Introduction to Data Mining Methods and Models

Dr. Yuzana Win (Nagasaki University, Japan)
Lecturer
Department of Computer Engineering and Information Technology

Basic Course Information

- Course Title
  - ***Introduction to Data Mining Methods and Models***
- Textbook and Reference Materials
  - ***Introduction to Data Mining***, First Edition by Pang-ning Tan, Michael Steinbach, Vipin Kumar
  - ***Data Mining Methods and Models,*** by Daniel T. Larose
  - ***Data Mining***, Concepts and Techniques, by J. Han and M. Kamber
  - ***Introduction to Machine Learning with Python***, First Edition by Andreass C. Mullter & Sarach Guido, 2016
- Course Duration
  - ***14 Weeks***

Course Objectives

- To understand students with comprehensive ***knowledge of data mining techniques and method***
- To analysis the issues of ***data pre-processing step*** for a large amount of data
- To have skills to ***demonstrate knowledge of basic algorithms***, in particular the application of ***supervised*** and ***unsupervised*** algorithms
- To apply popular algorithms of data mining to ***find important information or knowledge*** from real-world data sources

CEIT
Think smart, be different.

Course Schedule

Week	Topic
Week 1	Introduction to Data Mining and Methods
Week 2	Type of Data and Data Pre-processing
Week 3	Supervised Learning: Classification with Nearest-Neighbor Classifier
Week 4	Supervised Learning: Classification with Naïve Bayes Classifier
Week 5	Supervised Learning: Classification with Decision Tree
Week 6	Supervised Learning: Classification with Artificial Neural Network (ANN)
Week 7	Supervised Learning: Classification with Support Vector Machine (SVM)

Course Schedule

Week	Topic
Week 8	Supervised Learning: Simple Regression
Week 9	Unsupervised Learning: Association Analysis with Apriori Algorithm
Week 10	Unsupervised Learning: Association Analysis with FP-Growth Algorithm
Week 11	Unsupervised Learning: Cluster Analysis with K-means
Week 12	Dimensionality Reduction: Principal Component Analysis (PCA)
Week 13	Model Evaluation and Improvement
Week 14	Practical: Working with Text Data

CEIT
Think smart, be different.

Exam and Grading System

Task	Mark
Final Exam	80%
Attendance	10%
Assignment	10%

Mark	Grade
85 ~ 100	A+
80 ~ 85	A
70 ~ 80	B
50 ~ 70	C
<50	F

CEIT
Think smart, be different.

Lecture 1
Introduction to Data Mining and Methods

Lecture Objectives

- To introduce
  - What is Data Mining
  - How Data Mining is used for
  - Application of Data Mining
  - Data Mining Methods and Models

What is Data Mining?

- Data mining that finds "important information or knowledge" from real-world data sources, e.g., databases, **texts**, images and the **Web**, etc.

- Data Mining is the process of automatically searching large stores of data to discover patterns and evaluate the probability of future events.

What is Data Mining?

- Data Mining is also known as Knowledge Discovery in Data (KDD)
- Knowledge on data analysis is useful in massive data for helping people to have a better insight on it

