

*Tutorial on Statistics, Probability and  
Information Theory for Language Engineers*

*Prof. Ibrahim F. Imam*

Full Professor and Assistant Dean,  
College of Computing and Information Technology  
Arab Academy for Science, Technology & Maritime Transport, Cairo

Email: ifi05@yahoo.com

Phone: 012-2242929

# OUTLINE

1- Supporting Tools	3
2- Basic Concepts	13
3- Documents as Vectors	19
4- Text Mining Applications	32
5- Introduction to Probability	53
6- Preprocessing	60
7- Introduction to Statistics	73
8- Regression	91
9- Testing Measures	96
10- Test of Significance	103
11- The Information Theory	109
12- Association Rules	122
13- Decision Trees	131

# *Tutorial on Text Mining*

## *Part 0*

*Supporting Tools*  
*WordNet & SUMO*

# The WordNet

- WordNet is a semantic network encoding the words of a single (or multiple) language(s) using:
  - Synsets encoding the meanings for each word
  - Relations synonymy, antonymy, hypernymy, hyponymy, holonymy, meronymy, homonymy, troponymy, . . .
  - The English WordNet (v3) encodes 155287 words

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Totals	155287	117659	206941

- WordNet is organized by the concept of synonym sets (synsets), e.g.:
  - musician, instrumentalist, player
  - person, individual, someone

<http://wordnet.princeton.edu/>

# The WordNet Relations

Relation	Definition	Example
Hypernym	From lower to higher concepts	breakfast -> meal
Hyponym	From concepts to subordinates	meal -> lunch
Has-Member	From groups to their members	faculty -> professor
Member-Of	From members to their groups	copilot -> crew
Has-Part	From wholes to parts	table -> leg
Part-Of	From parts to wholes	course -> meal
Antonym	Opposites	leader -> follower

# The WordNet

Word: Cool

## Noun

S: (n) cool (the quality of being at a refreshingly low temperature) "*the cool of early morning*"

S: (n) [aplomb](#), [assuredness](#), cool, [poise](#), [sang-froid](#) (great coolness and composure under strain) "*keep your cool*"

## Verb

S: (v) cool, [chill](#), [cool down](#) (make cool or cooler) "*Chill the food*"

S: (v) cool, [chill](#), [cool down](#) (lose heat) "*The air cooled considerably after the thunderstorm*"

S: (v) cool, [cool off](#), [cool down](#) (lose intensity) "*His enthusiasm cooled considerably*"

## Adjective

S: (adj) cool (neither warm nor very cold; giving relief from heat) "*a cool autumn day*"; "*a cool room*"; "*cool summer dresses*"; "*cool drinks*"; "*a cool breeze*"

S: (adj) cool, [coolheaded](#), [nerveless](#) (marked by calm self-control (especially in trying circumstances); unemotional) "*play it cool*"; "*keep cool*"; "*stayed coolheaded in the crisis*"; "*the most nerveless winner in the history of the tournament*"

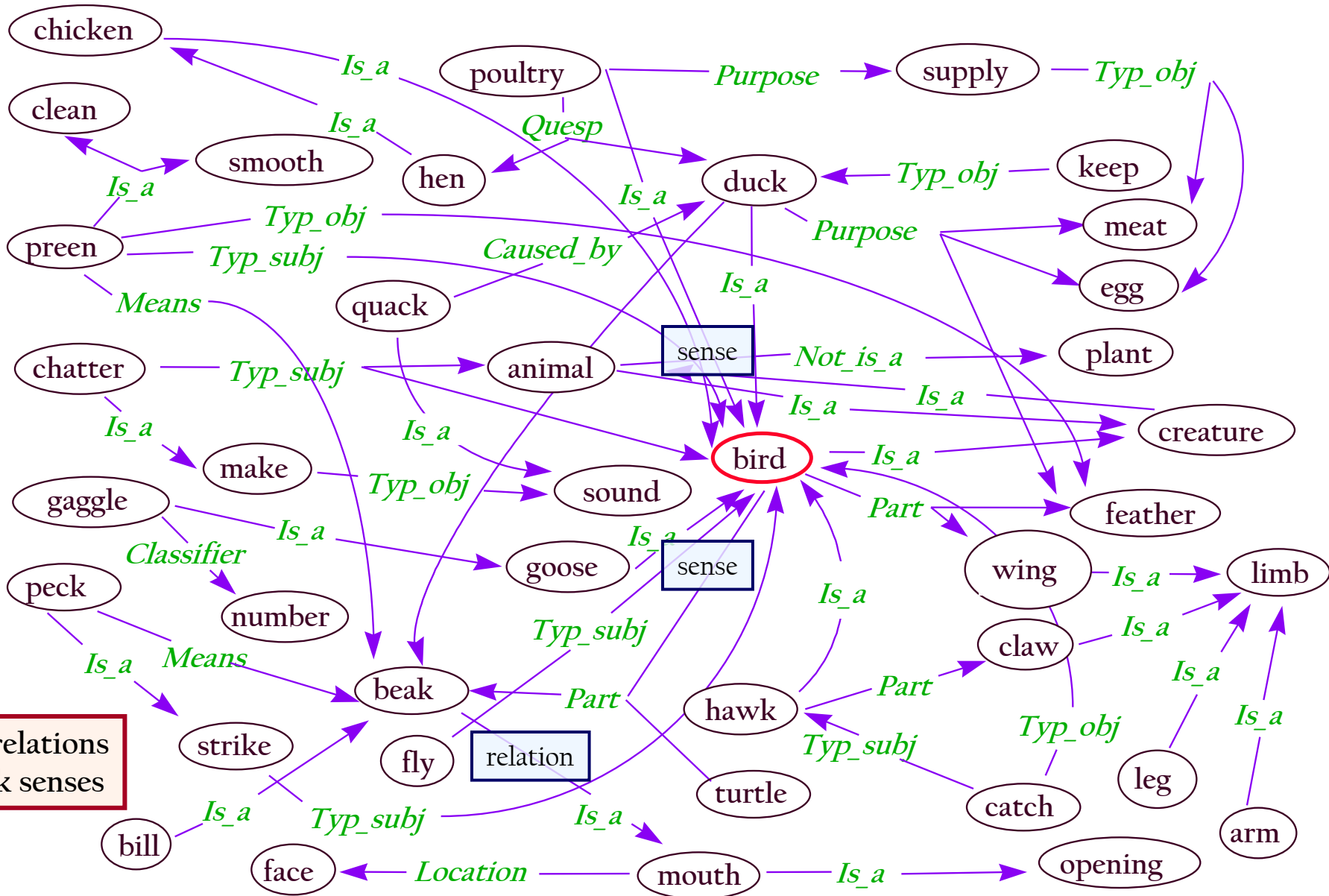
S: (adj) cool ((color) inducing the impression of coolness; used especially of greens and blues and violets) "*cool greens and blues and violets*"

S: (adj) cool (psychologically cool and unenthusiastic; unfriendly or unresponsive or showing dislike) "*relations were cool and polite*"; "*a cool reception*"; "*cool to the idea of higher taxes*"

S: (adj) cool ((used of a number or sum) without exaggeration or qualification) "*a cool million bucks*"

S: (adj) cool (fashionable and attractive at the time; often skilled or socially adept) "*he's a cool dude*"; "*that's cool*"; "*Mary's dress is really cool*"; "*it's not cool to arrive at a party too early*"

# Sample Graph from The WordNet



# *Suggested Upper Merged Ontology (SUMO)*

Suggested S

It is large, open source, and formal

+ Upper U

Focusing on *The most general* and reusable terms and definitions

+ Merged M

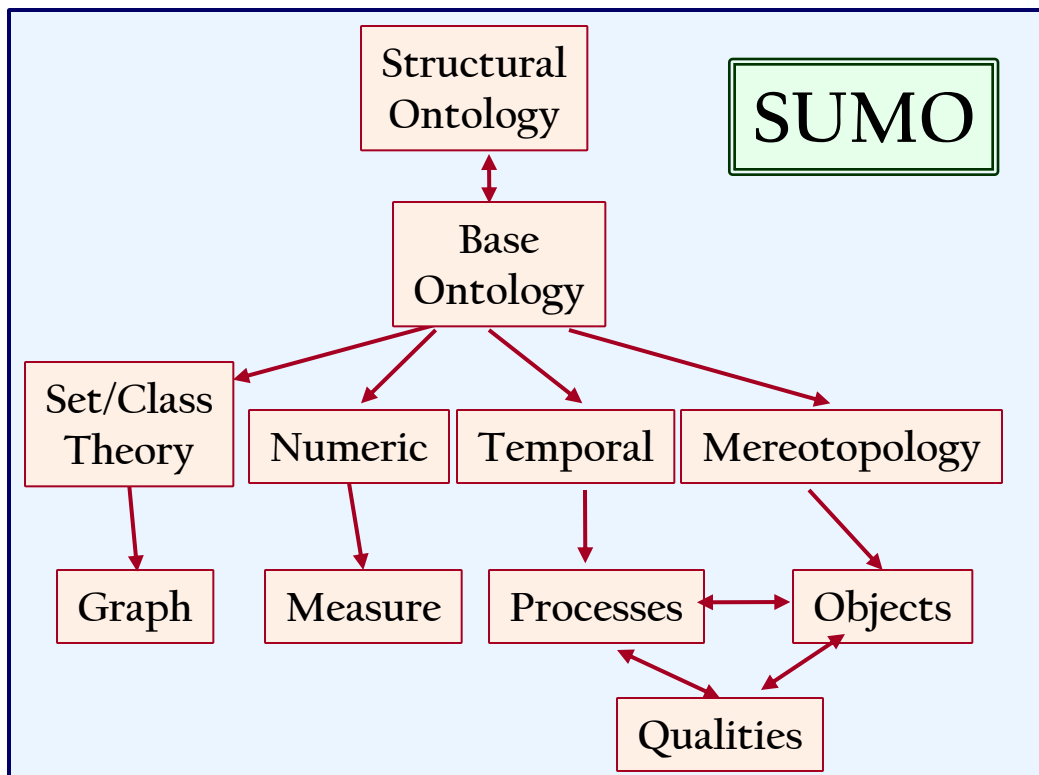
Mapped with large multi-lingual lexicon

+ Ontology O = SUMO

Ontology is a set of term definitions in a formal language describing the world

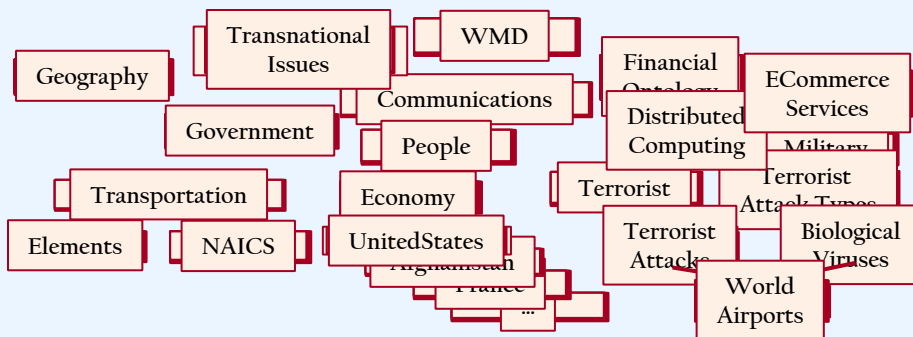


# Suggested Upper Merged Ontology (SUMO)



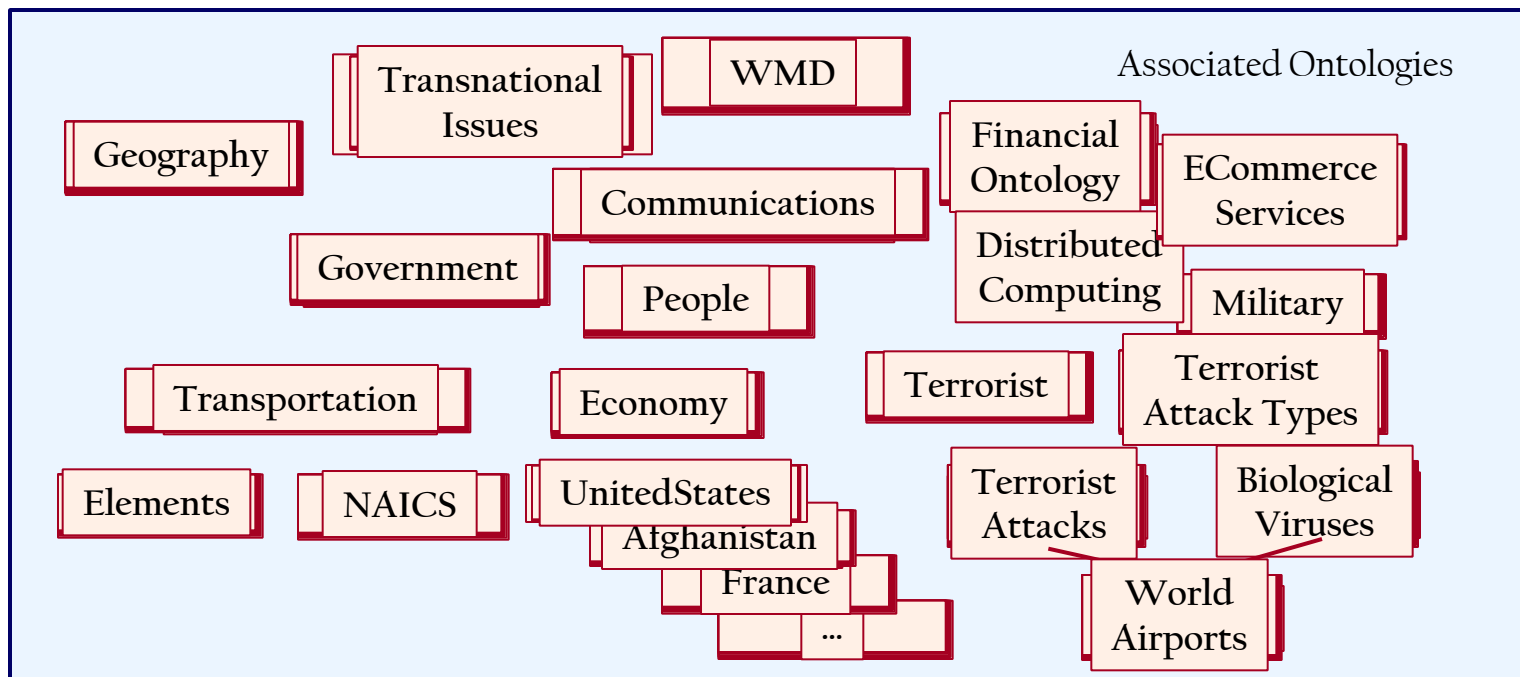
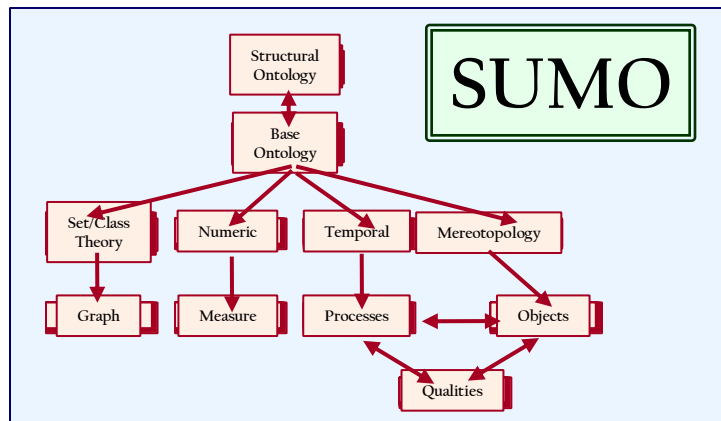
Total Terms = 20399  
Total Axioms = 67108  
Rules = 2500

[www.ontologyportal.org](http://www.ontologyportal.org)



*Associated  
Ontologies*

# Suggested Upper Merged Ontology (SUMO)



# *Suggested Upper Merged Ontology (SUMO)*

## SUMO Search Tool

This tool relates English terms to concepts from the [SUMO](#) ontology by means of mappings to [WordNet](#) synsets.

