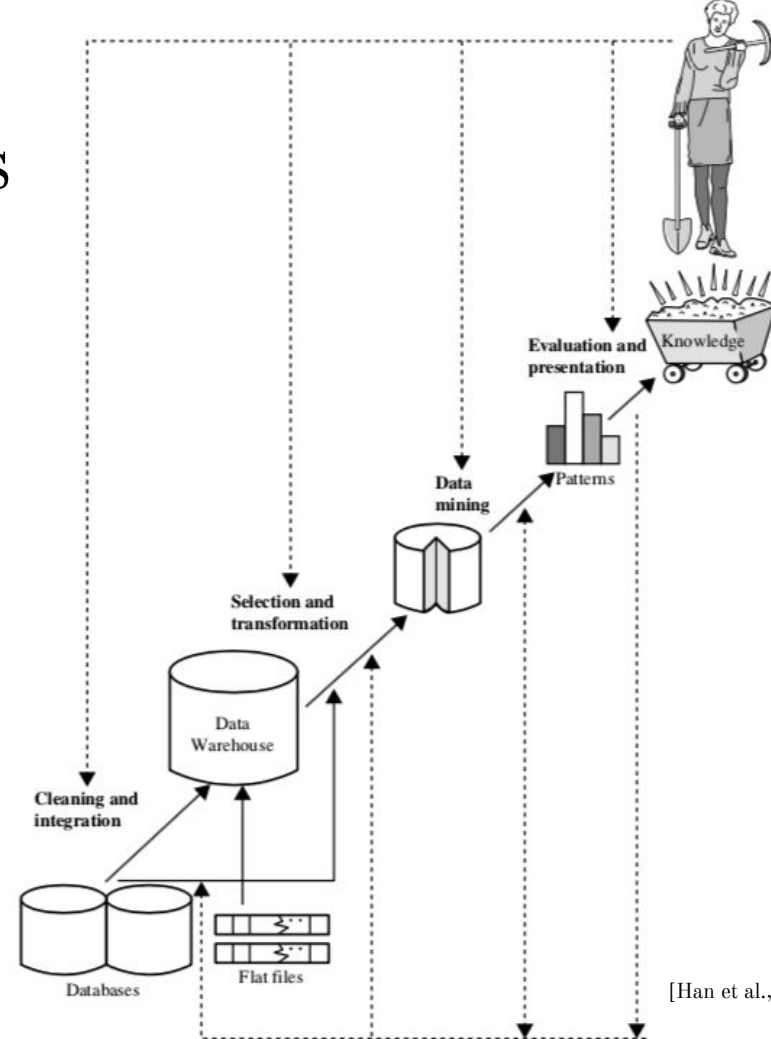
Knowledge discovery process

1. Data preprocessing
   - Data cleaning
   - Data integration
   - Data selection
   - Data transformation
2. Data mining
3. Data post-processing
   - Pattern evaluation
   - Pattern interpretation
   - Knowledge representation / Visualization

Evaluation and presentation

Knowledge

Data mining

Patterns

Selection and transformation

Data Warehouse

Cleaning and integration

Databases   Flat files

[Han et al., 2012]

Knowledge discovery process

- Data selection
  - Selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed
- Data cleaning and preprocessing
  - Removal of noise or outliers
  - Collecting necessary information to model or account for noise
  - Strategies for handling missing data fields
  - Accounting for time sequence information and known changes