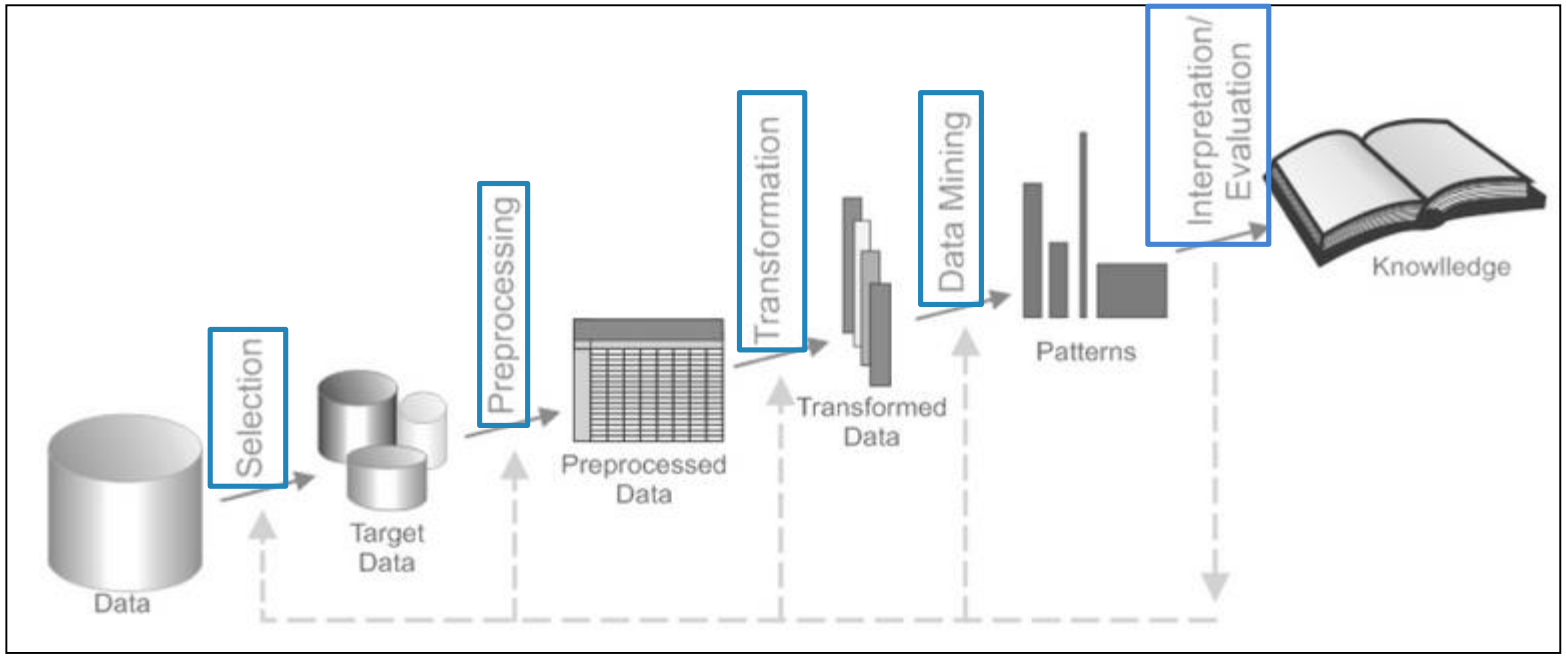
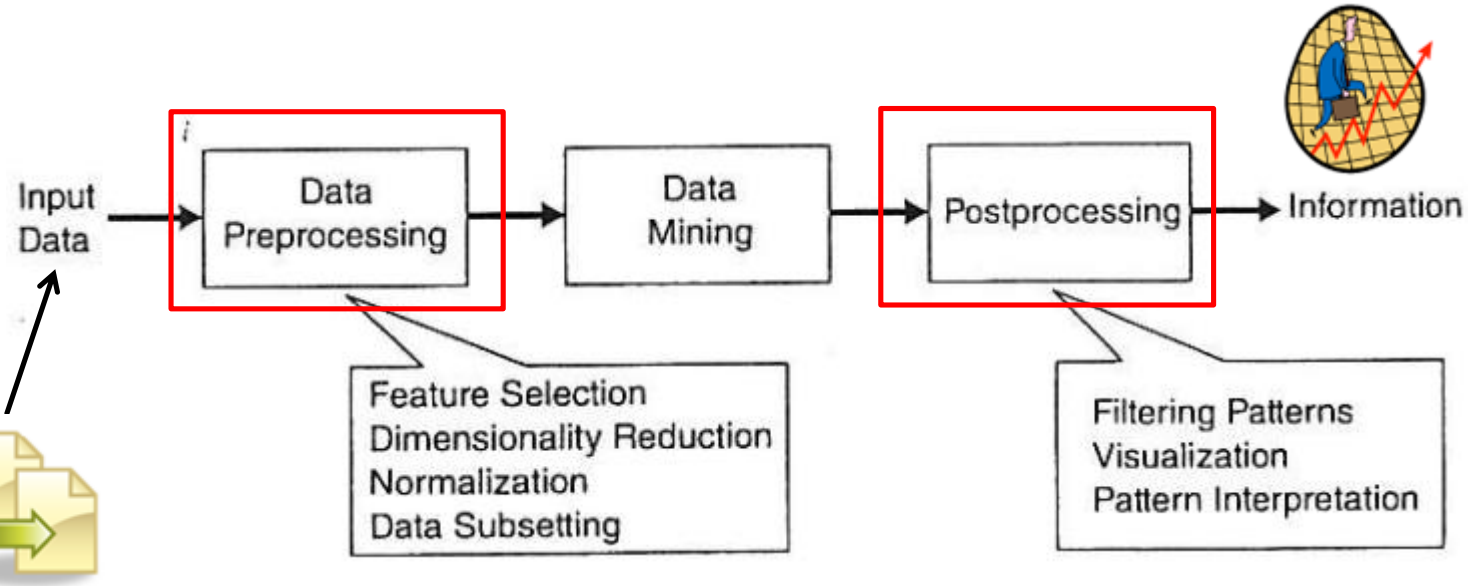


Figure: Stages in the Knowledge Discovery Process (KDD)
Source: Fayyad et al. (1995)

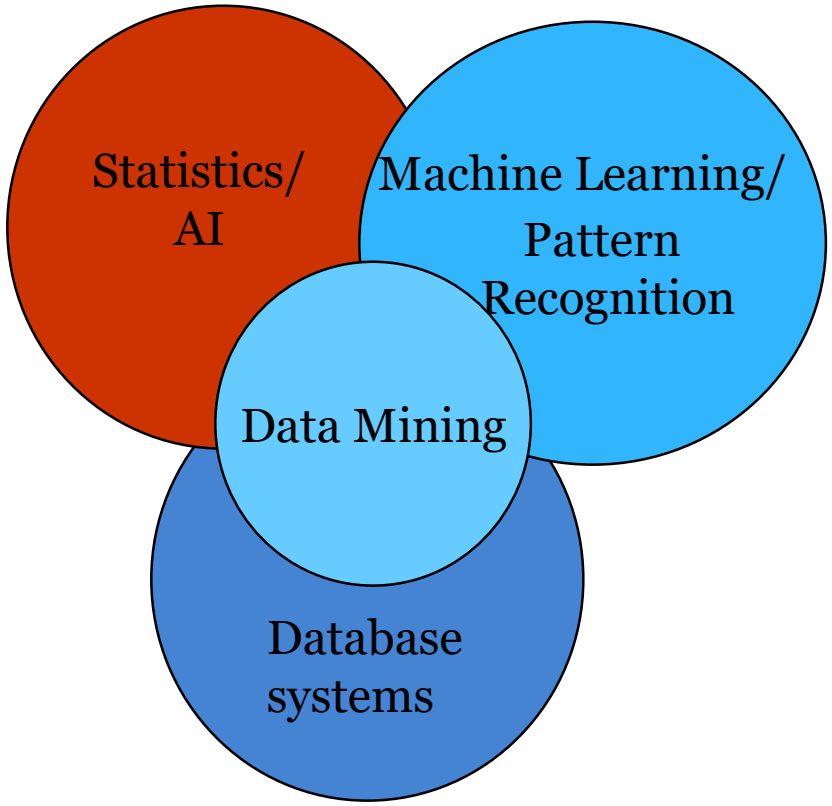
Data Mining and Knowledge Discovery



flat files, spread-sheets or relational tables

How Data Mining is used for?

- The ideal: Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Realistically: Process incoming data, deduce the future, based on previous/past patterns, i.e., learn from past 'experience' (data)
- In order to learn/deduce, system needs to be 'trained', i.e., what data to look at, how to process it, and how to output the results



CEIT
Think smart, be different.