**English Word:** *According to WordNet, the noun "table" has 6 sense(s).*

[104379243](#) a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs;  
"it was a sturdy table".

SUMO Mappings: [Table](#) (equivalent mapping)

[104379964](#) a piece of furniture with tableware for a meal laid out on it; "I reserved a table at my favorite restaurant".

SUMO Mappings: [Table](#) (subsuming mapping)

[107565259](#) food or meals in general; "she sets a fine table"; "room and board".

SUMO Mappings: [Food](#) (subsuming mapping)

[108266235](#) a set of data arranged in rows and columns; "see table 1".

SUMO Mappings: [ContentBearingObject](#) (subsuming mapping)

[108480135](#) a company of people assembled at a table for a meal or game; "he entertained the whole table with his witty remarks".

SUMO Mappings: [Meeting](#) (subsuming mapping)

[109351905](#) flat tableland with steep edges; "the tribe was relatively safe on the mesa but they had to descend into the valley for water".

SUMO Mappings: [Mesa](#) (equivalent mapping)

# *Suggested Upper Merged Ontology*



Table(table)

[\\_ King Arthur's Round Table](#), [Lord's table](#), [Parsons table](#), [Round Table](#), [altar](#), [board](#), [booth](#), [breakfast table](#), [card table](#), [cocktail table](#), [coffee table](#), [communion table](#), [conference table](#), [console](#), [console table](#), [council board](#), [council table](#), [counter](#), [dining-room table](#), [dining table](#), [dinner table](#), [dresser](#), [dressing table](#), [drop-leaf table](#), [gaming table](#), [gueridon](#), [high table](#), [kitchen table](#), [operating table](#), [pedestal table](#), [pier table](#), [refectory table](#), [stand](#), [table](#), [tea table](#), [toilet table](#), [trestle table](#), [triclinium](#), [vanity](#), [work table](#), [worktable](#)

appearance as argument number 1

([documentation Table EnglishLanguage](#) "A piece of [Furniture](#) with four legs and a flat top. It is used either for eating, paperwork or meetings.")[Mid-level-ontology.kif 1328-1329%3\(externalImage Table](#) "http://upload.wikimedia.org/wikipedia/commons/7/7a/ Table\_and\_chairs.jpg")

# *BASIC MATHEMATICS*

## *Part 1*

### *Basic Concepts*

# BASIC MATHEMATICS

$$\sum_{i=1}^n i = 1 + 2 + \dots + n$$

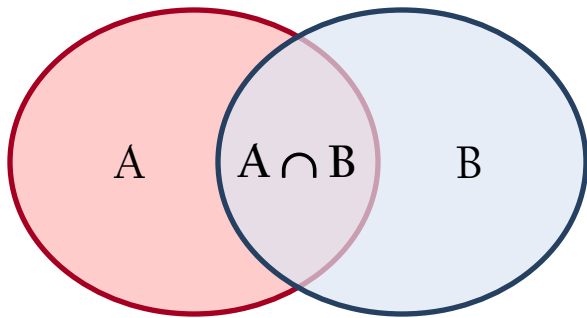
$$\prod_{i=1}^n i = 1 * 2 * \dots * n$$

$$\sum_{i=1}^n ki = k \sum_{i=1}^n i$$

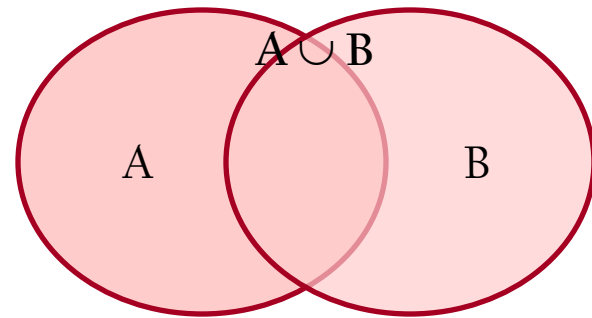
$$\prod_{i=1}^n ki = k \prod_{i=1}^n i$$

# Introduction to Set Theory

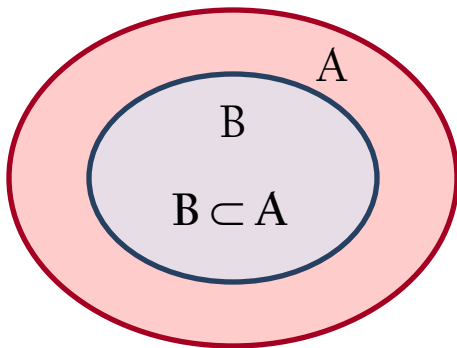
- A set is a collection of distinct items (Example:  $A = \{1, 2, 3, 4, 5\}$ )



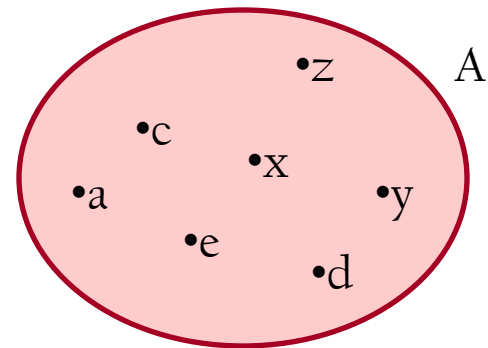
Intersection



Union



Sub-set & Super-set



$x \in A; a \in A; d \in A; \dots$

# Introduction to Set Theory

•  $A = \{a, c, e, d, x, y, z\}$

$$B = \{b, c, d, y, m, n\}$$

$$C = \{c, d\}$$

$$A \cap B = \{c, d, y\}$$

Intersection

$$A \cup B = \{a, b, c, d, e, m, n, x, y, z\}$$

Union

$$A \not\subset B \quad C \subset B \quad C \subset A$$

Sub-set & Super-set

$$x \in A; \quad x \notin B; \quad x \notin C$$

Belong Relationship

$\Phi/\phi$  is the empty set

$$\cap \cup \subset \not\subset \in \notin \neg \wedge \vee$$



# Introduction to Set Theory

- $A \cap (B \cap C) = (A \cap B) \cap C$       &       $A \cup (B \cup C) = (A \cup B) \cup C$
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
- $\neg(\neg A) = A$
- $\neg(A \cap B) = \neg A \cup \neg B$

# Introduction to Propositional Logic

- It is also called the Zero Order Logic
- A sentence  $X$  can be either true or false (1 or 0)

X	Y
0	0
1	1

X	Y	$X \wedge Y$
0	0	0
0	1	0
1	0	0
1	1	1

X	Y	$X \vee Y$
0	0	0
0	1	1
1	0	1
1	1	1

X	Y	$X \rightarrow Y$
0	0	1
0	1	1
1	0	0
1	1	1

X	Y	$X \text{ XOR } Y$
0	0	0
0	1	1
1	0	1
1	1	0

$X \rightarrow Y = \neg X \vee Y$
$\neg(X \wedge Y) = \neg X \vee \neg Y$
$X \wedge X = X \quad \& \quad X \vee X = X$
$X \vee (Y \wedge Z) = (X \vee Y) \wedge (X \vee Z)$
$\neg(\neg X) = X$

# *Introduction to Vectors*

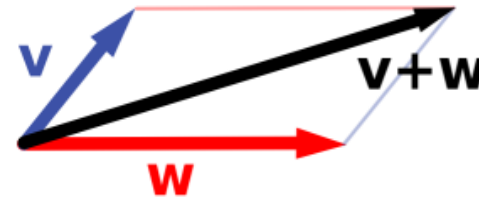
## *Part 2*

### *Representing Documents As Vectors*

# Introduction to Vectors

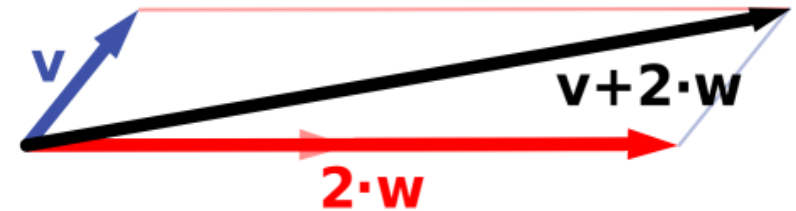
Adding two vectors

$$(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2)$$



Multiplying a vector by a constant and adding it to another vector

$$(x_1, y_1) + (2 \cdot x_2, 2 \cdot y_2) = (x_1 + 2x_2, y_1 + 2y_2)$$

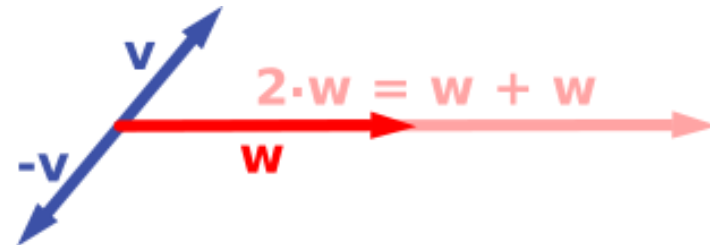


Multiplying a vector by -1

$$-(x_1, y_1) = (-x_1, -y_1)$$

Multiplying a vector by a constant

$$2 \cdot (x_2, y_2) = (2x_2, 2y_2)$$



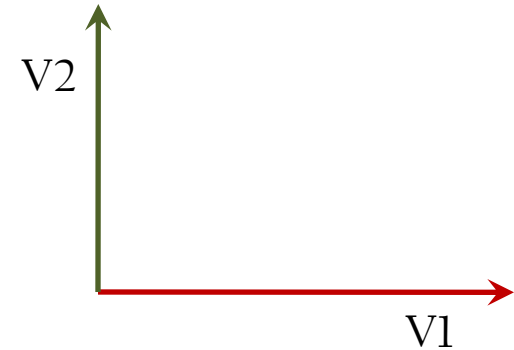
# Introduction to Vectors

Multiplying two orthogonal vectors equal to zero.

Examples:

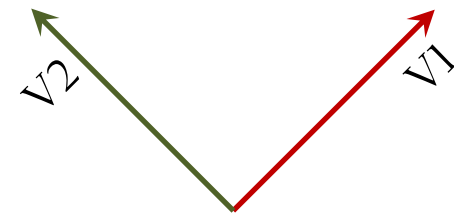
$$V1=(5, 0) \quad \& \quad V2=(0, 4)$$

$$V1 \cdot V2 = 0$$



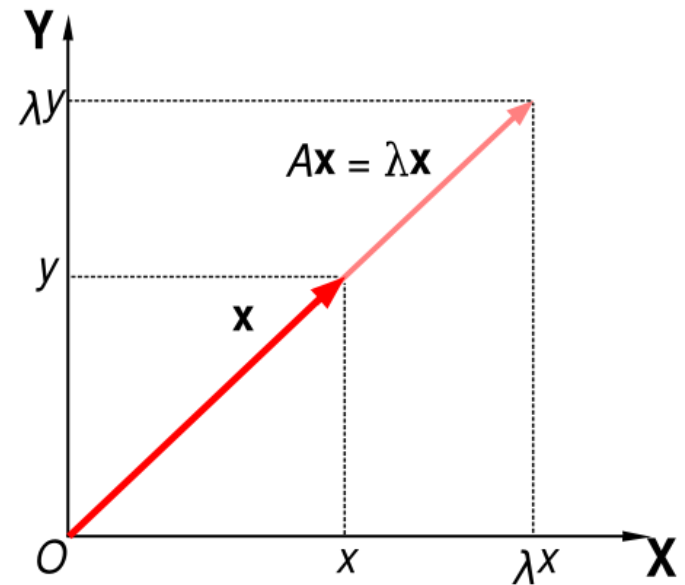
$$V1=(5, 4) \quad \& \quad V2=(-4, 5)$$

$$V1 \cdot V2 = 0$$



# Eigen Values & Eigen Vectors

- An eigenvector of a matrix  $\underline{A}$  is a nonzero vector  $\underline{x}$ , where  $\underline{A}\cdot\underline{x}$  is similar to applying a linear transformation  $\underline{\lambda}$  to  $\underline{x}$  which, may change in length, but not direction
- $\underline{A}$  acts to stretch the vector  $\underline{x}$ , not change its direction, so  $\underline{x}$  is an eigenvector of  $\underline{A}$



$$A\underline{x} - \lambda I\underline{x} = \underline{0}$$

$$(A - \lambda I)\underline{x} = \underline{0}$$

*if there exist an inverse  $(A - \lambda I)^{-1}$ , then  $\underline{x} = \underline{0}$*

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \end{bmatrix}$$

*we need  $\det(A - \lambda I) = 0$  to avoid the trivial solution  $\underline{x} = \underline{0}$*

$$\det(A - \lambda I) = 0$$

# Example on Eigen Values & Eigen Vectors

- Suppose  $A$  is 2x2 matrix

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

$$\det \begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix} = (2-\lambda)^2 - 1 = 0$$

$$\lambda = 1 \quad \text{or} \quad \lambda = 3$$

$$\text{for } \lambda = 3, \quad \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 3 \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\text{for } \lambda = 1, \quad \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 1 \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\begin{bmatrix} 2x + y \\ x + 2y \end{bmatrix} = \begin{bmatrix} 3x \\ 3y \end{bmatrix}$$

$$2x + y = 3x$$

$$x = y$$

$$\begin{bmatrix} 2x + y \\ x + 2y \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix}$$

$$2x + y = x$$

$$x = -y$$

The eigenvectors are:

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

# Representing Documents as Vectors

*Journal* of Artificial *Intelligence* Research

JAIR is a refereed *journal*, covering all areas of Artificial *Intelligence*, which is distributed free of charge over the *internet*. Each *volume* of the *journal* is also published by Morgan Kaufman...

Term Count	Term
0	learning
3	journal
2	intelligence
0	text
0	agent
1	internet
0	webwatcher
0	Perl5
:	:
:	:
:	:
1	volume

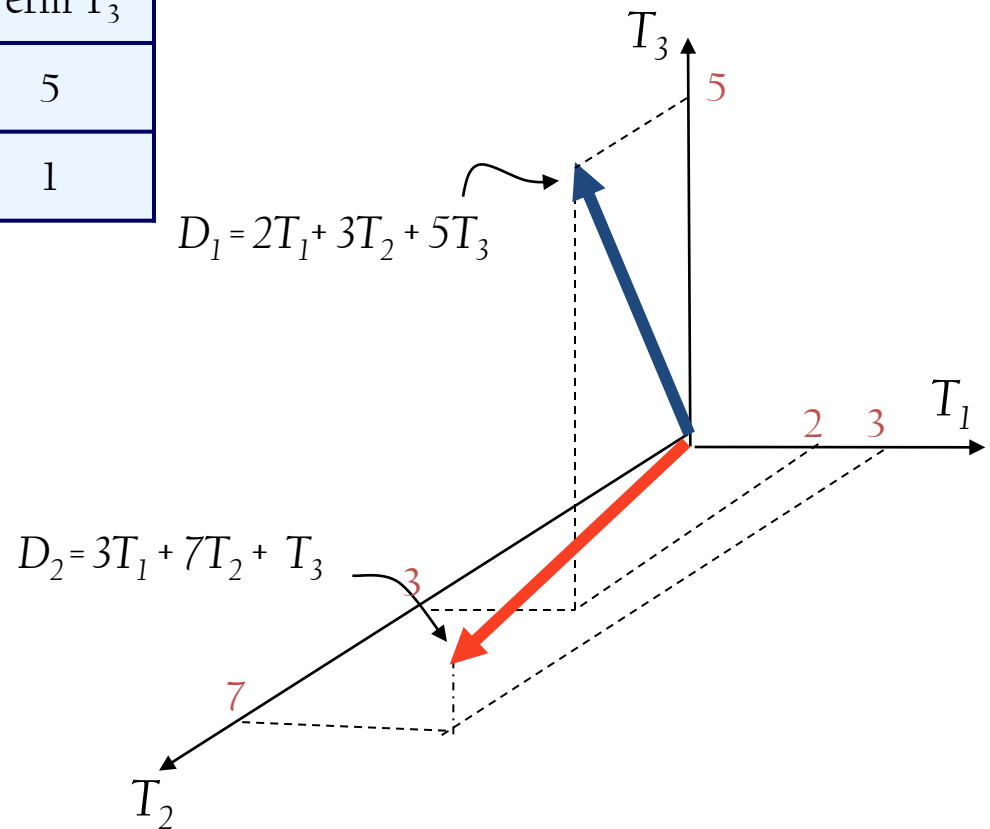


# Documents as Vectors

Suppose we have two documents containing three nouns only

	Term $T_1$	Term $T_2$	Term $T_3$
Document $D_1$	2	3	5
Document $D_2$	3	7	1

$$\begin{array}{c} D_1 \\ \left[ \begin{array}{c} 2 \\ 3 \\ 5 \end{array} \right] \end{array} \quad \left| \quad \begin{array}{c} D_2 \\ \left[ \begin{array}{c} 3 \\ 7 \\ 1 \end{array} \right] \end{array}$$