Knowledge discovery process

- Data transformation
  - Transforming data into forms appropriate for mining
  - Dimensionality reduction
- Data mining
  - Applying intelligent methods to extract data patterns
- Interpretation, presentation, and evaluation
  - Using visualization and knowledge representation techniques to present mined knowledge to users
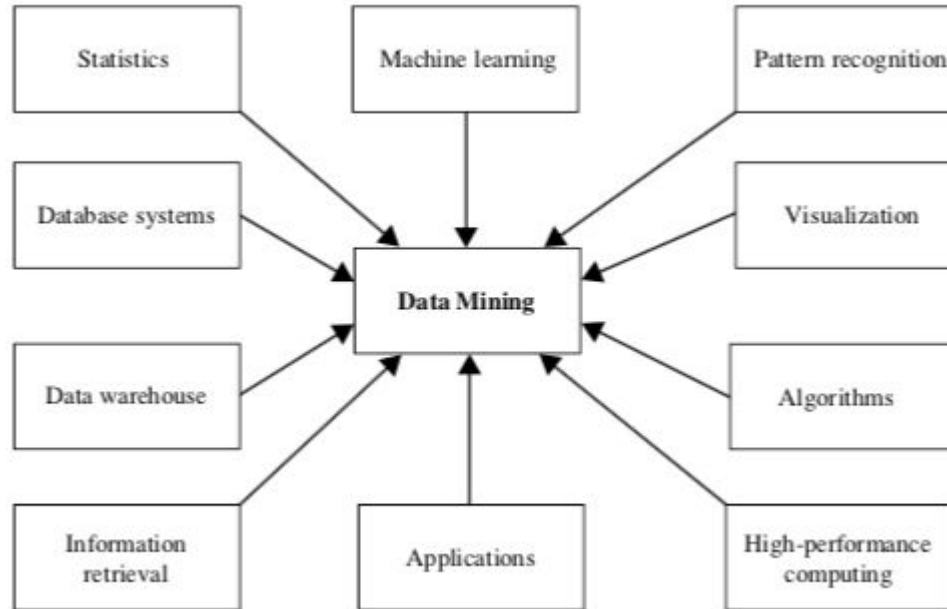
What kinds of data can be mined?

- Database data
  - Data stored in databases
- Data warehouses
  - A repository of information collected from multiple sources, stored under a unified schema, and usually residing at single site
  - Maintained separately from the operational databases
  - Constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing
- Transactional data
- Other kinds of data
  - Time-related or sequence data (e.g., historical records, biological sequence data etc.)
  - Data streams (e.g., video surveillance data)
  - Spatial data (e.g., maps)
  - Hypertext, multimedia data etc.
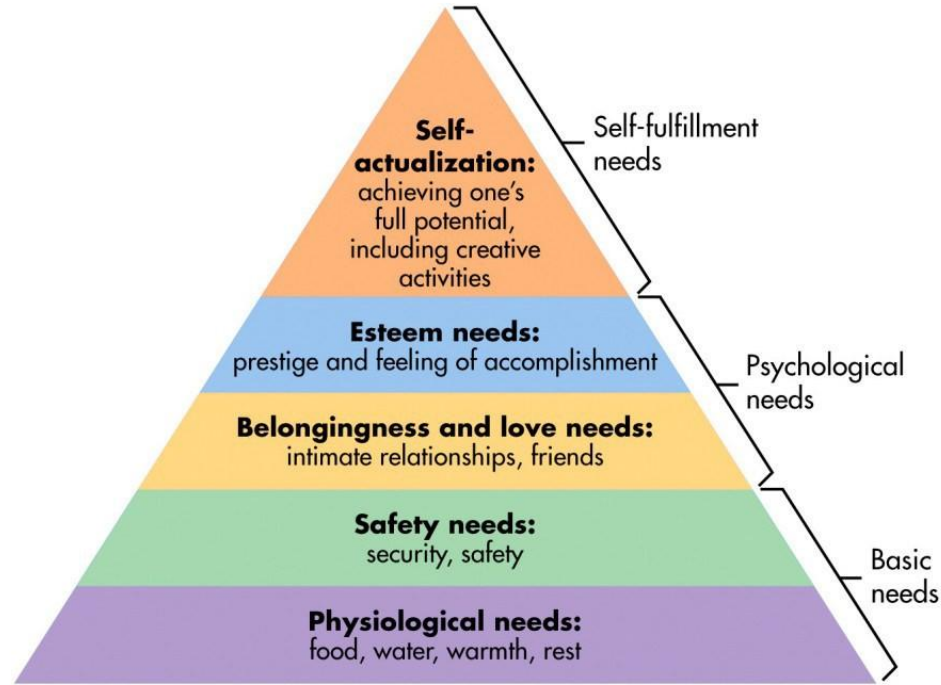
Data mining tasks