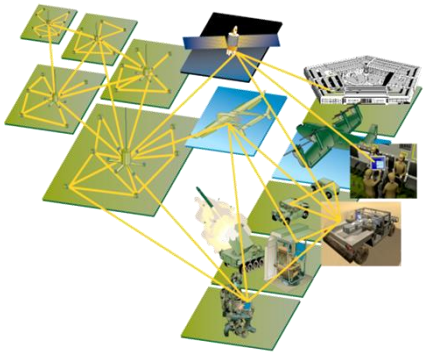
Application of Data Mining

Sales/Marketing

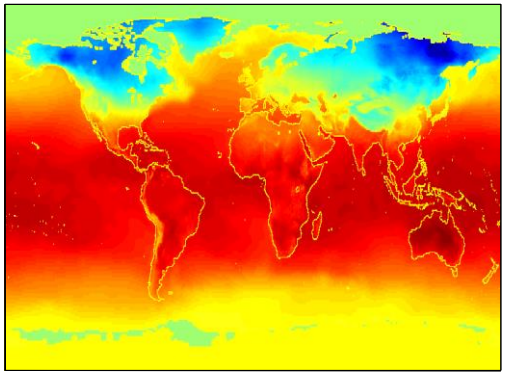
Banking/Finance

Health Care and Insurance

Medicine



Transportation



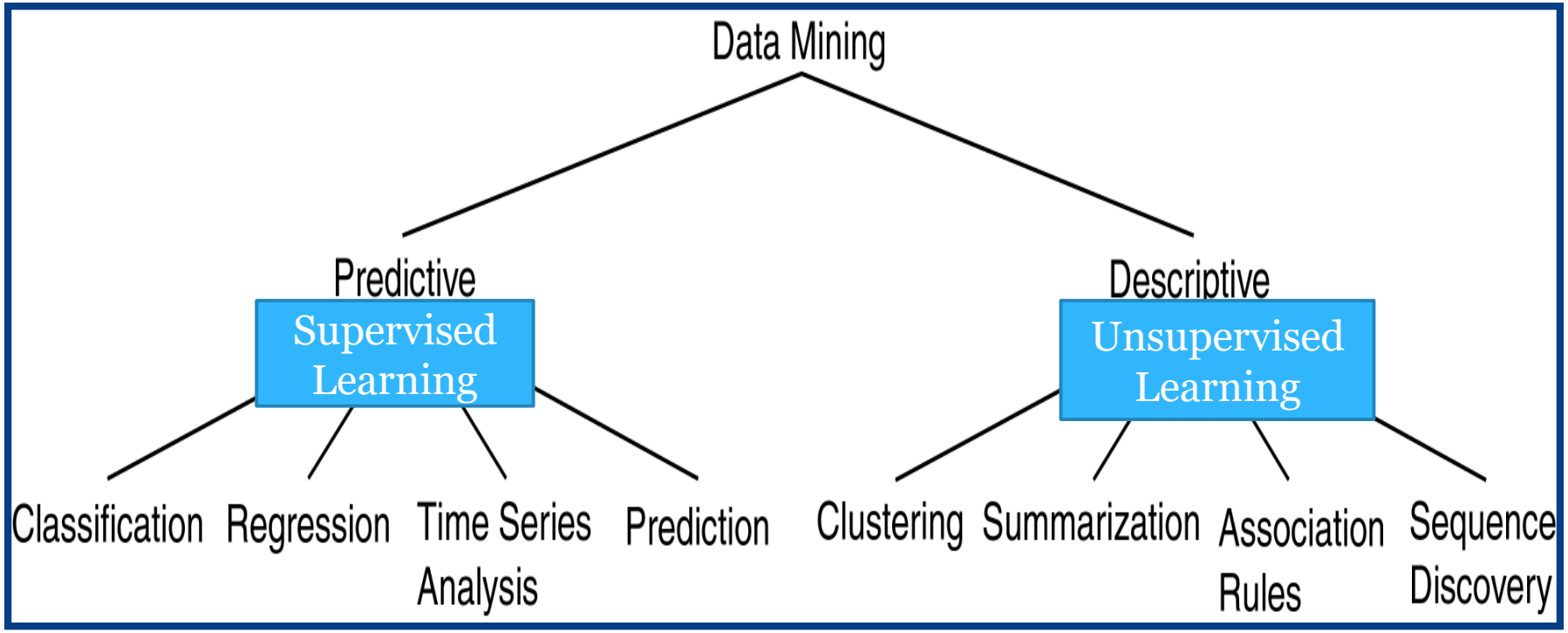
Research Analysis

Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables
- Description Methods
  - Find human-interpretable patterns that describe the data

CEIT
Think smart, be different.

Data Mining Methods and Models



Classification

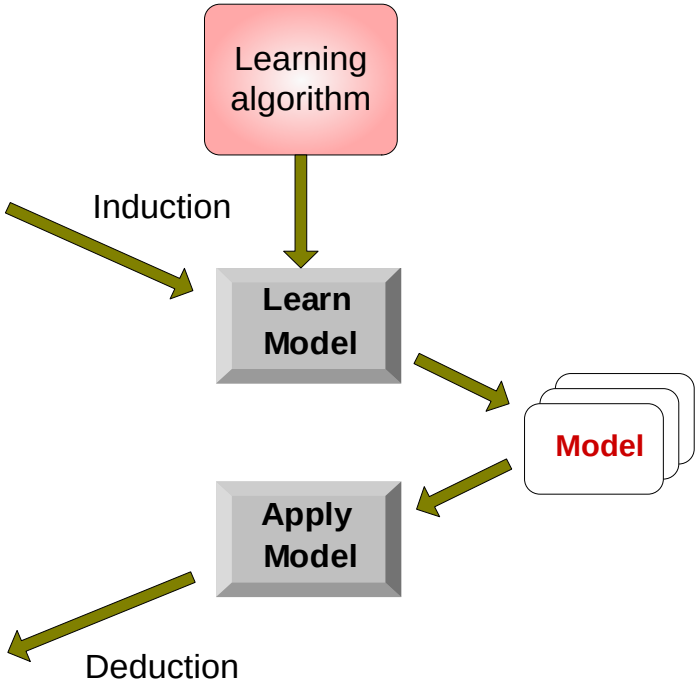
- Classification maps data into predefined groups or classes
  - Goal: previously unseen records should be assigned a class as accurately as possible