# *Dimensionality Reduction*

Term Count	Term
34	Home
32	Garden
15	Room
14	Window
11	Furniture
11	Restroom
6	Floor
5	Kitchen
5	Balcony
1	Chimney
1	Street
1	City
1	Dog
1	Lake



*Dimensionality Reduction*

- Term Count
- tfidf
- Chi-Square
- Information Gain
- Gain Ratio

Term Count	Term
15	Room
14	Window
11	Furniture
11	Restroom
6	Floor
5	Kitchen
5	Balcony

## *Term Frequency & Inverse Document Frequency*

Usually a combination of the term frequency and the inverse document frequency

$$TFIDF = w_{ik} = tf_{ik} \times idf_{ik}$$

$$tf_{ik} = 1 + \log_2(tr_{ik}) \quad \text{and zero when } \log = 0$$

$$idf_{ik} = \log_2\left(\frac{N}{n_{ik}}\right) \quad \text{and zero when } \log = 0$$

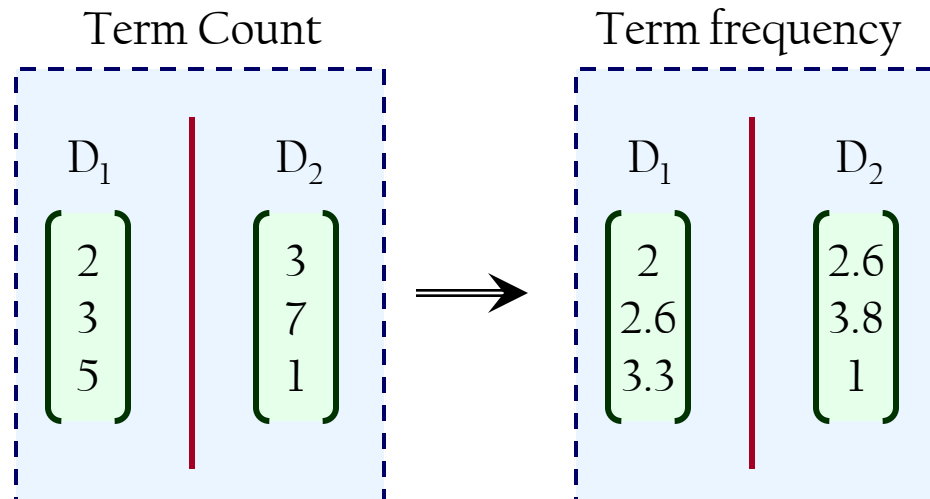
$tf_{ik}$  is the term frequency of term  $i$  in document  $k$ ,  $tr_{ik}$  is the count of term  $i$  in document  $k$ ,  $idf_{ik}$  is the inverse document frequency of term  $i$  in document  $k$ ,  $N$  is the total number of documents in the collection,  $n_{ik}$  is the number of occurrence of term  $i$  in document  $k$ ,  $w_{ik}$  is the weight of term  $i$  in document  $k$ . Logarithm has been used to reduce the difference between the weight of high and low frequency terms. Logarithm of base 2 is used when vectors are full of binary TFIDF weights 0 and 1. Logarithm of base 10 is used when vectors are full of TFIDF weights except binary ones. TFIDF weights values are not normalized.

## Term Frequency & Inverse Document Frequency

$$tf_{ik} = 1 + \log_2(tr_{ik}) \quad \text{and zero when } \log = 0$$

$$idf_{ik} = \log_2\left(\frac{N}{n_{ik}}\right) \quad \text{and zero when } \log = 0$$

$$\log_2 x = \log_{10} x / \log_{10} 2$$



# *The Chi-Square Distribution*

$$\chi^2(t_k, c_i) = \frac{[P(t_k, c_i)P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i)P(\bar{t}_k, c_i)]^2}{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)}$$

$P(t_k, c_i)$  → probability document x contains term t and belongs to category c.

$P(\bar{t}_k, c_i)$  → probability document x does not contain term t and belongs to category c.

$P(t_k, \bar{c}_i)$  → probability document x contains term t and does not belong to category c.

$P(\bar{t}_k, \bar{c}_i)$  → probability document x does not contain term t and does not belong to category c.

$P(t)$  → probability of term t

$P(c)$  → probability of category c

# *The Information Gain*

It measures the classification power of a term

$$IG(t_k, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} P(t, c) \log_2 \frac{P(t, c)}{P(t)P(c)}$$

$P(t_k, c_i)$  → probability document x contains term t and belongs to category c.

$P(\bar{t}_k, c_i)$  → probability document x does not contain term t and belongs to category c.

$P(t_k, \bar{c}_i)$  → probability document x contains term t and does not belong to category c.

$P(\bar{t}_k, \bar{c}_i)$  → probability document x does not contain term t and does not belong to category c.

$P(t)$  → probability of term t.

$P(c)$  → probability of category c.

# The Gain Ratio

$$GR(t_k, c_i) = \frac{\sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} P(t, c) \log_2 \frac{P(t, c)}{P(t)P(c)}}{- \sum_{c \in \{c_i, \bar{c}_i\}} P(c) \log_2 P(c)}$$

$P(t_k, c_i)$  → probability document  $x$  contains term  $t$  and belongs to category  $c$ .

$P(\bar{t}_k, c_i)$  → probability document  $x$  does not contain term  $t$  and belongs to category  $c$ .

$P(t_k, \bar{c}_i)$  → probability document  $x$  contains term  $t$  and does not belong to category  $c$ .

$P(\bar{t}_k, \bar{c}_i)$  → probability document  $x$  does not contain term  $t$  and does not belong to category  $c$ .

$P(t)$  → probability of term  $t$ .

$P(c)$  → probability of category  $c$ .

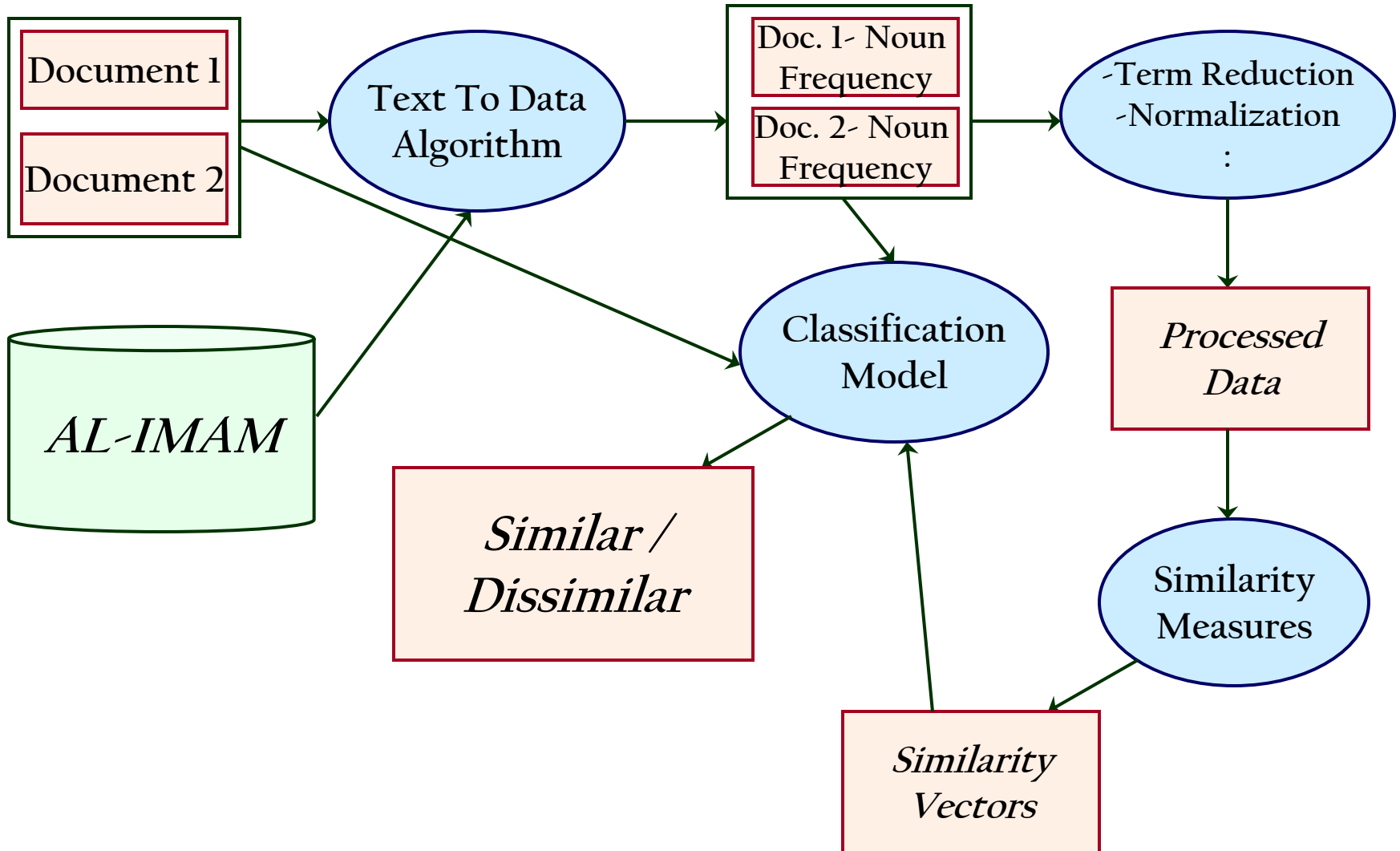
# *Tutorial on Text Mining*

## *Part 3*

### *Text Mining Applications*



# Text Similarity



# Text Similarity

- Each Document is represented by a vector of terms
- Each Term is considered as a dimension in the space
- Terms in the space are uncorrelated so the dimensions are orthogonal on each other
- Each element of the vector has a value (Term Weight)

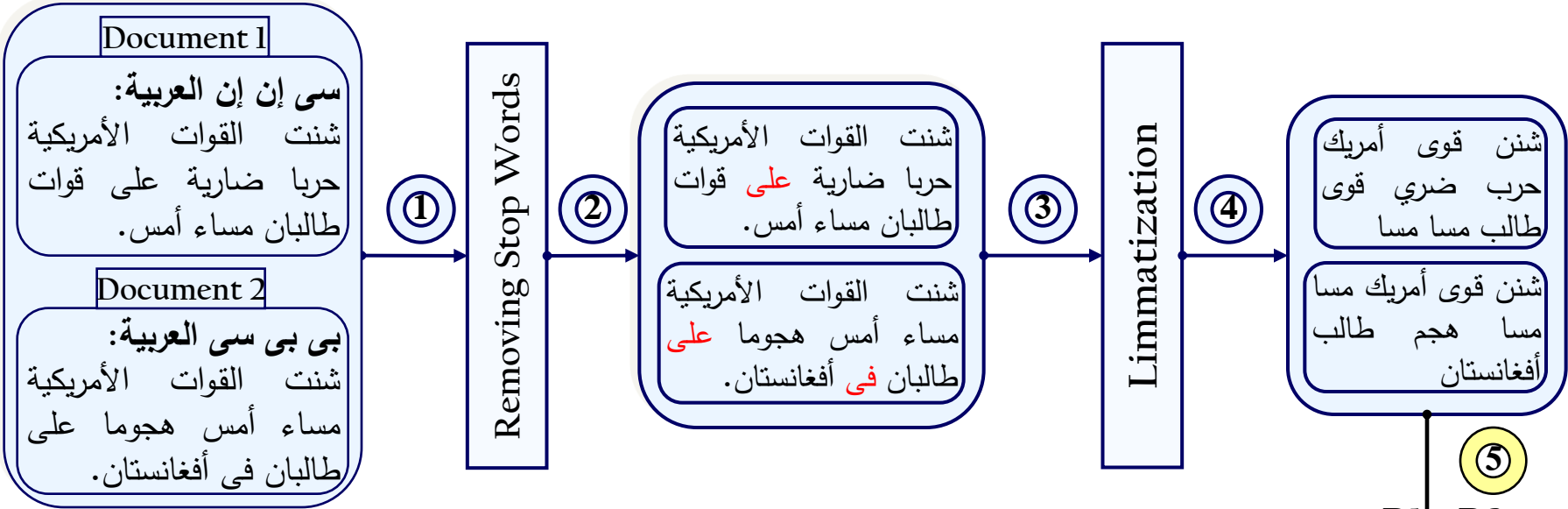
- Document A
  - “A dog and a cat.”

A	Dog	and	Cat	Frog
2	1	1	1	0

- Document B
  - “A frog.”

A	Dog	and	Cat	Frog
1	0	0	0	1

# Text Similarity



Weight indicates term importance either locally or globally

**Similar Documents**  $tf_{ik}$  is the term frequency of term  $i$  in document  $k$ .

**Dissimilar Documents**  $idf_{ik}$  is the inverse document frequency of term  $i$  in the corpus.  $N$  is the total number of documents in the corpus.  $n_i$  is the number of occurrence of term  $i$  in document  $k$ ,  $w_{ik}$  is the weight of term  $i$  in document  $k$ .

**Measuring Text Similarity between Document 1 & 2 vectors using Cosine Criterion**

	D1	D2
شنت	1	1
قوى	1	1
أمريك	1	1
حرب	1	0
ضري	1	0
قوى	1	0
طالب	1	1
مسا	1	1
هجم	0	1
أفغانستان	0	1

$W_{ik} = tf_{ik} \times idf_{ik} = 1 + \log(n_i / n_{ik})$

# Text Similarity

$$\text{Cosine}(D_j, D_k) = \frac{\sum_{i=1}^n w_{ij} \times w_{ik}}{\sqrt{\sum_{i=1}^n w_{ij}^2} \sqrt{\sum_{i=1}^n w_{ik}^2}}$$

$$\text{Euclidean}(D_j, D_k) = \sqrt{\sum_{i=1}^n (w_{i,j} - w_{i,k})^2 / n}$$

$$\text{Dice}(D_j, D_k) = \frac{2 \sum_{i=1}^n w_{i,j} \times w_{i,k}}{\sqrt{\sum_{i=1}^n w_{i,j}^2} + \sqrt{\sum_{i=1}^n w_{i,k}^2}}$$

$$\text{Overlap}(D_j, D_k) = \frac{\sum_{i=1}^n w_{i,j} \times w_{i,k}}{\min(\sqrt{\sum_{i=1}^n w_{i,j}^2}, \sqrt{\sum_{i=1}^n w_{i,k}^2})}$$

$$\text{Jaccard}(D_j, D_k) = \frac{\sum_{i=1}^n w_{i,j} \times w_{i,k}}{\sum_{i=1}^n w_{i,j}^2 + \sum_{i=1}^n w_{i,k}^2 - \sum_{i=1}^n w_{i,j} \times w_{i,k}}$$