Two "high-level" primary goals of data mining, in practice:

- **Description**
  - Find human-interpretable patterns that describe the data
  - Example: Clustering, association rules etc.
- **Prediction**
  - Use some variables to predict unknown or future values of other variables
  - Example: Classification, recommender systems etc.

Technologies adopted

Statistics	Machine learning	Pattern recognition
Database systems		Visualization
	Data Mining	
Data warehouse		Algorithms
Information retrieval	Applications	High-performance computing

[Han et al., 2012]

Maslow's Hierarchy of Needs

https://www.simplypsychology.org/maslow.html

The Data Science Hierarchy of Needs

https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007
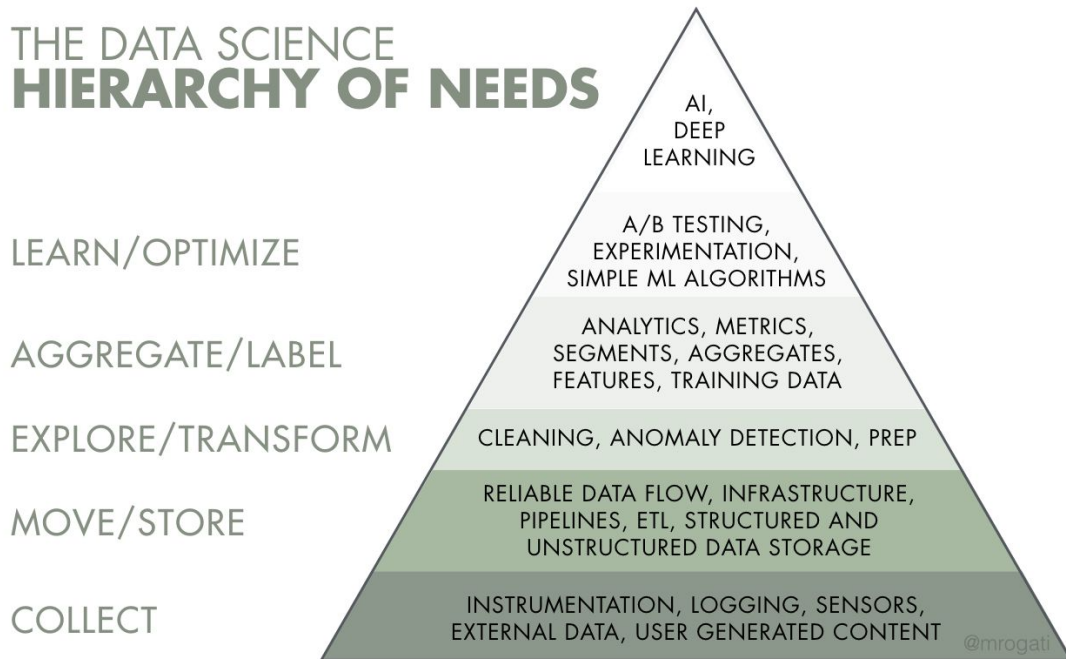
Tools