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Examples of Classification Task

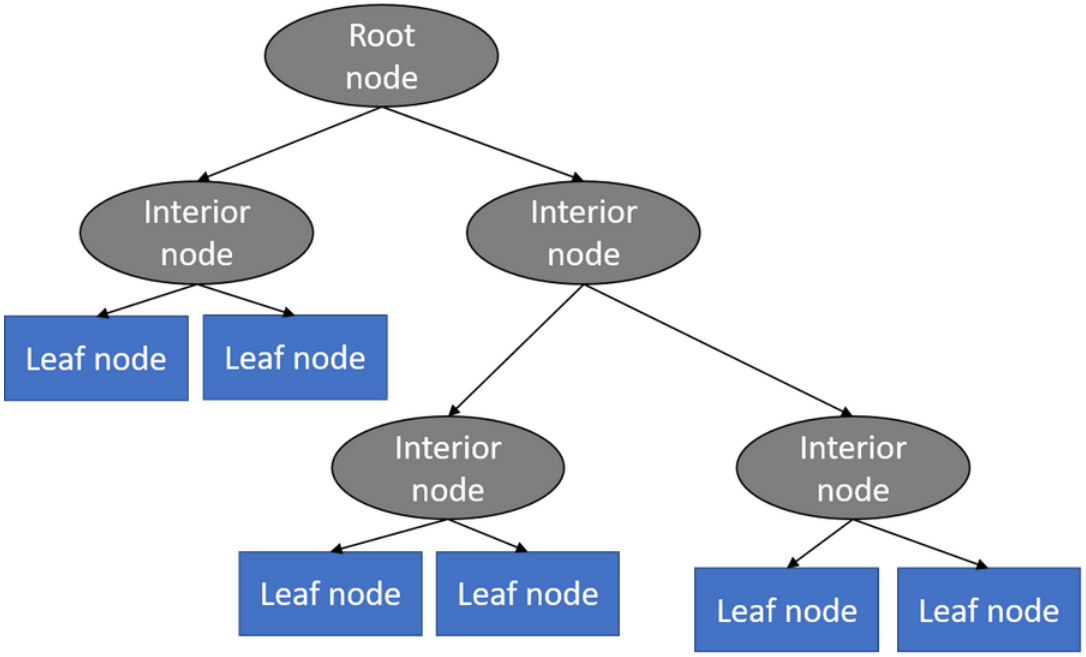
- Categorizing news articles as finance, weather, entertainment, sports, etc.
- Predicting tumor cells as benign or malignant
- Classifying credit card transactions as legitimate or fraudulent

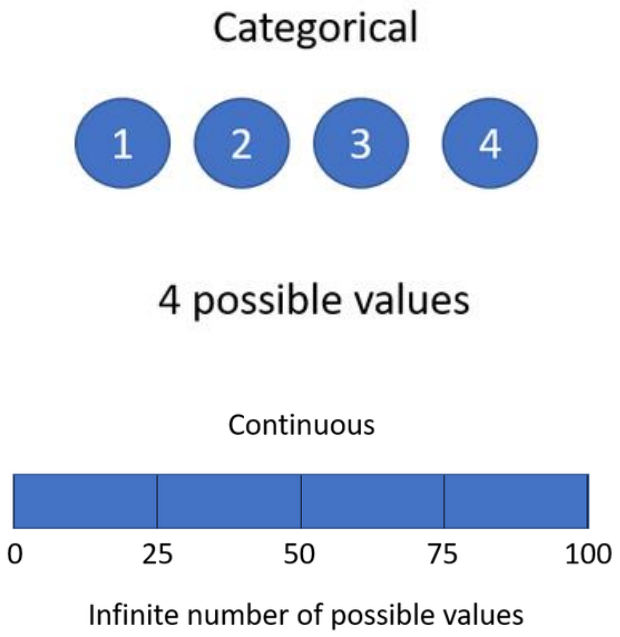
Classification Techniques or Methods

- Naïve Bayes Classifier *
- Support Vector Machines *
- Neural Networks (Deep Learning)
- Decision Tree based Methods
- Rule-based Methods
- K-Nearest Neighbors *