Term Weight ( $w_{ik}$ ) =  $tf_{ik} \times idf_{ik}$ ,

Term Frequency ( $tf_{ik}$ ) =  $1 + \log(tr_{ik})$

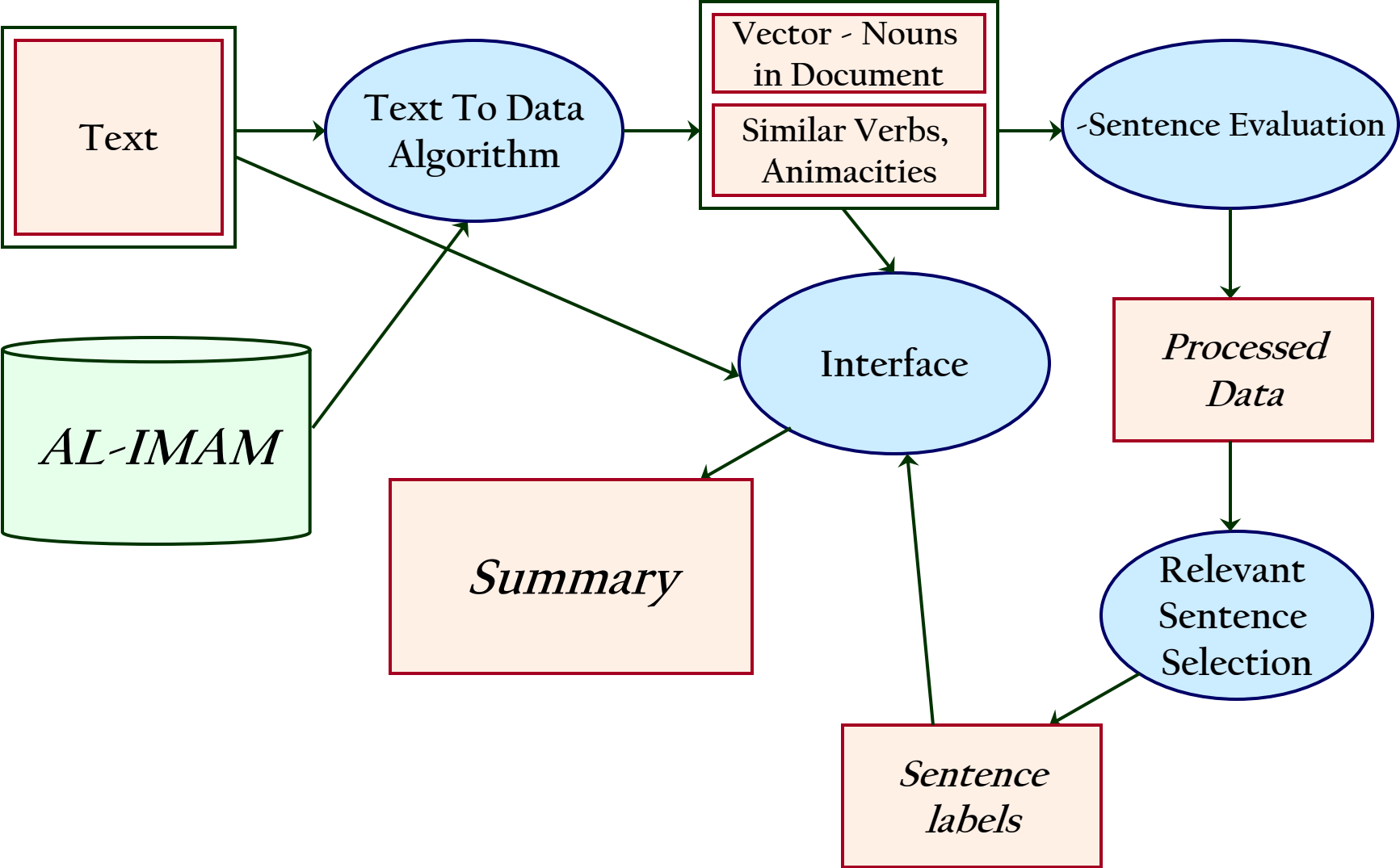
Inverse Doc. Freq. ( $idf_{ik}$ ) =  $\log(N/n_{ik})$

( $tr_{ik}$ ) is the count of term  $i$  in doc.  $k$ .

( $N$ ) is the total # of docs

( $n_{ik}$ ) is the # of occur. of term  $i$  in doc.  $k$

# Text Summarization



# Text Summarization Approaches

## SUMMARIZATION APPROACHES

```
graph TD; A[SUMMARIZATION APPROACHES] --> B[Syntactic-Based]; A --> C[Semantic-Based];
```

### Syntactic-Based

Selecting sentences from the original document according to an evaluation function

### Semantic-Based

Measuring the relevancy of sentences based on their meaning, synonyms, etc.

# Microsoft Word Summarizer

The screenshot displays the Microsoft Word interface with a document titled "TMLA-WWW2003-TutorialProposal.doc". The document content is as follows:

**Tutorial title**  
**Text Mining and Link Analysis for Web Data**

**Presenter contact information including the e-mail address**

**Dunja Mladenic**  
Address: J. Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia  
E-mail: [Dunja.Mladenic@ijs.si](mailto:Dunja.Mladenic@ijs.si)  
Phone: +386 1 4773 377

Marko Grobelnik  
Address: J. Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia  
E-mail: [Marko.Grobelnik@ijs.si](mailto:Marko.Grobelnik@ijs.si)  
Phone: +386 1 4773 778

**Aims/Learning objectives;**

**The aim of this tutorial is to present topics from the areas of text mining and link analysis in the relationship to the web data.** The goal is to show the whole list of nontrivial problems appearing in everyday life and occasionally in professional work with the web and to show how they can be approached using text mining and link analysis techniques and tools. The goal is to make an overview of the available approaches, which are potentially useful for solving interesting problems connected to the documents and their linkage coming from the web structure.

**Duration (half or full day)**  
Half day, but it could be scaled to full day

**Scope (general topic area) and why it is relevant for WWW2004;**

**The tutorial's relevance for the WWW2004 is in the presentation of analytic approaches used on the web data (text+links).** In particular, the tutorial will focus on the possibilities offered by two very active and relevant subfields of data mining: text mining and link analysis. The relevance of these topics to the WWW2004 public is in extending possible activities, which could be used in shaping, understanding and potentially predicting the static and dynamic nature of the web. **Analysis of such data offers typically new insights in the nature of the complex web data.** Suitability of the tutorial for the WWW2004

On the left side, a box labeled "Selected Summary" has arrows pointing to the highlighted text in the document. On the right side, a box labeled "Threshold" has an arrow pointing to the "AutoSummarize" dialog box, which is currently set to 30%.

The Windows taskbar at the bottom shows the "start" button, several open applications, and the system clock displaying "10:30 AM".

# Example of Semantic Summarization

- Summarize the following article in 10 words

HOUSTON – The Hubble Space Telescope got smarter and better able to point at distant astronomical targets on Thursday as spacewalking astronauts replaced two major pieces of the observatory’s gear. On the second spacewalk of the shuttle Discovery’s Hubble repair mission, the astronauts, C. Michael Foale and Claude Nicollier, swapped out the observatory’s central computer and one of its fine guidance sensors, a precision pointing device. The spacewalkers ventured into Discovery’s cargo bay, where Hubble towers almost four stories above, at 2:06 p.m. EST, about 45 minutes earlier than scheduled, to get a jump on their busy day of replacing some of the telescope’s most important components. . . .

*Space News*: [the shuttle Discovery’s Hubble repair mission, the observatory’s central computer]

Taken from: Ren´e Witte, “Introduction to Text Mining”, <http://rene-witte.net>, 2006



# Example of Semantic Summarization

1. Input document is split into sentences
2. Each sentence is deep-parsed
3. Name-entities are disambiguated:
  - Determining that 'George Bush' == 'Bush' == 'U.S. president'
4. Performing Anaphora resolution:
  - Pronouns are connected with named-entities
5. Extracting of **Subject-Predicate-Object** triples
6. Constructing a **graph** from triples
7. Each triple in the graph is described with features for learning
8. Using machine learning train a model for classification of triples into the summary
9. Generate a summary graph from selected triples
10. From the summary graph generate textual summary document

Tom went to town. In a bookstore he bought a large book.

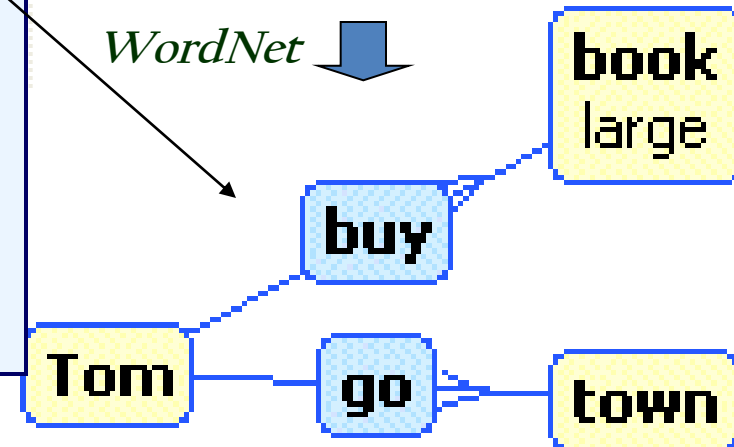
*NLPWin* ↓

**Tom** went to town. In a bookstore he [**Tom**] bought a large book.



Tom ← go → town  
Tom ← buy → book

*WordNet* ↓





# Example of Arabic Summarization

Page 1/5

## انهيار البورصة المصرية. تصحيح أم هبوط.

انهيار أسعار أسهم البورصة المصرية.

للمرة الثانية خلال شهرين يتظاهر مستثمرو البورصة المصرية مطالبين بإقالة رئيس البورصة، وجاءت التظاهرة الثانية على أثر تراجع البورصة والانخفاض بقيمة أغلب الأسهم المتداولة بمنتصف هذا الشهر (مايو 2006) بنسبة 4%؛ وهو ما أعاد للذاكرة ما حدث يوم الثلاثاء الأسود بشهر مارس لنفس العام من هبوط شديد بمؤشر البورصة، والتي حدثت بهينة سوق المال بإيقاف التداول ذلك اليوم.

وتعود طفرة التعاملات خلال 2005 إلى تضخم السيولة بالسوق؛ نتيجة طرح الحكومة أسهم شركتي أموك وسيدبك للجمهور، وتحقيق المشترين لتلك الأسهم لأرباح وصلت إلى حوالي ضعف ثمن الشراء خلال أسابيع قليلة. وفي هذا الجو من توقع تكرار تلك الأرباح العالية من شراء الأسهم الحكومية، طرحت الحكومة نسبة 20% من أسهم الشركة المصرية للاتصالات، وهي الشركة الوحيدة المحكرة لخطوط التليفونات الثابتة وكذلك الاتصالات الدولية؛ وهو ما جعل الجمهور يتكالب عليها. أسباب طفرة 2005

وساهم تضخم الشركة المروجة لأسهم الاتصالات لنسب الإقبال، ومبالغة وسائل الإعلام الرسمية في التوقعات لقيمة السهم بعد طرحه. في حدوث إقبال كبير على شراء أسهم شركة الاتصالات من جانب فئات شعبية تدخل البورصة للمرة الأولى، وليس لديها أي ثقافة استثمارية. ومع تخصيص عدد محدود من الأسهم لطالبي الشراء اتجه هؤلاء الداخلون الجدد لتوجيه فوائض الاكتتاب لشراء أسهم أخرى أو لإعادة شراء أسهم الاتصالات بأسعارها المرتفعة توقعاً لارتفاع أسعارها. وعلى صعيد المستثمرين العرب ساعدت الفوائض البترولية العربية في اتجاه كثيرين منهم للشراء بالبورصة المصرية؛ وهو ما زاد من الطلب خاصة مع انخفاض سعر قيمة الأسهم المصرية النسبي بالنسبة للمستثمرين العرب والأجانب. وزاد دور المضاربين في توجيه السوق -والذي يخلو من وجود صانع سوق يمكنه ترشيد الطفرات السعريّة- وسادت سياسة القطيع في الشراء دون الاستناد إلى المعلومات أو البيانات المالية للشركات أو للتحليل الأساسي أو الفني. حتى زادت أسعار شركات بنسب عالية لا تتناسب بالمرّة مع أدائها، بل إن بعض أسهم شركات الدواجن كانت تتجه للصعود رغم كارثة إنفلونزا الطيور التي شهدتها مصر. **وزاد عدد الأسهم المقيدة بالبورصة بنسبة 41% ليصل إلى 9 316** مليارات سهم. كما زاد رأس المال السوقي للشركات المقيدة بالبورصة بنسبة 95% ليصل إلى 456 مليار جنيه.

وبدأ التصحيح...

وبلغ عدد الشركات المقيدة عام 2005 بالبورصة 744 شركة بنقص 48 شركة عن العام السابق، وهي شركات محدودة التعامل تم شطبها لأسباب تتعلق بنقص شروط القيد، وهو أمر لم يؤثر على السوق التي تتميز بظاهرة تركيز النشاط في نحو 50 شركة فقط. **وارتفع مؤشر أسعار البورصة المصرية (CASE30) بنسبة 146%.**

إلا أن الأسعار لم تأخذ نفس الاتجاه الصعودي بعد أن بلغت مستويات لا تتفق مع واقع الشركات التي تنتمي إليها، ومن هنا فقد كان من الطبيعي أن تصحح السوق نفسها. خاصة مع حدوث نفس التصحيح بالأسواق الخليجية التي كانت قد شهدت طفرة في أسعارها خلال العام الماضي. وتضافر ذلك مع عدم تنسيق هيئة سوق المال نزول عدد من الاكتتابات في زيادة رؤوس أموال الشركات في نفس الوقت؛ وهو ما أدى لزيادة العرض.

وتعود طفرة التعاملات خلال 2005 إلى تضخم السيولة بالسوق؛ نتيجة طرح الحكومة أسهم شركتي أموك وسيدبك للجمهور، وتحقيق المشترين لتلك الأسهم لأرباح وصلت إلى حوالي ضعف ثمن الشراء خلال أسابيع قليلة. **وفي هذا الجو من توقع تكرار تلك الأرباح العالية من شراء الأسهم الحكومية، طرحت الحكومة نسبة 20% من أسهم الشركة المصرية للاتصالات، وهي الشركة الوحيدة المحكرة لخطوط التليفونات الثابتة وكذلك الاتصالات الدولية؛ وهو ما جعل الجمهور يتكالب عليها.**

# Example of Arabic Summarization (Cont.)

Page 2/5

محاولات للإنعاش أسباب طفرة 2005. كما استخدمت الحكومة سلطانها في توجيه محافظ الأوراق المالية الضخمة بالبنوك الحكومية العامة للشراء، ونفس الأمر لبعض صناديق الاستثمار التابعة للبنوك العامة. ومن هنا تماسكت السوق بل اتجهت للارتفاع بعض الوقت. إلا أن قوى السوق كان لا بد لها من أن تؤدي دورها فاستمرت الأسعار في التراجع. حتى إنه مع إعلان وزير الاستثمار -الذي يشرف على السوق من قبل الحكومة- في احتفال كبير عن بدء إطلاق مؤشر داو جونز الخاص بالأسهم المصرية اتجهت الأسعار للتراجع في اليوم التالي مباشرة لإطلاق المؤشر. وحدث نفس الأثر للإعلان عن تكوين محافظ في أسواق دولية تستند محافظها إلى مكونات مؤشر البورصة الذي يضم الشركات الثلاثين الأكثر نشاطا. أسباب الانخفاض

وجاءت انفجارات مدينة ذهب السياحية خلال شهر إبريل 2006 وكان من الطبيعي أن تؤثر على الأسعار بالبورصة. إلا أن الحكومة تدخلت أيضا في إطار سياستها التي تتجه إلى الدعوى بأن أحداث ذهب لم تؤثر على حركة السياحة أو الطيران وبالتالي على البورصة رغم أن واقع الحال الحقيقي غير ذلك. ومع عودة سياسة القمع الحكومية تجاه حركات المجتمع المدني كان من الطبيعي أيضا أن تتأثر البورصة باعتبارها المرأة لكل ما يحدث بالمجتمع من مؤثرات على مناخ الاستثمار. وساهمت عدة عوامل في تراجع ثقة المستثمرين بالسوق. منها تراجع سعر أسهم المصرية للاتصالات لأقل من سعر الطرح الحكومي؛ وهو ما ألحق خسائر كبيرة لحائزيه، خاصة لمن اشتروه بقيم عالية من السوق. كذلك انخفاض سعر سهم هيرميس القابضة كسهم قائد للسوق، وزادت حالة التشاؤم لدى صغار المتعاملين الذين أصبحت لهم النسبة الكبرى من التعامل بعد ابتعاد كثير من المؤسسات المالية عن السوق توقعًا لاستمرار حالة الهبوط السعري حتى شهر أكتوبر القادم. دور بورصات الخليج وبدأ التصحيح.