- Orange
- Weka
- TANAGRA
- RapidMiner
- SAS Enterprise Miner
- KNIME

- KEEL
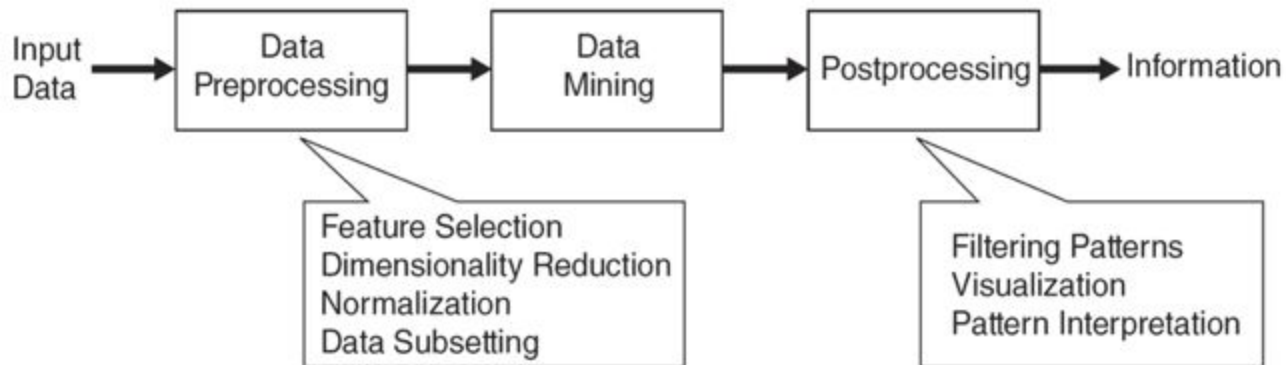- IBM SPSS Modeler
- Coheris SPAD
- R
- Matlab, Octave etc.

For more, check https://www.kdnuggets.com/software/index.html

Data repositories

- [UCI KDD repository](#)
- [UCI machine learning repository](#)
- [Yahoo sandbox datasets](#)
- [Yelp academic datasets](#)
- [Kaggle datasets](#)

For more, check https://www.kdnuggets.com/datasets/index.html

Knowledge Discovery Process (Revisited)

Input Data → Data Preprocessing → Data Mining → Postprocessing → Information

Data Preprocessing:
Feature Selection
Dimensionality Reduction
Normalization
Data Subsetting

Postprocessing:
Filtering Patterns
Visualization
Pattern Interpretation

[Tan et al., 2016]

Input

Data can often be represented or abstracted as an **_n_ x _d_ data matrix** (aka a **matrix of instances versus attributes**, or a **single relation** or a **flat file**), with n rows and d columns.

$$\mathbf{D} = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

Where x_i denotes the ith row, which is a d-tuple given as $x_i = (x_{i1}, x_{i2}, ..., x_{id})$
And X_j denotes the jth column, which is an n-tuple given as $X_j = (x_{1j}, x_{2j}, ..., x_{nj})$

Input

$$D = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

Depending on the application domain, rows may be referred to as *entities, instances, records, examples, transactions, objects, feature-vectors, tuples*, and so on.

Likewise, columns may also be called *attributes, properties, features, dimensions, variables, fields*, and so on.

The number of instances n is referred to as the **size** of the data.

The number of attributes d is called the **dimensionality** of the data.

Problems often involve relationships between objects rather than separate, independent instances. Mining such data = **relational data mining**

Input: Example

Instance is characterized by the values of a set of predetermined **attributes**

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no

Input: Attributes

- **Continuous / numeric**: measures numbers, either real or integer valued. Example: temperature and humidity in the following dataset
- **Nominal / categorical / enumerated / discrete**: takes on values in a prespecified, finite set of possibilities. Example: outlook