Decision Tree Method

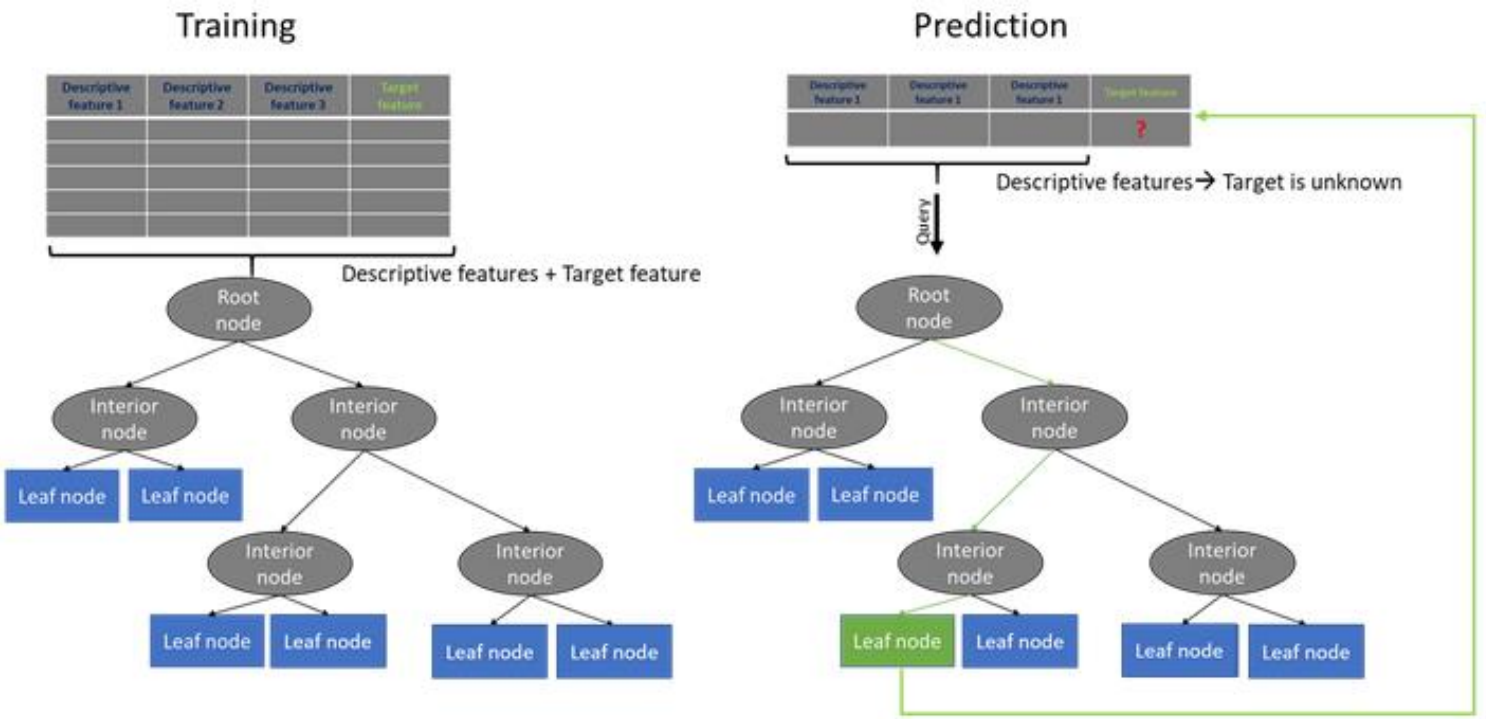
- Decision trees are supervised learning algorithms used for both, classification and regression tasks
- The main idea of decision trees is to find the most "information".
- A decision tree mainly contains of a **root node**, **interior nodes**, and **leaf nodes** which are then connected by **branches**.





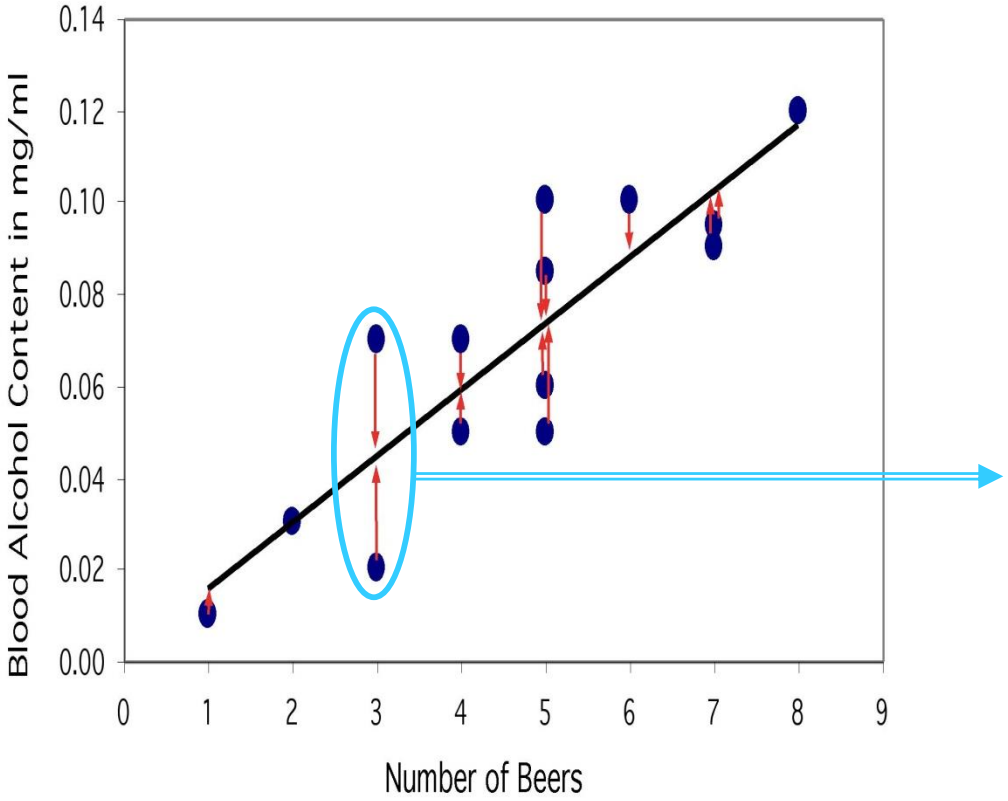
Decision Tree Method

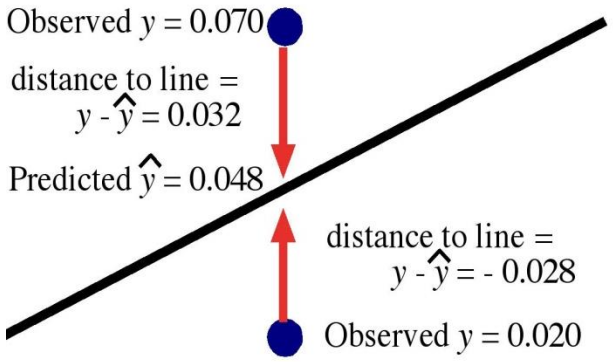
1. Present a dataset containing of a number of training instances characterized by a number of descriptive features and a target feature
2. Train the decision tree model by using a measure of information gain during the training process
3. Grow the tree until we accomplish a stopping criteria --> create leaf nodes which represent the *predictions* we want to make for new query instances
4. Show query instances to the tree and run down the tree until we arrive at leaf nodes
5. DONE



Regression

- Regression is used to map a data item to a real valued prediction variable.