وذكر هؤلاء أن كثيرا من المستثمرين الخليجيين كانوا مقترضين جانبا من قيمة مشترياتهم من الأسهم، وأنه مع انخفاض الأسعار بأسواقهم طلبتهم البنوك المقرضة لهم بسداد الفرق عن أسعار الأسهم المنخفضة. لذا اتجهوا لتسييل محافظهم في مصر لتدبير سيولة لدفعها لتلك البنوك. **ولقد استمرت كثير من مؤشرات التعامل بالبورصة في النمو مع بداية العام الحالي 2006؛ ففي الثلث الأول من العام زادت قيمة التعامل بنسبة 207% لتصل إلى 119 مليار جنيه مقابل 39 مليار تحققت خلال الثلث الأول من 2005.** وارتفع المتوسط اليومي لقيمة التعامل إلى 1.457 مليار جنيه مقابل 491 مليون جنيه عن نفس الفترة العام الماضي. كما زاد عدد الأوراق المالية المتداولة بنسبة 78% وارتفع عدد الصفقات بنسبة 117%. مع الأخذ في الاعتبار انخفاض مؤشرات التعامل تدريجيا من يناير إلى إبريل.

توقيت حرج: جاء توقيت انهيار البورصة حرجا للحكومة المصرية التي تبنت تماسك الأسعار بالبورصة، والتي تستعد لافتتاح مؤتمر دافوس الشرق الأوسط بمدينة شرم الشيخ بعد 5 أيام من التظاهر في العشرين من مايو. وهو المؤتمر الذي تريد من خلاله الحكومة أن تؤكد ثقة المستثمرين العالميين بها خاصة بعد توالي أحداث العنف تجاه السياحة والشرطة وارتفاع حالة الاحتقان السياسي من جانب قطاعات من القضاة والصحفيين والأطباء ونقابات أخرى وبعض جمعيات حقوق الإنسان. ومن هنا تدخلت الحكومة لتتجه الأسعار للارتفاع بشكل واضح في اليوم التالي للتظاهرة مباشرة. وهذا التدخل الحكومي بسوق الأوراق المالية المصرية يمنع حركتها من التعبير الحقيقي عن آليات السوق، والبورصة الطبيعية تحركها قوى العرض والطلب والمعلومات. حتى تكون مرآة صادقة عن الاقتصاد. ونظرا لأن الاقتصاد المصري يعاني من عجز مزمن بالميزان التجاري، وعجز مزمن بالموازنة العامة، ومن دين عام متزايد، ونسب عالية من البطالة والفقر وحالة من الغلاء، هذا بالإضافة إلى حالة احتقان سياسي غير مسبوق بالمجتمع المصري. فان هذه العوامل لا بد أن تلقي بظلالها على البورصة في الأجل القصير على الأقل، ومهما تدخلت الحكومة فإن قوى السوق لا بد أن تؤدي دورها ويكون لها الكلمة الأخيرة.

## Example of Arabic Summarization (Cont.)

### انهيار البورصة المصرية

انهيار أسعار أسهم البورصة المصرية .

للمرة الثانية خلال شهرين يتظاهر مستثمرو البورصة المصرية مطالبين بإقالة رئيس البورصة، وجاءت التظاهرة الثانية على أثر تراجع البورصة والانخفاض بقيمة أغلب الأسهم المتداولة بمنتصف هذا الشهر (مايو 2006) بنسبة 4%؛ وهو ما أعاد للذاكرة ما حدث يوم الثلاثاء الأسود بشهر مارس لنفس العام من هبوط شديد بمؤشر البورصة، والتي حدثت بهيئة سوق المال بإيقاف التداول ذلك اليوم .

وزاد عدد الأسهم المقيدة بالبورصة بنسبة 41% ليصل إلى  
وارتفع مؤشر أسعار البورصة المصرية (CASE30) بنسبة 146 %.

وفي هذا الجو من توقع تكرار تلك الأرباح العالية من شراء الأسهم الحكومية، طرحت الحكومة نسبة 20% من أسهم الشركة المصرية للاتصالات، وهي الشركة الوحيدة المحكرة لخطوط التليفونات الثابتة وكذلك الاتصالات الدولية؛ وهو ما جعل الجمهور يتكالب عليها .

ولقد استمرت كثير من مؤشرات التعامل بالبورصة في النمو مع بداية العام الحالي 2006؛ ففي الثلث الأول من العام زادت قيمة التعامل بنسبة 207% لتصل إلى 119 مليار جنيه مقابل 39 مليار تحققت خلال الثلث الأول من 2005

After Using Sentence-Base Summarization Algorithm:

Number of Pages in the Summary: ½ out of 5

Number of Paragraphs in the Summary: 7 out of 33

Number of Sentences in the Summary: 7 out of 73

# Example of Arabic Summarization (Cont.)

## بعض الجمل التي تم حذفها لعدم أهميتها

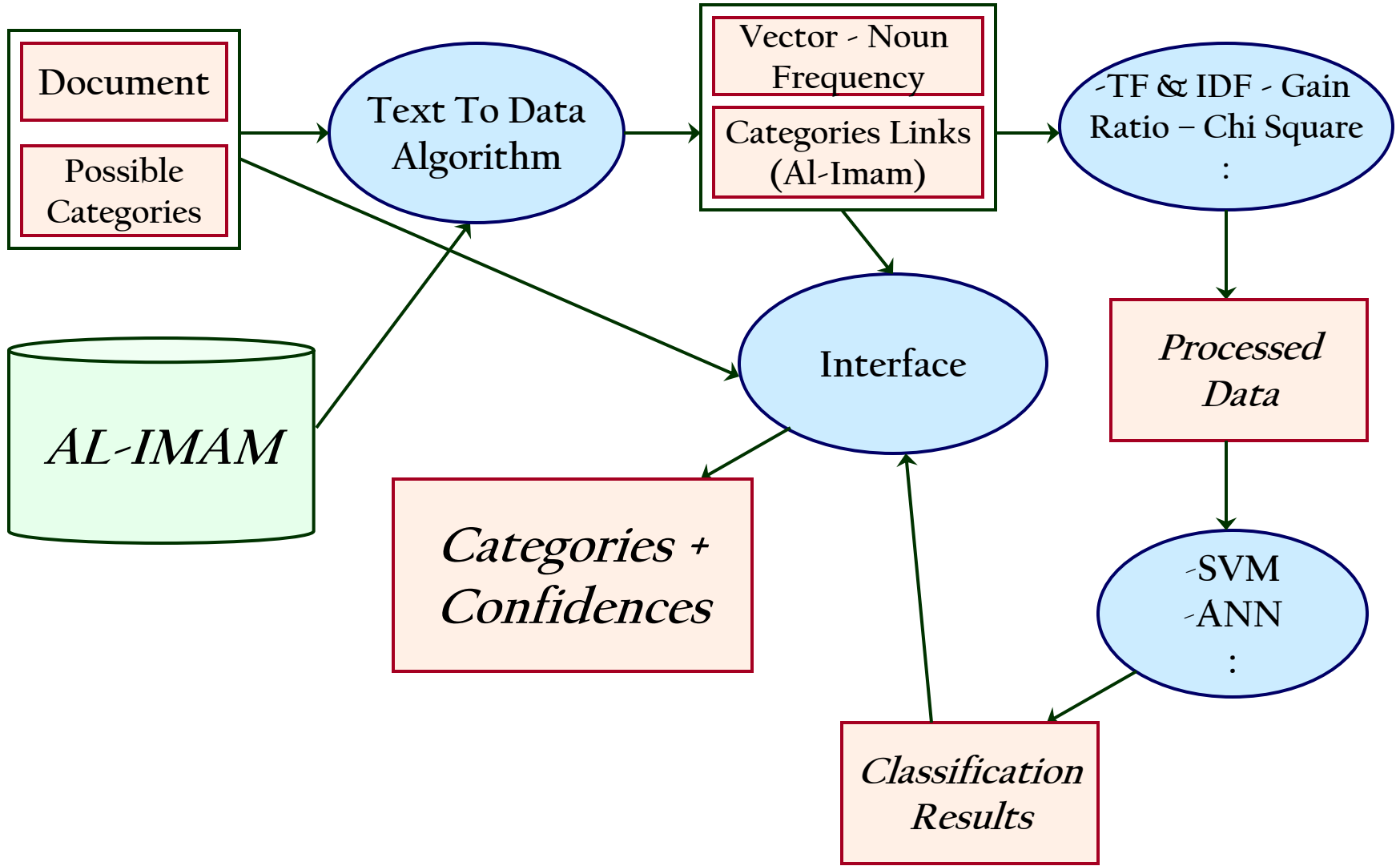
وتعود طفرة التعاملات خلال 2005 إلى تضخم السيولة بالسوق؛ نتيجة طرح الحكومة أسهم شركتي أموك وسيدبك للجمهور، وتحقيق المشتريين لتلك الأسهم لأرباح وصلت إلى حوالي ضعف ثمن الشراء خلال أسابيع قليلة. وفي هذا الجو من توقع تكرار تلك الأرباح العالية من شراء الأسهم الحكومية، طرحت الحكومة نسبة 20% من أسهم الشركة المصرية للاتصالات، وهي الشركة الوحيدة المحتركة لخطوط التليفونات الثابتة وكذلك الاتصالات الدولية؛ وهو ما جعل الجمهور يتكالب عليها.

أسباب طفرة 2005

وساهم تضخم الشركة المروجة لأسهم الاتصالات لنسب الإقبال، ومبالغة وسائل الإعلام الرسمية في التوقعات لقيمة السهم بعد طرحه. في حدوث إقبال كبير على شراء أسهم شركة الاتصالات من جانب فئات شعبية تدخل البورصة للمرة الأولى، وليس لديها أي ثقافة استثمارية. ومع تخصيص عدد محدود من الأسهم لطالبي الشراء اتجه هؤلاء الداخلون الجدد لتوجيه فوائض الاكتتاب لشراء أسهم أخرى أو لإعادة شراء أسهم الاتصالات بأسعارها المرتفعة توقعاً لارتفاع أسعارها.

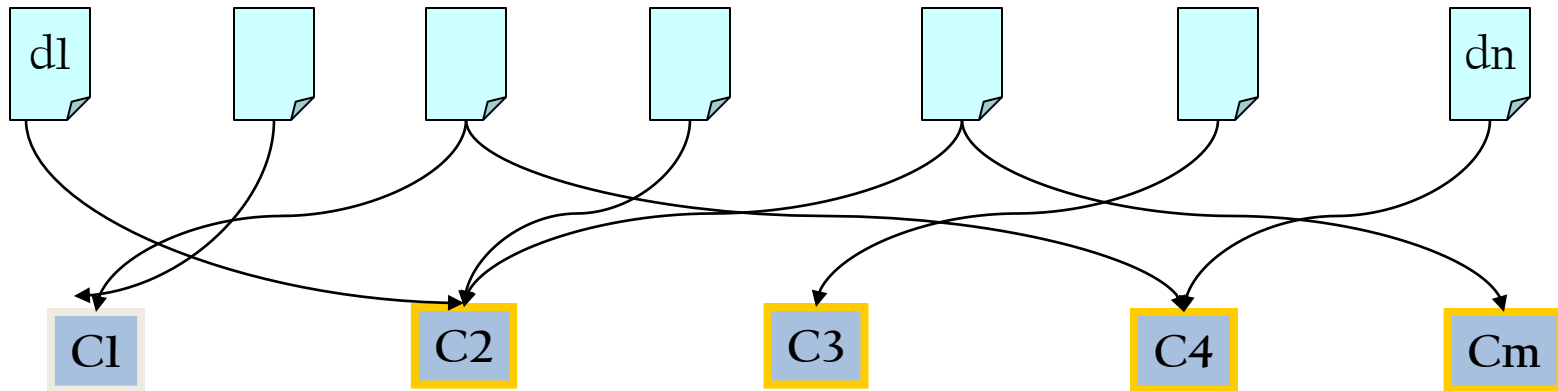
وعلى صعيد المستثمرين العرب ساعدت الفوائض البترولية العربية في اتجاه كثيرين منهم للشراء بالبورصة المصرية؛ وهو ما زاد من الطلب خاصة مع انخفاض سعر قيمة الأسهم المصرية النسبي بالنسبة للمستثمرين العرب والأجانب. وزاد دور المضاربين في توجيه السوق -والذي يخلو من وجود صانع سوق يمكنه ترشيد الطفرات السعرية- وسادت سياسة القطيع في الشراء دون الاستناد إلى المعلومات أو البيانات المالية للشركات أو للتحليل الأساسي أو الفني. حتى زادت أسعار شركات بنسب عالية لا تتناسب بالمرّة مع أدائها، بل إن بعض أسهم شركات الدواجن كانت تتجه للصعود رغم كارثة إنفلونزا الطيور التي شهدتها مصر.

# Supervised Text Categorization



# Supervised Text Categorization

Text Categorization (TC) is the process of labeling electronic text documents with different labels



	$d_1$	...	...	$d_j$	...	...	$d_n$
$c_1$	$a_{11}$	...	...	$a_{1j}$	...	...	$a_{1n}$
...	...	...	...	...	...	...	...
$c_i$	$a_{i1}$	...	...	$a_{ij}$	...	...	$a_{in}$
...	...	...	...	...	...	...	...
$c_m$	$a_{m1}$	...	...	$a_{mj}$	...	...	$a_{mn}$