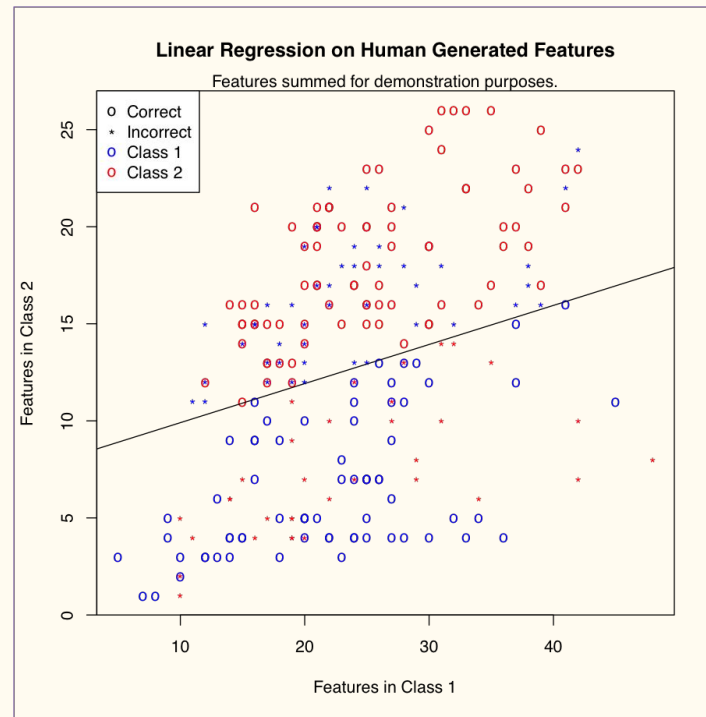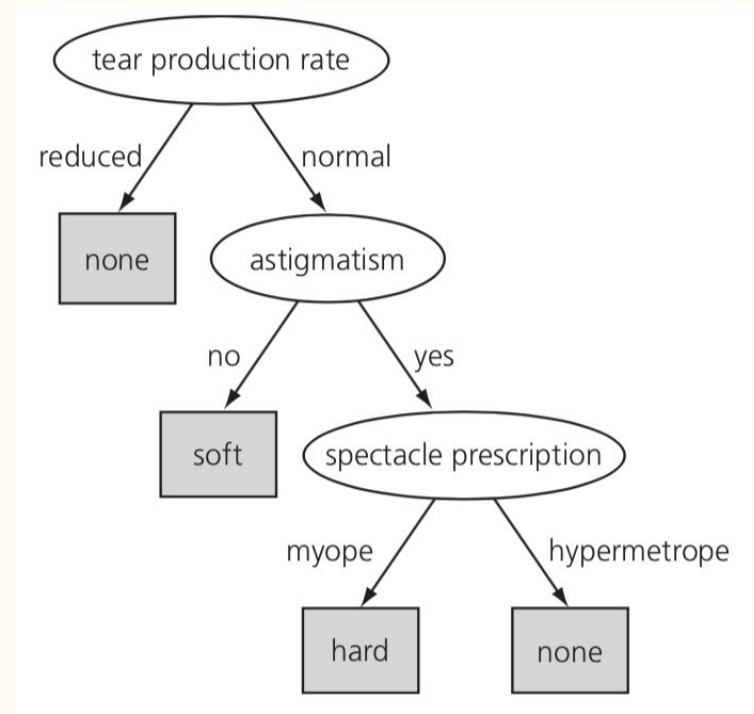| Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes

Output

Common ways of representing the output:

- Linear models
- Trees
- Rules
  - Classification rules
  - Association rules
- Clusters

Output: Linear models

Linear Regression on Human Generated Features
Features summed for demonstration purposes.

Output: Trees

Output: Rules

Classification rules:

- A popular alternative to decision trees
- Contains two parts:
  - Antecedent: a series of tests just like the tests at nodes in decision trees
  - Consequent: gives the class or classes that apply to instances covered by that rule

```
If outlook = sunny and humidity = high  then play = no
If outlook = rainy and windy = true     then play = no
If outlook = overcast                   then play = yes
If humidity = normal                    then play = yes
If none of the above                    then play = yes
```
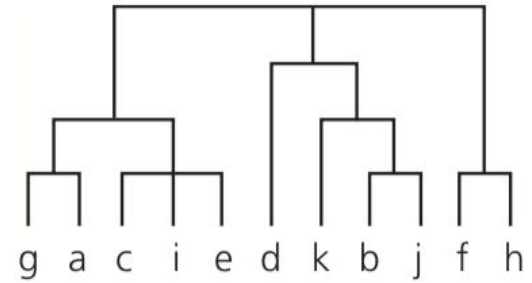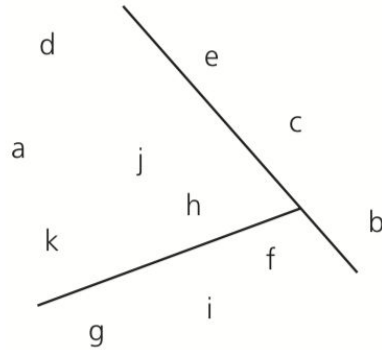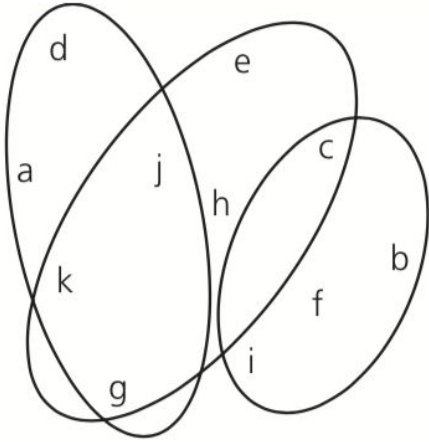
Output: Rules

Association rules

- Unlike classification rules, association rules can predict any attribute, not just the class

```
If windy = false and play = no then outlook = sunny
                                        and humidity = high
```

Output: Clusters

Clustering involves associating a cluster number with each instance

Use of data mining in real-world

"Diapers and beer"
Observation that customers who buy diapers are more likely to by beer than average allowed supermarkets to place beer and diapers nearby, knowing many customers would walk between them. Placing potato chips between increased sales of all three items.

Decision trees constructed from bank-loan histories to produce algorithms to decide whether to grant a loan.

Alibaba established **a model for monitoring, analyzing and cracking down on counterfeit goods systems** and is currently in use when cooperating with the police to boycott counterfeit goods.

Use of data mining in real-world

Patterns of travellers' behavior mined to manage the sale of discounted seats on planes, rooms in hotels, etc.

Data mining is also used to **forecast crime scenes**, and to make strategies for reducing crime rates.

Sources:

https://www.octoparse.com/blog/data-mining-explained-with-10-interesting-stories
http://infolab.stanford.edu/~ullman/mining/allnotes.pdf

More data mining success stories:

http://dml.cs.byu.edu/~cgc/docs/mldm_tools/Reading/DMSuccessStories.html

Data mining and ethics

- The use of data—particularly data about people—for data mining has serious ethical implications
- Data mining is frequently used to discriminate
- Certain kinds of discrimination (racial, sexual, religious, and so on) are not only unethical but also illegal
  - However, it depends on the application
  - Using sexual and racial information for medical diagnosis is certainly ethical, but using the same information when mining loan payment behavior is not

Data mining and ethics

Machine bias

- Biased data $\Rightarrow$ biased models
  - Google's face recognition software tagged a lot of black faces as gorillas
  - On LinkedIn, high-paying jobs were not displayed as frequently for women as they were for men
  - There were fewer Pokémon locations in primarily black neighborhoods
  - More: https://github.com/MachineBias/machinebias.github.io
- Ensure that the data is checked for bias: no simple task

Data mining and ethics

Privacy

- Privacy-preserving data mining (PPDM): The philosophy of PPDM is to observe data sensitivity and preserve people's privacy while performing successful data mining

General Data Protection Regulation (GDPR) compliance

- GDPR directly impacts any company handling EU residents' personal data

"... data mining is just a tool in the whole process: it is people who take the results, along with other knowledge, and decide what action to apply." [Witten et al 2014]

References

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques (3rd ed.). Elsevier.

Tan, P.-N., Steinbach, M., Karpatne, A.,Kumar, V. (2019). Introduction to Data Mining (2nd ed.). Pearson.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques (4th ed.). Morgan Kaufmann.

Zaki, M. J., Meira, W. (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.