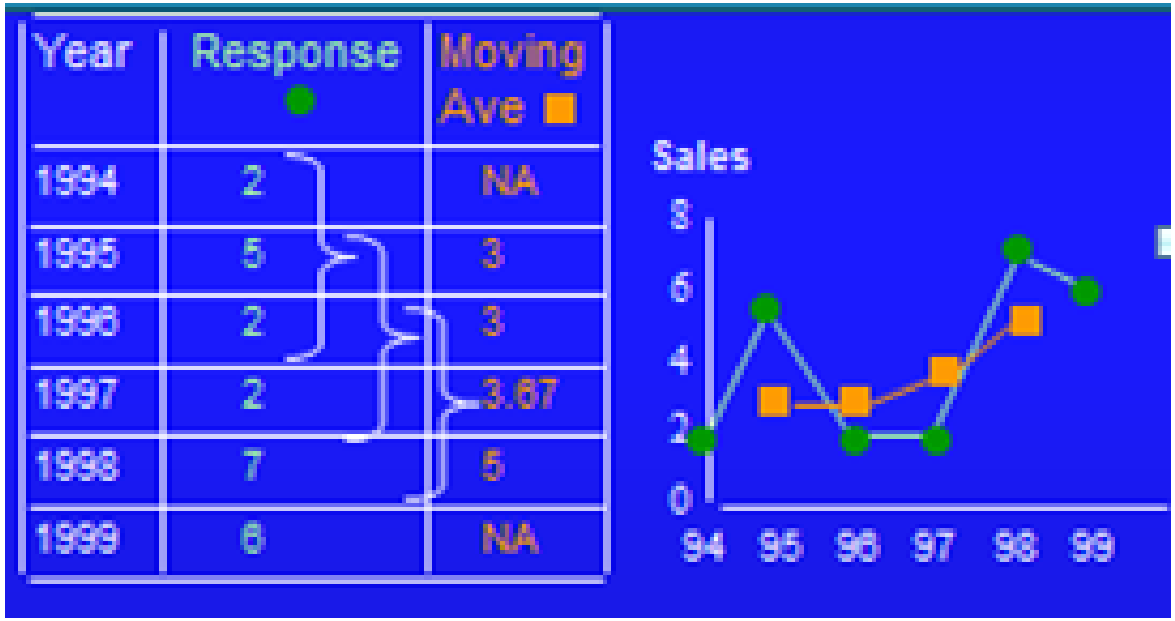


Regression Techniques or Methods

- Linear Regression
- Multiple Regression
- Logistic Regression
- LASSO

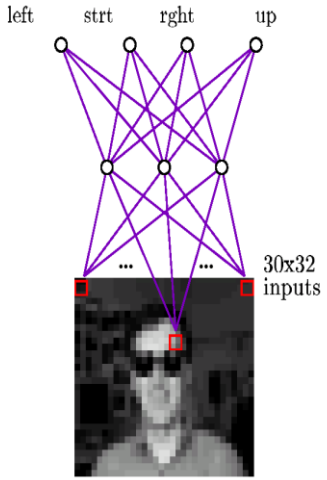
Time-Series

- Regression is used to map a data item to a real valued prediction variable.
- A quantitative forecasting method to predict future values (based on past and present observations)



Prediction

- Prediction attempts to form patterns that permit it to predict the next events(s) given the available input data.