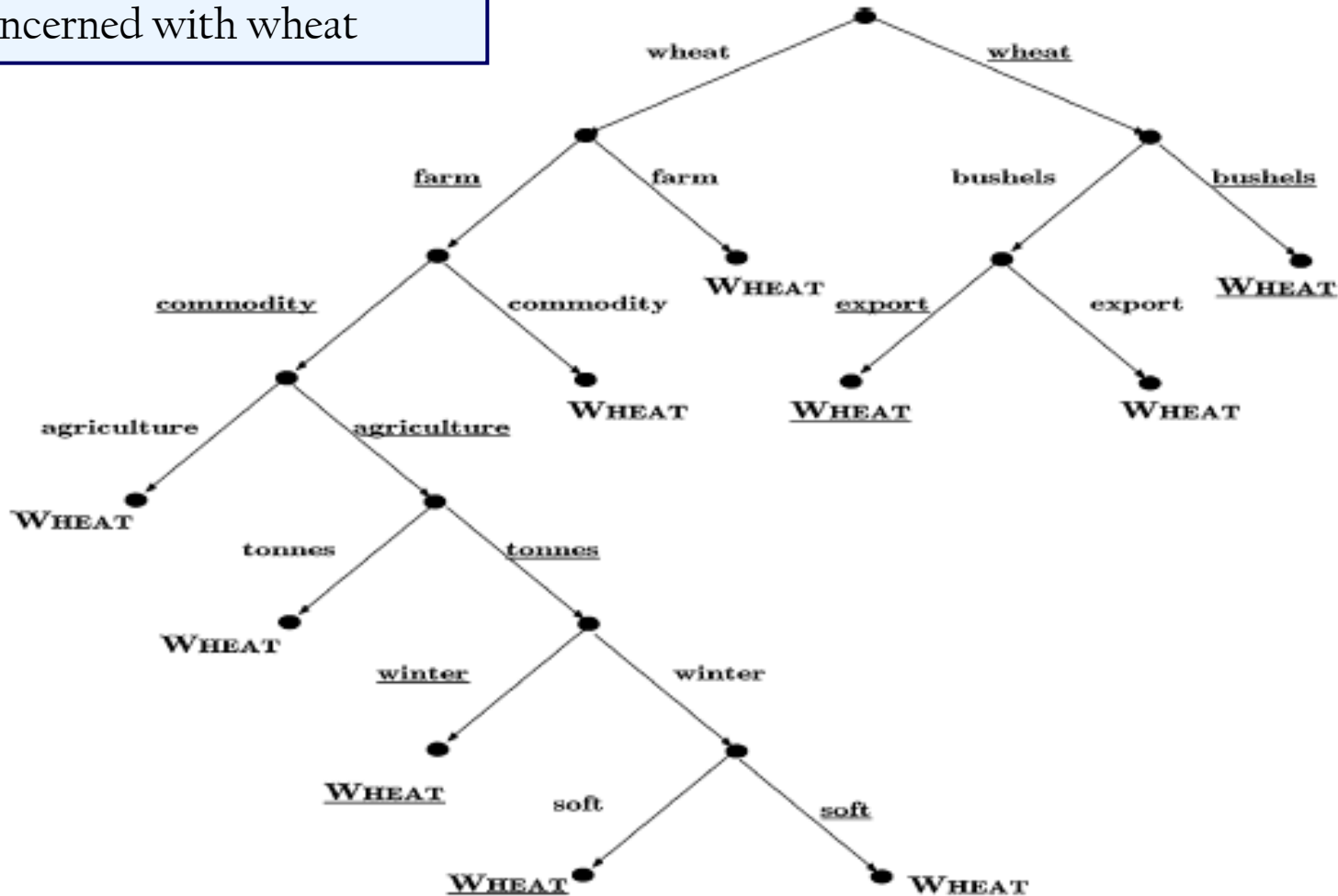


# Supervised Text Categorization

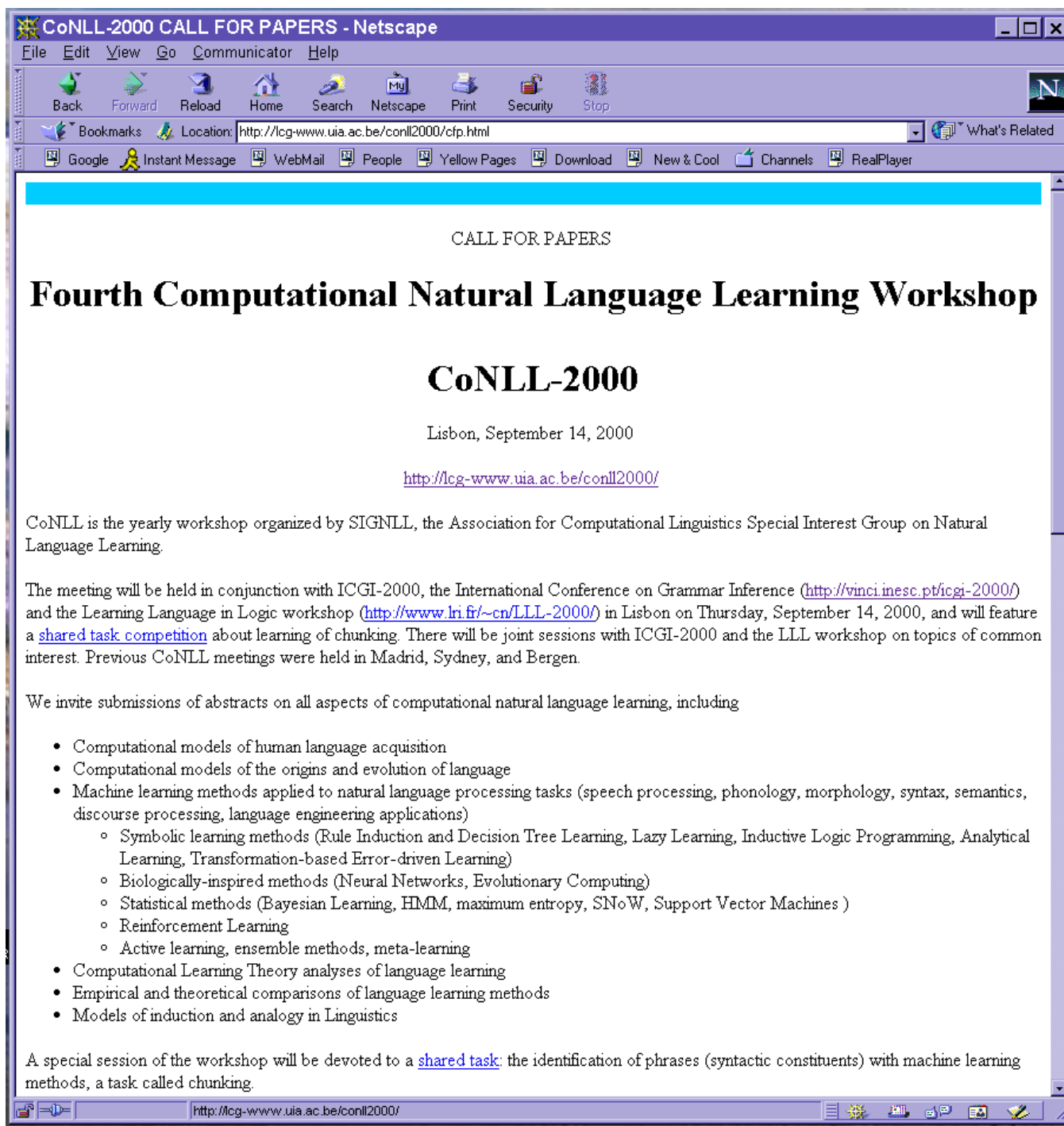
	Supervised	Semi-supervised	Unsupervised
Input Documents	Labeled documents	Labeled and Unlabeled Documents	Unlabeled documents
Method	Machine Learning / Statistical Approaches	Clustering / Machine Learning / Statistical Approaches	Clustering / SOM / Similarity
References	(Deng, Z. 2004) (Sebastiani F., 2003) (Yang, Y. & Pederson, J. 1997)	(Zeng, et al. 2003) (Nigam, et al. 2000)	(Gliozzo, et al. 2005) (Zhao, Y. & Karypis, G. 2005)

# Learning Tree Categorization

Categorizing documents concerned with wheat



Document to categorize:  
CFP for CoNLL-2000



The screenshot shows a Netscape browser window with the title "CoNLL-2000 CALL FOR PAPERS - Netscape". The address bar shows the URL "http://cg-www.uia.ac.be/conll2000/cfp.html". The page content is as follows:

CALL FOR PAPERS

## Fourth Computational Natural Language Learning Workshop

### CoNLL-2000

Lisbon, September 14, 2000

<http://cg-www.uia.ac.be/conll2000/>

CoNLL is the yearly workshop organized by SIGNLL, the Association for Computational Linguistics Special Interest Group on Natural Language Learning.

The meeting will be held in conjunction with ICGI-2000, the International Conference on Grammar Inference (<http://vinci.inesc.pt/icgi-2000/>) and the Learning Language in Logic workshop (<http://www.lri.fr/~cn/LL-2000/>) in Lisbon on Thursday, September 14, 2000, and will feature a [shared task competition](#) about learning of chunking. There will be joint sessions with ICGI-2000 and the LLL workshop on topics of common interest. Previous CoNLL meetings were held in Madrid, Sydney, and Bergen.

We invite submissions of abstracts on all aspects of computational natural language learning, including

- Computational models of human language acquisition
- Computational models of the origins and evolution of language
- Machine learning methods applied to natural language processing tasks (speech processing, phonology, morphology, syntax, semantics, discourse processing, language engineering applications)
  - Symbolic learning methods (Rule Induction and Decision Tree Learning, Lazy Learning, Inductive Logic Programming, Analytical Learning, Transformation-based Error-driven Learning)
  - Biologically-inspired methods (Neural Networks, Evolutionary Computing)
  - Statistical methods (Bayesian Learning, HMM, maximum entropy, SNoW, Support Vector Machines)
  - Reinforcement Learning
  - Active learning, ensemble methods, meta-learning
- Computational Learning Theory analyses of language learning
- Empirical and theoretical comparisons of language learning methods
- Models of induction and analogy in Linguistics

A special session of the workshop will be devoted to a [shared task](#): the identification of phrases (syntactic constituents) with machine learning methods, a task called chunking.

Some  
predicted  
categories

Document Keywords - Netscape

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Stop

Bookmarks Location: <http://alchemist.ijs.si/yqint/yqint.exe> What's Related

Google Instant Message WebMail People Yellow Pages Download New & Cool Channels RealPlayer

### Best Categories

Rank	Prob.	Word [Weight]	Category Path
1.	1.00	LANGUAGE [0.0714]	/Computers_and_Internet/Software/Natural_Language_Processing/
2.	1.00	NATURAL LANGUAGE [0.0429]	/Computers_and_Internet/Internet/World_Wide_Web/Information_and_Documentation/
3.	0.99	PROCESSING [-0.0004]	/Computers_and_Internet/Supercomputing_and_Parallel_Computing/
4.	0.99	GROUP [0.0087]	/Computers_and_Internet/Mobile_Computing/
5.	0.99	SEPTEMBER [0.0089]	/Computers_and_Internet/Software/Programming_Tools/Object_Oriented_Programming/Conferences/
6.	0.99	PROCESSING [0.0041]	/Computers_and_Internet/Information_and_Documentation/Product_Reviews/Buyer_s_Guides/Software/
7.	0.98	GROUP [0.0056]	/Computers_and_Internet/Graphics/
8.	0.98	SEPTEMBER [0.0087]	/Computers_and_Internet/Conventions_and_Conferences/
9.	0.97	GROUP [0.0055]	/Computers_and_Internet/Software/
10.	0.97	LEARNING [0.0022]	/Computers_and_Internet/Internet/Information_and_Documentation/
11.	0.95	SEPTEMBER [0.0084]	/Computers_and_Internet/Communications_and_Networking/Conferences/
12.	0.95	SPECIAL [0.0121]	/Computers_and_Internet/Internet/World_Wide_Web/Conferences/Past_Events/
13.	0.93	PROCESSING [0.0256]	/Computers_and_Internet/Supercomputing_and_Parallel_Computing/Conferences/
14.	0.92	MAXIMUM [0.0019]	/Computers_and_Internet/Hardware/Peripherals/Modems/
15.	0.92	SUBMISSION [0.0857]	/Computers_and_Internet/Internet/World_Wide_Web/Announcement_Services/Robots/

Document: Done

# *PROBABILITY*

## *Part 4*

- Introduction*
- Terminology*

# What Is Probability?

- A priori probability  $P(e)$ : The chance that  $e$  happens
- Conditional probability  $P(f | e)$ : The chance of  $f$  given  $e$
- Joint probability  $P(e, f)$ : The chance of  $e$  and  $f$  both happening; If  $e$  and  $f$  are independent, then  $P(e, f) = P(e) * P(f)$ ; If  $e$  and  $f$  are dependent then  $P(e, f) = P(e) * P(f | e)$

For example, if  $e$  stands for “the first roll of the die comes up 5” and  $f$  stands for “the second roll of the die comes up 3,” then  $P(e, f) = P(e) * P(f) = 1/6 * 1/6 = 1/36$ .

$$\sum_e P(e) = 1$$

$$\sum_e P(e | f) = 1$$

# BASIC Probabilities

$$P(A \cup B) = \begin{cases} P(A) + P(B) & A \& B \text{ are not dependant} \\ P(A) + P(B) - P(A, B) & A \& B \text{ are dependant} \end{cases}$$

- For example, when drawing a single card at random from a regular deck of cards, the chance of getting a heart or a face card (J,Q,K) (or one that is both) is

$$\frac{13}{52} + \frac{12}{52} - \frac{3}{52} = \frac{22}{52}$$

A	$P(A) \in [0, 1]$
not A	$P(A') = 1 - P(A)$
A or B	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ $= P(A) + P(B) \quad \text{if A and B are mutually exclusive}$
A and B	$P(A \cap B) = P(A B)P(B)$ $= P(A)P(B) \quad \text{if A and B are independent}$
A given B	$P(A   B) = \frac{P(A \cap B)}{P(B)}$

# Inference Using Probability

	Toothache		~Toothache	
	Catch	~Catch	Catch	~Catch
Cavity	0.108	0.012	0.072	0.008
~Cavity	0.016	0.064	0.144	0.576

$$P(\text{Cavity} \vee \text{Toothache}) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$$

$$P(\text{Cavity}) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2$$

$$P(\text{Cavity} | \text{Toothache}) = \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6$$

$$P(\sim \text{Cavity} | \text{Toothache}) = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$



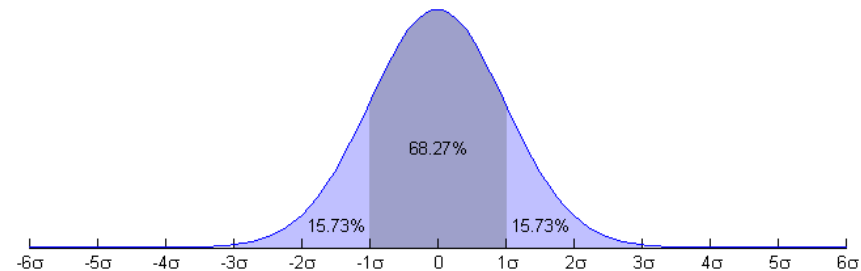
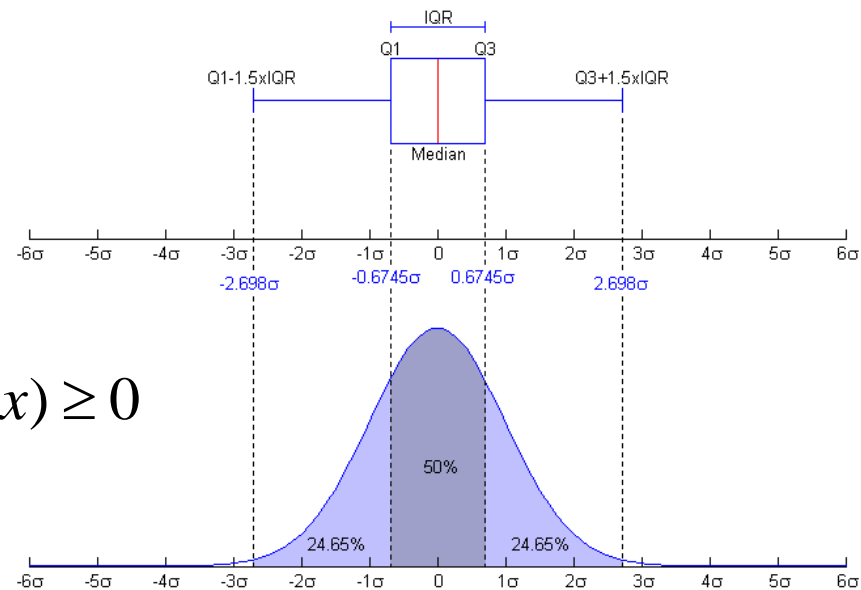
# Probability Density Function PDF