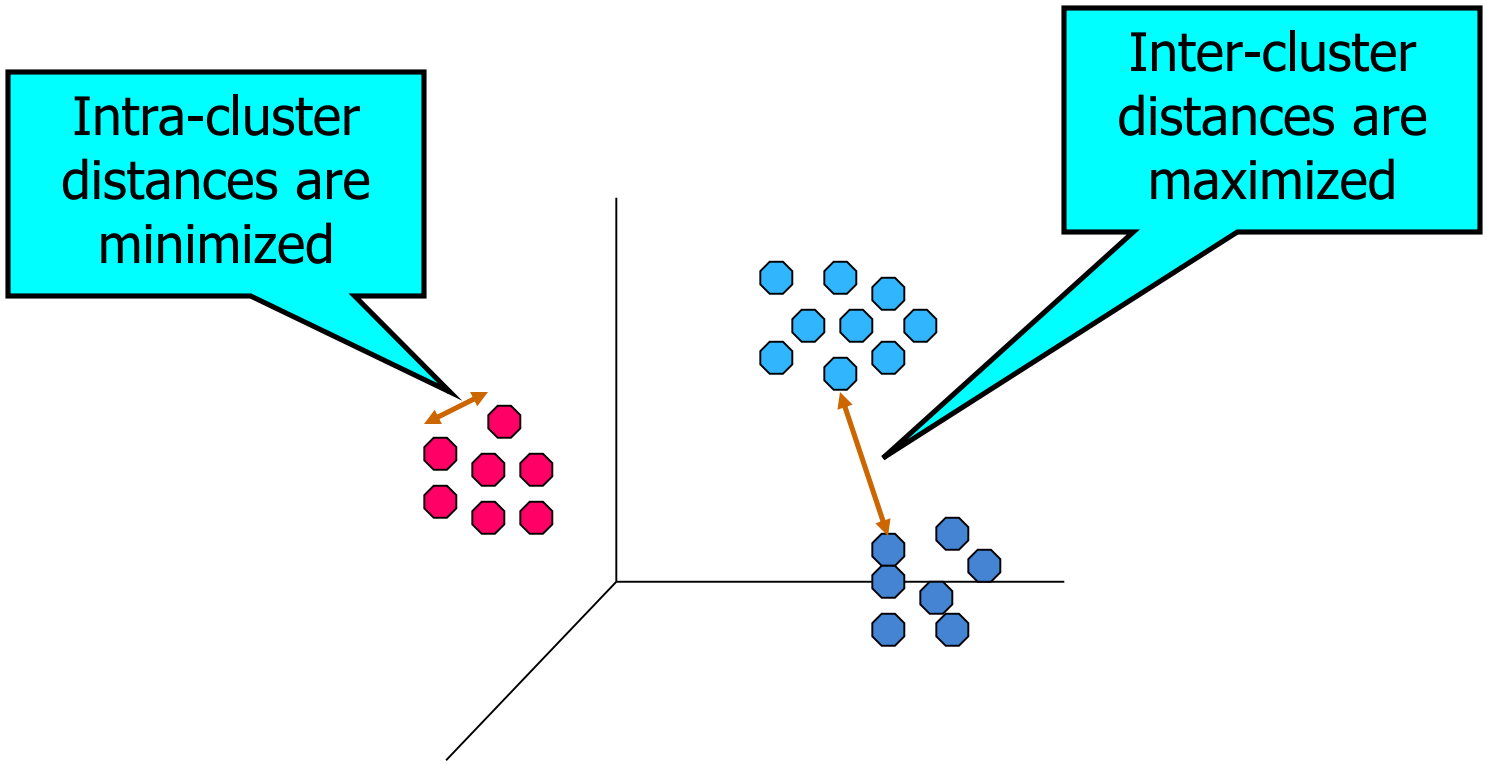
Typical input images

Prediction Techniques

- Classification-Based Approach
  - Nearest Neighbor
  - Neural Networks
  - Bayesian Classifiers
  - Decision Trees
- Sequential Behavior Modeling
  - Hidden Markov Models
  - Temporal Belief Networks

Clustering

- Clustering groups similar data together into clusters.



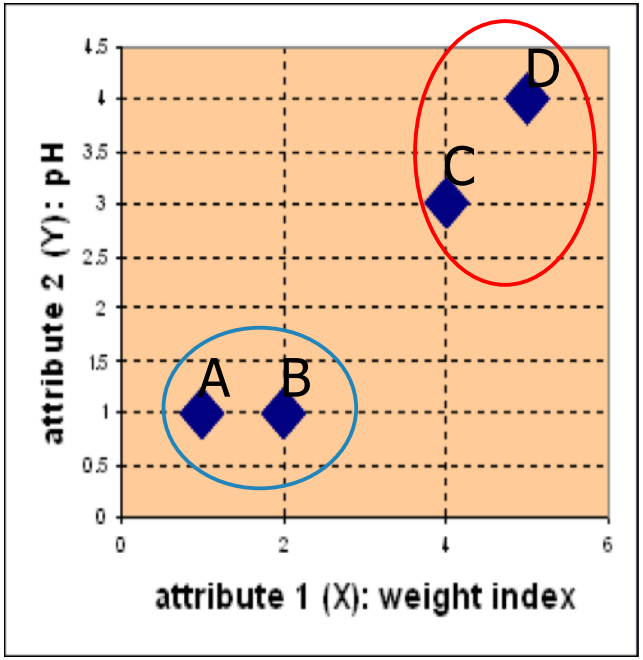
Clustering Algorithms

- K-means clustering *
- Hierarchical clustering
- Density-based clustering (DBSCAN)

K-means Clustering

- we have 4 types of medicines and each has two attributes (pH and weight index). Our goal is to group these objects into *K=2* group of medicine.

Medicine	Weight	pH-Index
A	1	1
B	2	1
C	4	3
D	5	4

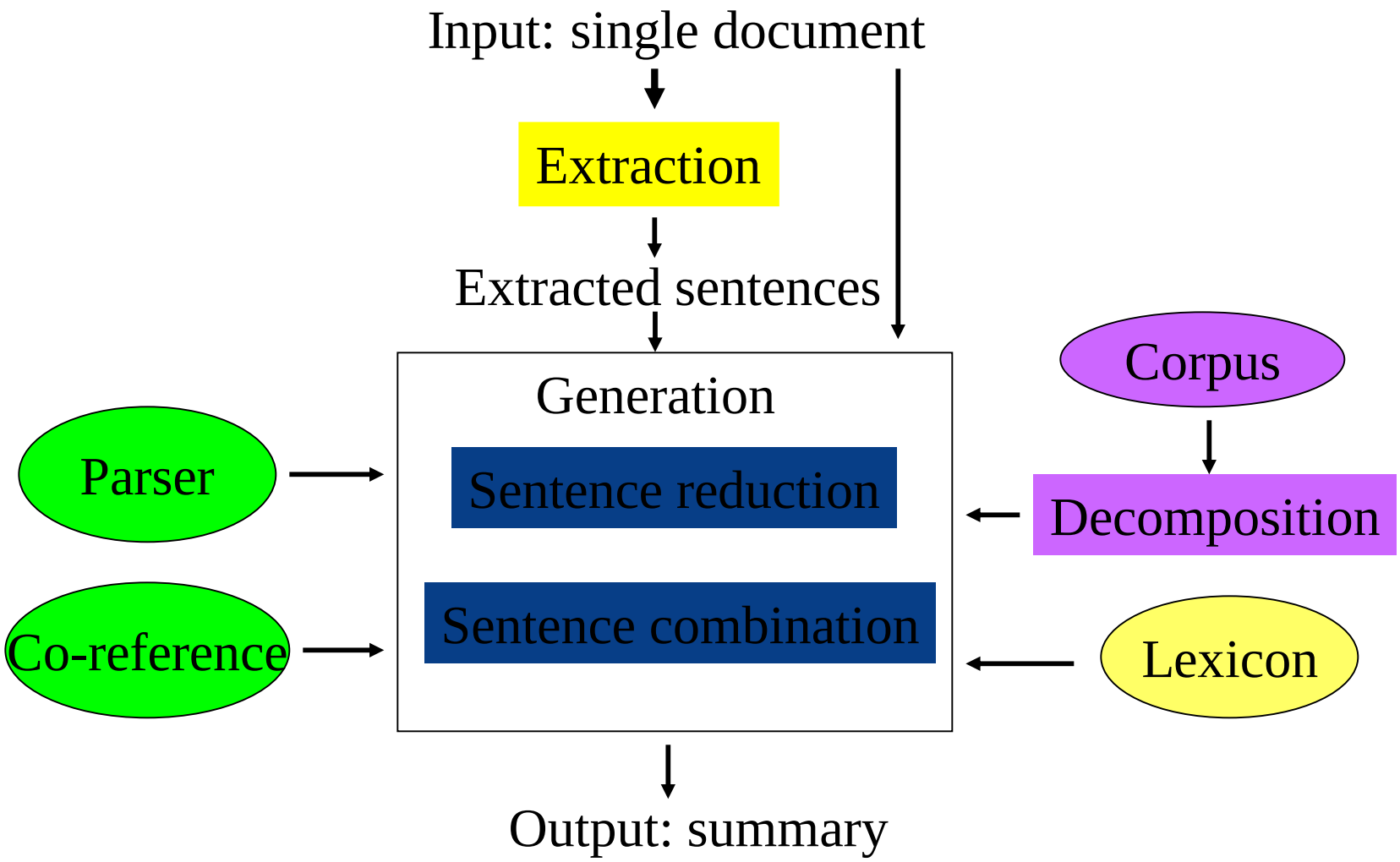


Object	Feature1(X): weight index	Feature 2 (Y): pH	Group (result)
Medicine A	1	1	1
Medicine B	2	1	1
Medicine C	4	3	2
Medicine D	5	4	2

Summarization

- Summaries must convey maximal information in minimal space
  - Data as input (database, software trace, expert system), text summary as output
  - Text as input (one or more articles), paragraph summary as output

Summarization Architecture



CEIT
Think smart, be different.

Association Rules

- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Association Rules

{Diaper} → {Beer},
{Milk, Bread} → {Eggs,Coke},
{Beer, Bread} → {Milk},

Association Rules Algorithm

- Apriori Algorithm
- FP-growth Algorithm
- ECLAT

Sequence Discovery

- A sequence discovery find the ordered elements or events that is given a set of sequences to find the complete set of frequent subsequences.

TID	itemsets
10	a, b, d
20	a, c, d
30	a, d, e
40	b, e, f

SID	sequences
10	<a(abc)(ac)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(ab)(df)cb>
40	<eg(af)cbc>

Sequential Pattern Mining Methods

- Apriori-based Approaches
  - GSP (Generalized Sequential Pattern)
  - SPADE (Sequential Pattern Discovery using Equivalent Class)
- Pattern-Growth-based Approaches
  - FreeSpan
  - PrefixSpan

Dimensionality Reduction

- PCA (Principle Component Analysis)
- SVD (Singular Value Decomposition)
- LSA (Latent Semantic Analysis)
- LDA (Latent Dirichlet Allocation)

Latent Dirichlet Allocation (LDA)

- well-known topic model for text mining
- special feature for finding the mixture of topics in a text corpus
- each **topic** is represented as a probability distribution defined over words
- each **word** is extracted from one of those topics

Topic

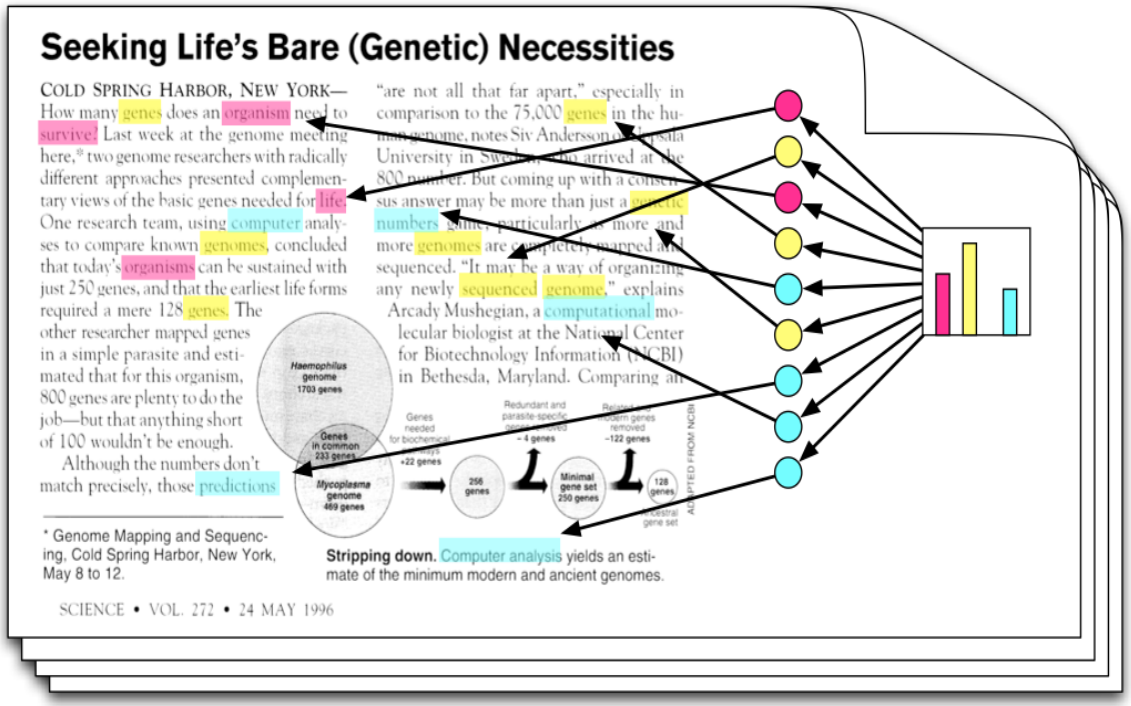
gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Document



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

[Karina Bunyik, 2014]

Next Week Lecture

- Type of Data
- Data Preprocessing

Thank You