- Probability density function (pdf) is a function that represents a probability distribution in terms of integrals

$$\int_a^b f(x) dx$$

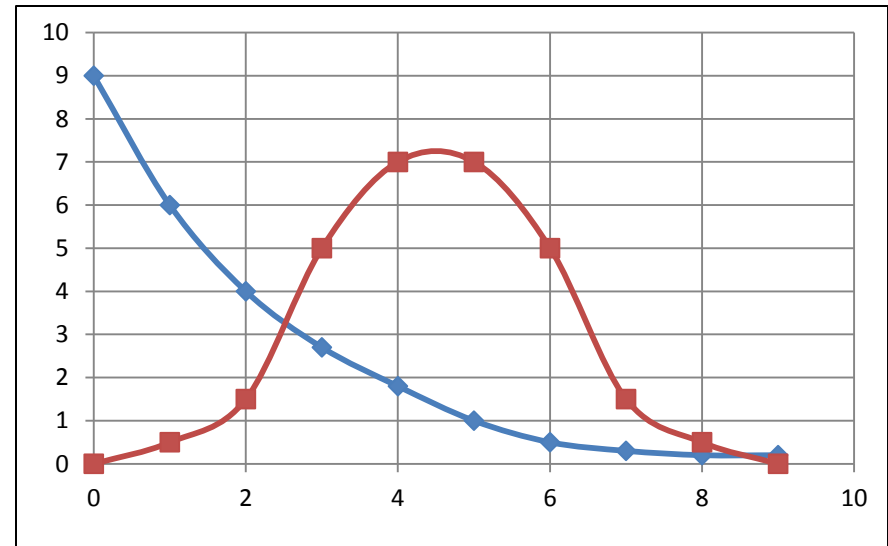
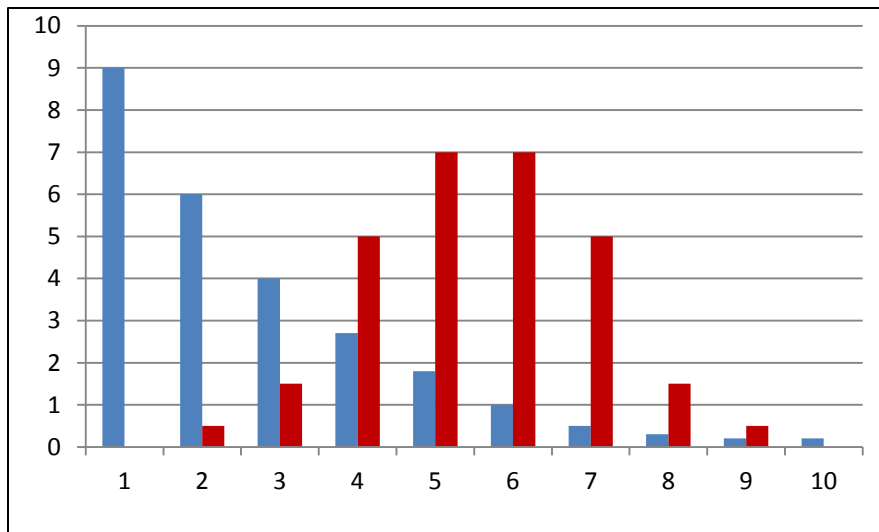
$$\int_{-\infty}^{\infty} f(x) dx = 1$$

$$\& \quad f(x) \geq 0$$



# Probability Density Function PDF

- The Summation is used with Discrete Data



# Conditional & Bayesian Probability

- Conditional probability is the probability of some event  $A$ , given the occurrence of some other event  $B$ ; it is written  $P(A|B)$ , and is read “the probability of  $A$ , given  $B$ ”

$$P(A | B) = \frac{P(A, B)}{P(B)}$$

- Bayesian probability, the probability of a hypothesis given the data (the *posterior*), is proportional to the product of the likelihood times the prior probability (often just called the *prior*)
- The likelihood brings in the effect of the data, while the prior specifies the belief in the hypothesis before the data was observed

$$P(A | B) = \frac{P(A)P(B | A)}{P(B)}$$

- If two variables  $A$  and  $B$  are independent

$$P(A \wedge B | C) = P(A | C)P(B | C)$$

# *Text Mining*

## *Part 5*

### *Preprocessing*

# Text Preprocessing

- Remove “fluff” if exists (e.g., ads, navigation bars, pictures, etc.)
- Convert to plain text (i.e., from PDF, DOC, or other formats)
- Check words correctness (in case of erroneous text or using OCR)
- Handle tables, numbers, and equations

- حذف التشكيل ( َ ِ ُ ً )
- حذف الرموز الخاصة ( / & # @ \$ \* % )
- حذف الأرقام
- حذف الزوائد في بداية الكلمة وآخرها ( است ها )
- تحويل همزات القطع إلى همزات وصل ( أحمد احمد )
- تحويل الألف اللينة إلى ألف العالية
- حذف الكلمات الزائدة Stop words

# Preprocessing: Sentence Splitter

## Sentence Splitting

- Sentences end with “.”, “!”, or “?”
- Difficult when a “.” do not indicate an EOS: “MR. X”, “3.14”, “Y Corp.”, etc.
- We can detect common abbreviations (“U.S.”), but what if a sentence ends with one? “. . .announced today by the U.S. The ...

توجد نفس المشاكل في اللغة العربية:

- “وقدم أ.د. إبراهيم إمام درس عن ...”
- الجمل في اللغة العربية تتداخل بصورة أكثر تعقيدا

Google n-gram corpus Statistics: <http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html#links> Size = 24 GB

Number of tokens:	1,024,908,267,229
Number of sentences:	95,119,665,584
Number of unigrams:	13,588,391
Number of bigrams:	314,843,401
Number of trigrams:	977,069,902
Number of fourgrams:	1,313,818,354
Number of fivegrams:	1,176,470,663

# Samples of Google n-gram Data

## 3-gram samples

Freq.

ceramics collectables collectibles	55
ceramics collectables fine	130
ceramics collected by	52
ceramics collectible pottery	50
ceramics collectibles cooking	45
ceramics collection ,	144
ceramics collection .	247
ceramics collection </S>	120
ceramics collection and	43
ceramics collection at	52
ceramics collection is	68
ceramics collection of	76
ceramics collection	59
ceramics collections ,	66
ceramics collections .	60
ceramics combined with	46
ceramics come from	69
ceramics comes from	660
ceramics community ,	109
ceramics community .	212
ceramics community for	61
ceramics companies .	53
ceramics companies consultants	173
ceramics company !	4432
ceramics company ,	133
ceramics company .	92

## 4-gram samples

Freq.

serve as the incoming	92
serve as the incubator	99
serve as the independent	794
serve as the index	223
serve as the indication	72
serve as the indicator	120
serve as the indicators	45
serve as the indispensable	111
serve as the indispensable	40
serve as the individual	234
serve as the industrial	52
serve as the industry	607
serve as the info	42
serve as the informal	102
serve as the information	838
serve as the informational	41
serve as the infrastructure	500
serve as the initial	5331
serve as the initiating	125
serve as the initiation	63
serve as the initiator	81
serve as the injector	56
serve as the inlet	41
serve as the inner	87
serve as the input	1323
serve as the inputs	189

# Preprocessing: Word Tokenizers

**Tokenization is difficult.** For example,

“John’s sick” shall we split “John’s” into one token or two?

If one ! problems in parsing (where’s the verb?)

If two ! what do we do with John’s house?

**Heavy Compounding** في اللغة العربية توجد مشاكل أكثر تعقيدا من ذلك  
مثلا:

• جملة “يلعبونها في الملاعب” عند حذف السوابق واللواحق يتبقى “لعب” وتم حذف  
الفاعل “هم” والمفعول به “هي”

أيضا إذا كان الكلام يحتوي على تركيبية كيميائية، أو هياكل خاصة بالعلوم:

1,4--xylanase II from *Trichoderma reesei*

When N-formyl-L-methionyl-L-leucyl-L-phenylalanine (fMLP) was injected. . .

Technetium-99m-CDO-MeB [Bis[1,2-cyclohexanedionedioximato(1-)-O]-[1,2-cyclohexanedione dioximato(2-)-O]methyl-borato(2-)-

N,N0,N00,N000,N0000,N00000)-

chlorotechnetium) belongs to a family of compounds. . .



# Preprocessing: Morphological Analyzers

## Morphological Analyzer

- Reflects changes in case, gender, number, tense, etc.  
give → gives, gave, given
- **Stemming** reduce words to a base form
- **Lemmatization** reduce words to their lemma (root)

التحليل الصرفي لكلمة: الفِلاحة									
الكلمة	النوع	السوابق	اللواحق	الساق	الجذر	الوزن	الجنس	معرف	إنساني
الفِلاحة	مصدر	ال	ة	فلاح	فلح	فعال	مؤنث	✓	✓

## Advantages of Using the Stem as a Word Representative:

- Simple and Fast

## Disadvantages of Using the Stem as a Word Representative:

- Can create words that do not exist in the language, e.g., computers → comput
- Often reduces different words to the same stem, e.g., army, arm → arm;  
stocks, stockings → stock

# Preprocessing: Morphological Analyzers (Cont.)

## **Advantages of Using the Root as a Word Representative:**

- The root is an actual word
- Usually provide better accuracy than the stem

## **Disadvantages of Using the Root as a Word Representative:**

- Significantly complex
- Requires language dependent resources

Get a copy of Porter stemmer (For English) at:

<http://www.tartarus.org/~martin/PorterStemmer/>

# Preprocessing: Part of Speech Tagging (POS)

- A Tagger algorithm assigns a tag for each word statistically
- calculated based on different word order probabilities

part of speech	function or "job"	example words	example sentences
<u>Verb</u>	action or state	(to) be, have, do, like, work, sing, can, must	EnglishClub.com <b>is</b> a web site. I <b>like</b> EnglishClub.com.
<u>Noun</u>	thing or person	pen, dog, work, music, town, London, teacher, John	This is my <b>dog</b> . He lives in my <b>house</b> . We live in <b>London</b> .
<u>Adjective</u>	describes a noun	a/an, the, 69, some, good, big, red, well, interesting	My dog is <b>big</b> . I like <b>big</b> dogs.
<u>Adverb</u>	describes a verb, adjective or adverb	quickly, silently, well, badly, very, really	My dog eats <b>quickly</b> . When he is <b>very</b> hungry, he eats <b>really</b> quickly.
<u>Pronoun</u>	replaces a noun	I, you, he, she, some	Tara is Indian. <b>She</b> is beautiful.
<u>Preposition</u>	links a noun to another word	to, at, after, on, but	We went <b>to</b> school <b>on</b> Monday.
<u>Conjunction</u>	joins clauses or sentences or words	and, but, when	I like dogs <b>and</b> I like cats. I like cats <b>and</b> dogs. I like dogs <b>but</b> I don't like cats.
<u>Interjection</u>	short exclamation, sometimes inserted into a sentence	oh!, ouch!, hi!, well	<b>Ouch!</b> That hurts! <b>Hi!</b> How are you? <b>Well</b> , I don't know.

# Preprocessing: Part of Speech Tagging (POS)

Verb
work!

Noun	Verb
John	works.

Pronoun	Verb	Noun
He	loves	cats.

Noun	Verb	Verb
John	is	working.

Noun	Verb	Noun	Adverb
Ahmed	speaks	French	well.

Noun	Verb	Adjective	Noun
cats	like	nice	children.

Pronoun	Verb	Preposition	Adjective	Noun	Adverb
She	ran	to	the	station	quickly.

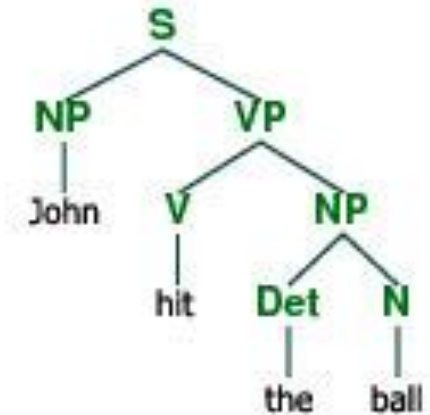
Pronoun	Verb	Adjective	Noun	Conjunction	Pronoun	Verb	Pronoun
She	likes	big	snakes	but	I	hate	them.

Interjection	Pronoun	Conjunction	Adjective	Noun	Verb	Prep.	Noun	Adverb
Well,	she	and	young	John	walk	to	school	Slowly.

# Preprocessing: Syntactic Analysis

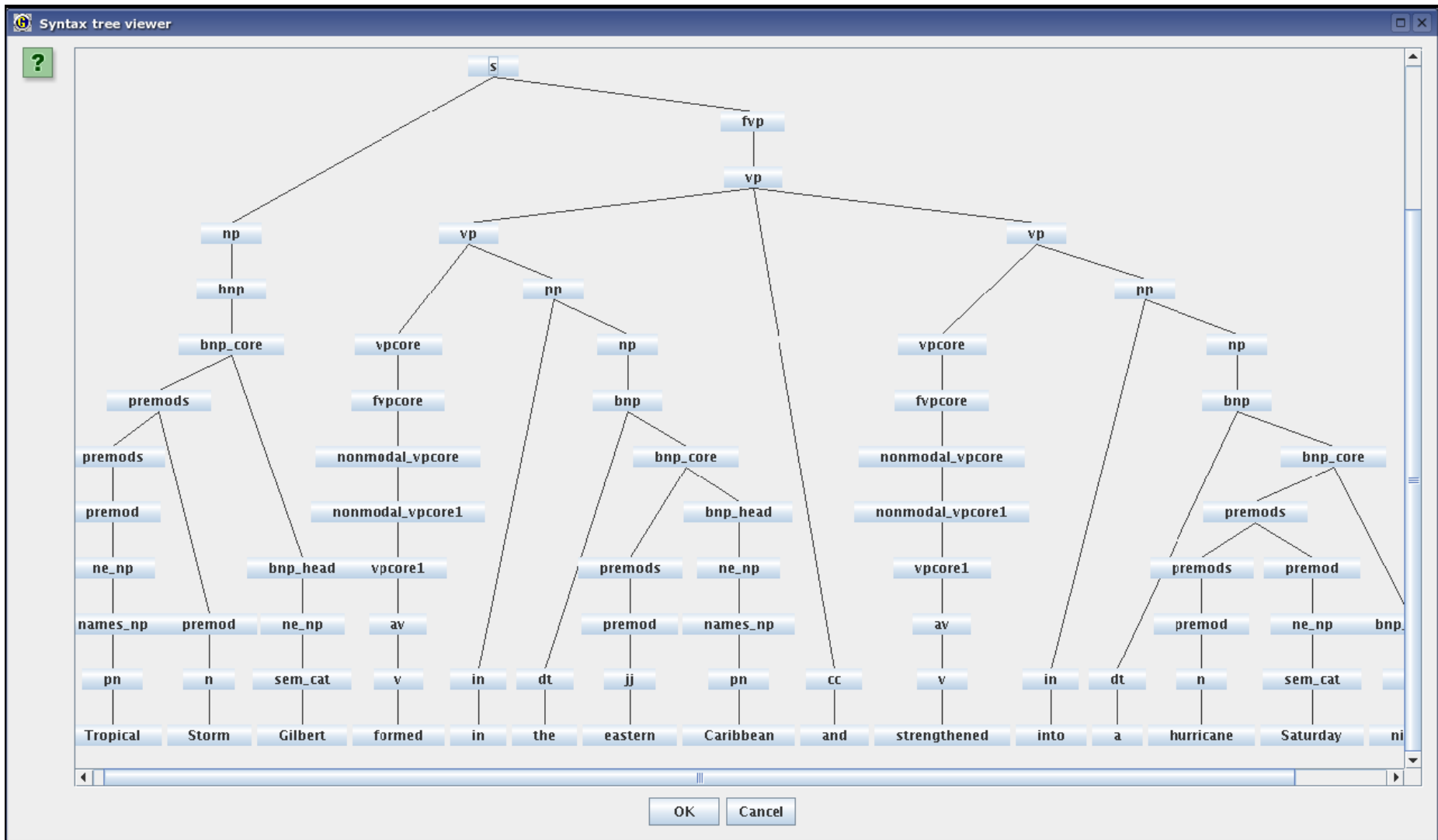
- **Parsing:** generating a parse tree for the given sentence (needs a grammar, and a lexicon)
- **Chunking:** finding syntactic constituents like Noun Phrases (NPs) or Verb Groups (VGs) within a sentence

- Parse trees can help in determining relationships such as:  
Who invented X?  
What company created product Y?  
Which organism is this protein coming from?
- Chunks are very useful in finding named entities (NEs), e.g., Persons, Companies, Locations, Patents, Organisms,



A Parse Tree

# Another Example of a Parse Tree





# AL-IMAM Database

Arabic-English Dictionary				Arabic Morphological Analysis						English Synonyms	
<i>A_ID</i>	<i>A_Word</i>	<i>E_Trans.</i>	<i>A_ID</i>	<i>Type</i>	<i>Root</i>	<i>Stem</i>	<i>Prefix</i>	<i>Suffix</i>	<i>Weigh.</i>	<i>E_ID</i>	<i>Synonym</i>
247	الفاحة	Planting	247	مصدر	فَلَح	فلاح	الـ	ة	فَعَلٌ	978	Farming
248	الفاحة	Farmer	Arabic Categorization (Learned)						978	Cultivating	
249	الفاحة	Success	<i>A_ID</i>	<i>Category</i>	<i>%</i>	<i>Disamb.</i>	<i>W_Code</i>		978	Agriculture	
English-Arabic Dictionary			247	زراعة	70	5%	TBD		978	Tilling	
<i>E_ID</i>	<i>E_Word</i>	<i>A_Trans.</i>	247	إنسان	5	?	TBD		Arabic Synonyms		
978	Planting	فِلاحة	247	الريف	25	10%	TBD		<i>A_ID</i>	<i>Synonym</i>	
Word Path in English Tree			English-Tree Titles			Arabic Tree Titles			247	حرثة	
<i>E_ID</i>	Tree Key		<i>E.T_ID</i>	Title		<i>A.T_ID</i>	Title		247	زراعة	
978	1.4.11.33.76.128.591		1	Action		1	شئ		Arabic Tree Links		
Human Factors			4	Group Action		3	شئ معنوى		<i>A_ID</i>	Tree Key	
<i>ID</i>	<i>Ani</i>	<i>ID</i>	<i>Gen.</i>	11	Commerce Trans.		10	مأكولات		247	1.3.10.31.65.97.154
274	Y	274	F	33	Industry		31	زراعة		SUMO Category	
Word Information			76	Production		65	محاصيل زراعية		Code	SUMO Categ.	
<i>ID</i>	<i>S/D/P</i>	<i>ID</i>	<i>Def.</i>	128	Cultivation		97	متطلبات زراعة		10837 4773	Subsuming Mapping (Putting)
274	S	274	Y	591	Farming		154	أشخاص			
WordNet Meaning			978	Planting		WordNet Sense (Glosses)					
978	108374773	putting seeds or young plants in the ground to grow				978	the planting of corn is hard work				



# *STATISTICS*

## *Part 6*

### *Introduction*

# Statistics

- Statistics is a Mathematical Science pertaining to the collection, analysis, interpretation or explanation, and presentation of data

# Statistical Terminologies

- Measures of Central Tendency (Mean, Median, Mode)
- Population Variance measures statistical dispersion of data points from the expected value (mean)
- Standard Deviation is a measure of the variability or dispersion of a population; Low SD indicates very close data points to the mean; High SD indicates spread out data points
- Covariance measures how much two variables change together
- Correlation (coefficient) indicates the strength and direction of a *linear* relationship between two random variables

$$\bar{x} = (1/n) \sum_{i=1}^n x_i$$

$$\begin{aligned} \text{Var}(X) &= E[(X - E(X))^2] \\ &= (1/n) \sum_{i=1}^n (x_i - \bar{x})^2 = \sigma^2 \end{aligned}$$

$$\text{sd}(X) = \sqrt{\sigma^2}$$

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{sd}(X) * \text{sd}(Y)} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

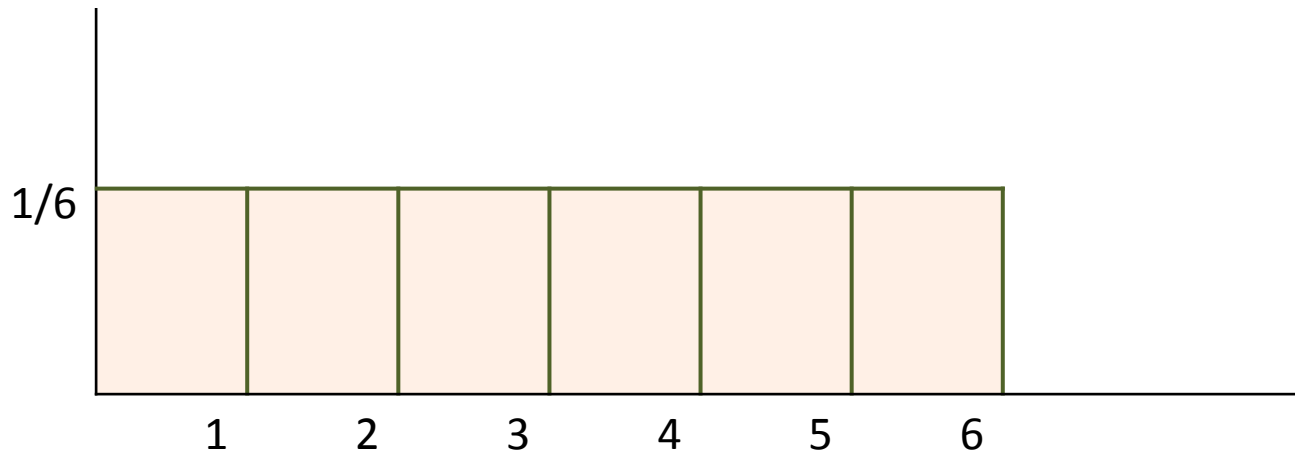
# Popular Distributions

Probability Distribution identifies the probability of each value of an unidentified random variable

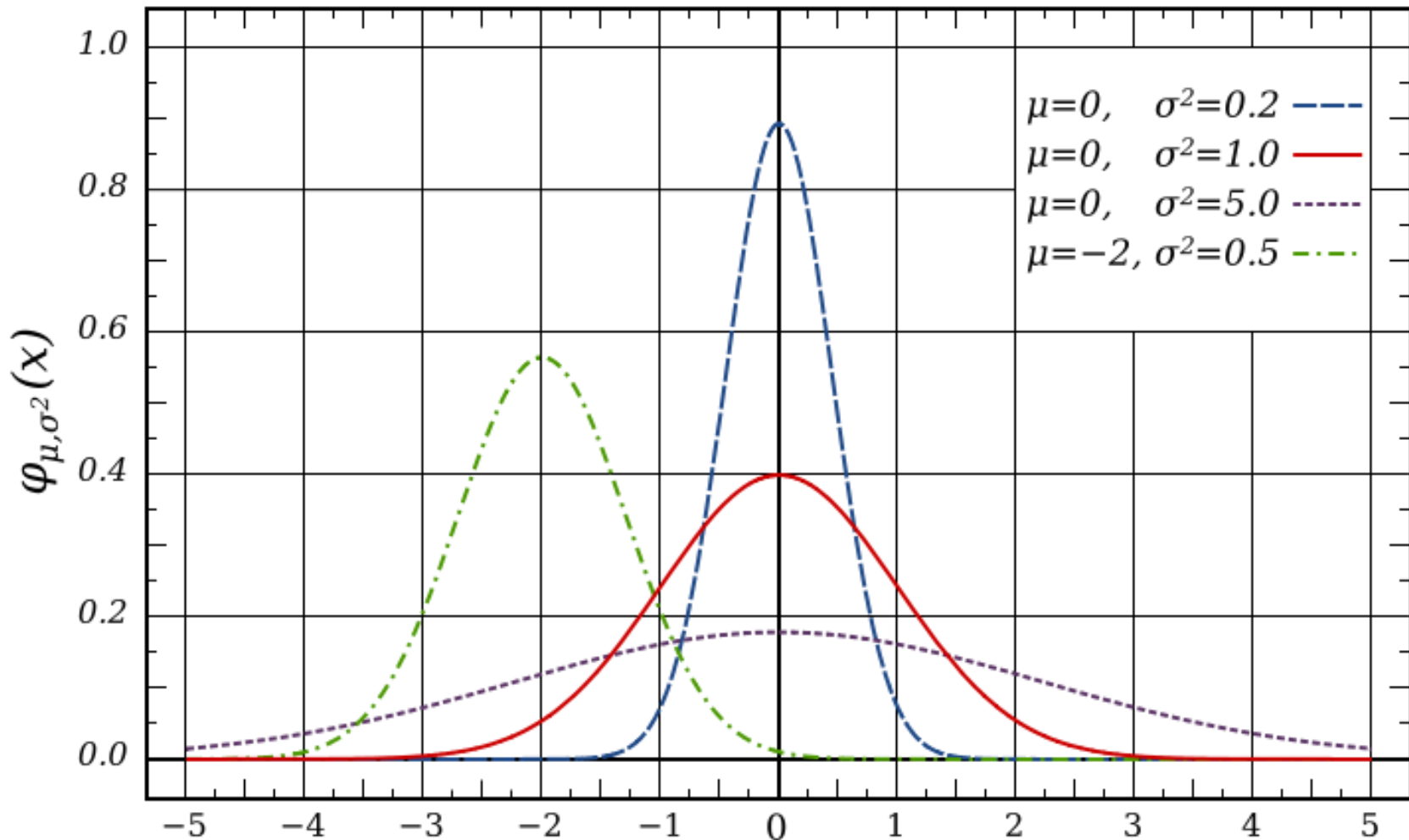
- *Uniform Distribution*
- *Normal (Gaussian) Distribution*
- *Chi-Square Distribution*
- *Exponential Distribution*
- *Poisson Distribution*
- *T Distribution*
- *F Distribution*

# The Uniform Distribution

- The probability is equal for all outcomes
- Suppose a fair dice is thrown, the probability of getting any of its 6 faces equal to  $1/6$
- The area under the line equal to 1

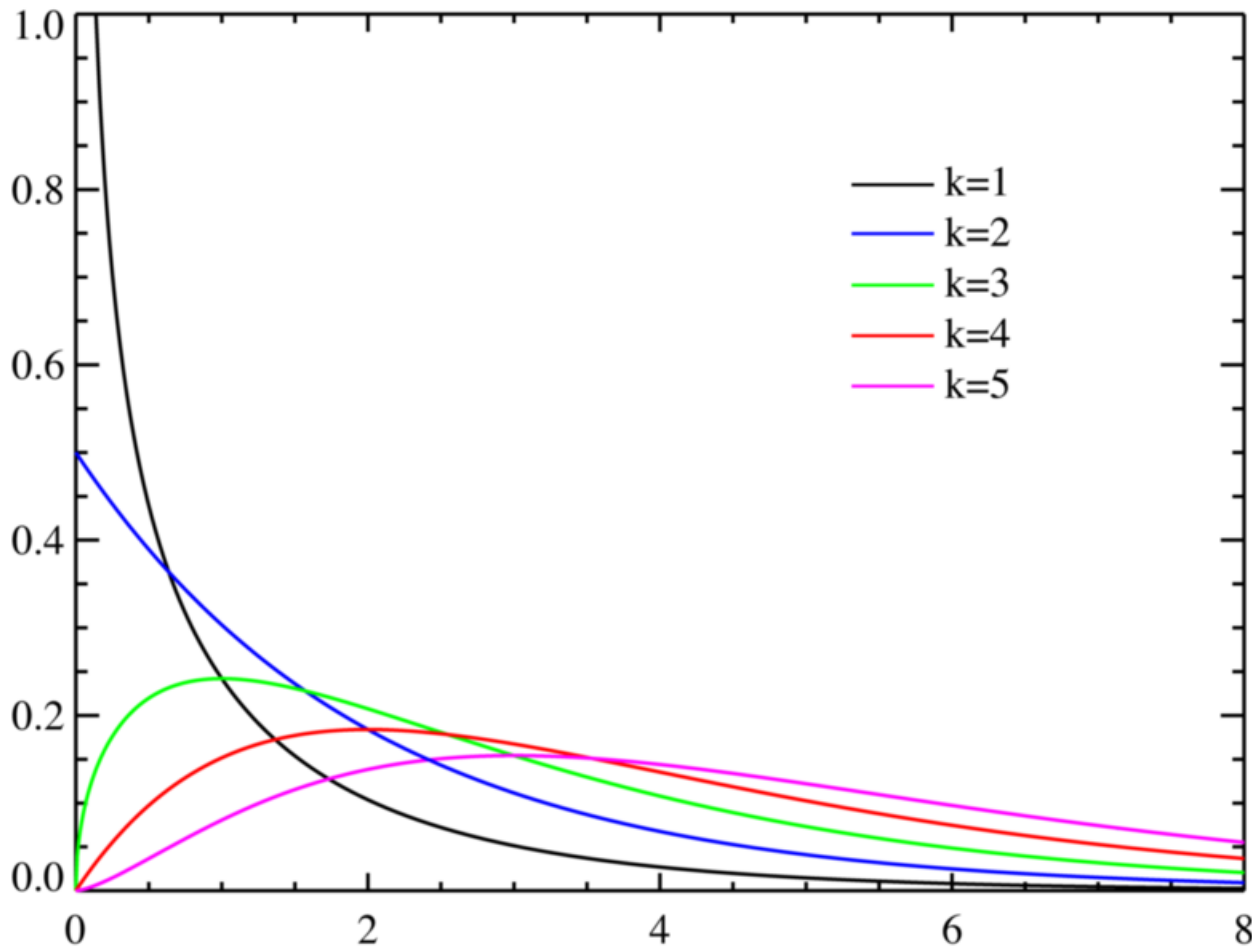


# The Normal/Gaussian Distribution



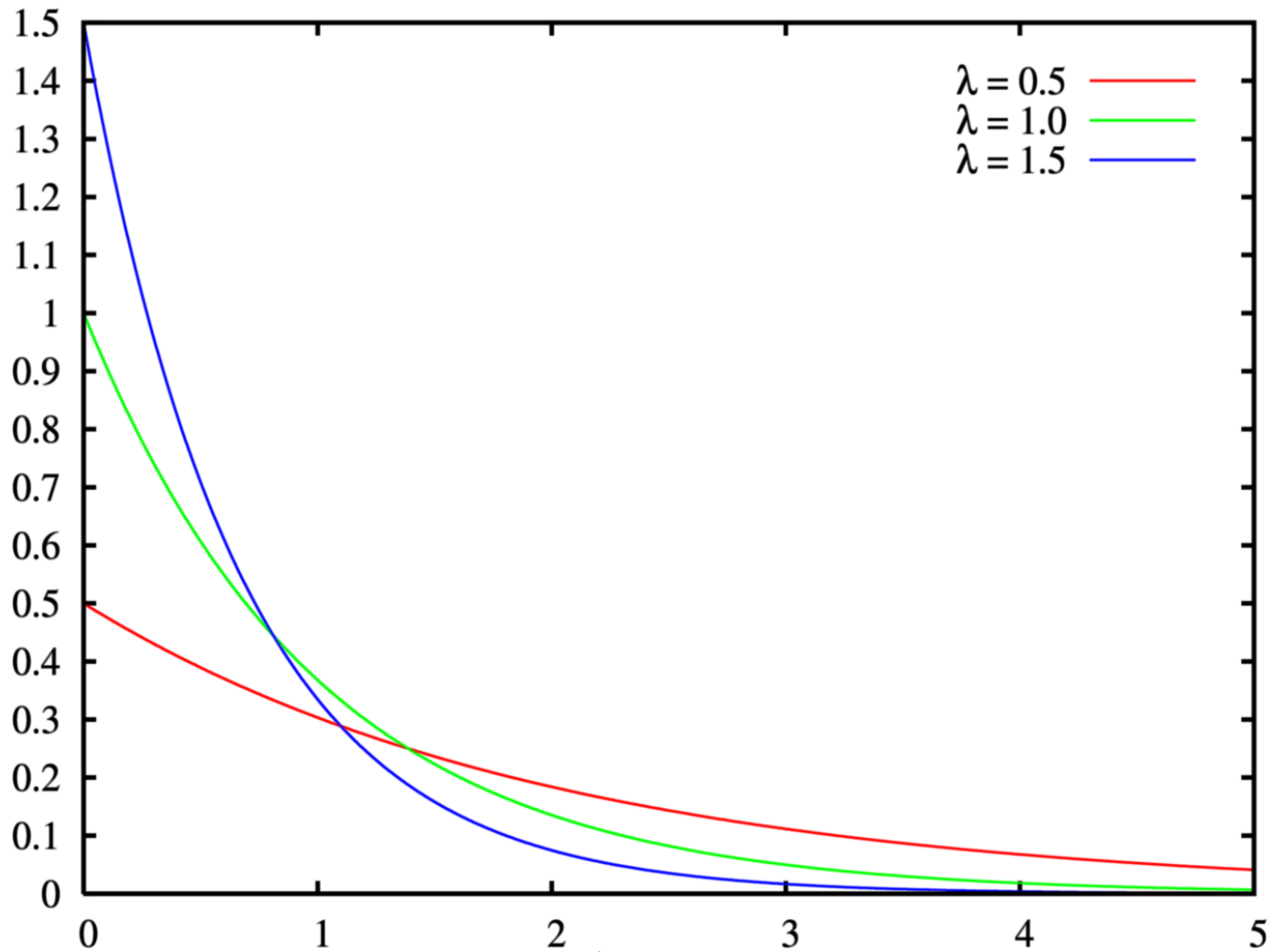
$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

# The Chi-Square Distribution



$$f(x; k) = \begin{cases} \frac{1}{2^{k/2} \Gamma(k/2)} x^{(k/2)-1} e^{-x/2} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

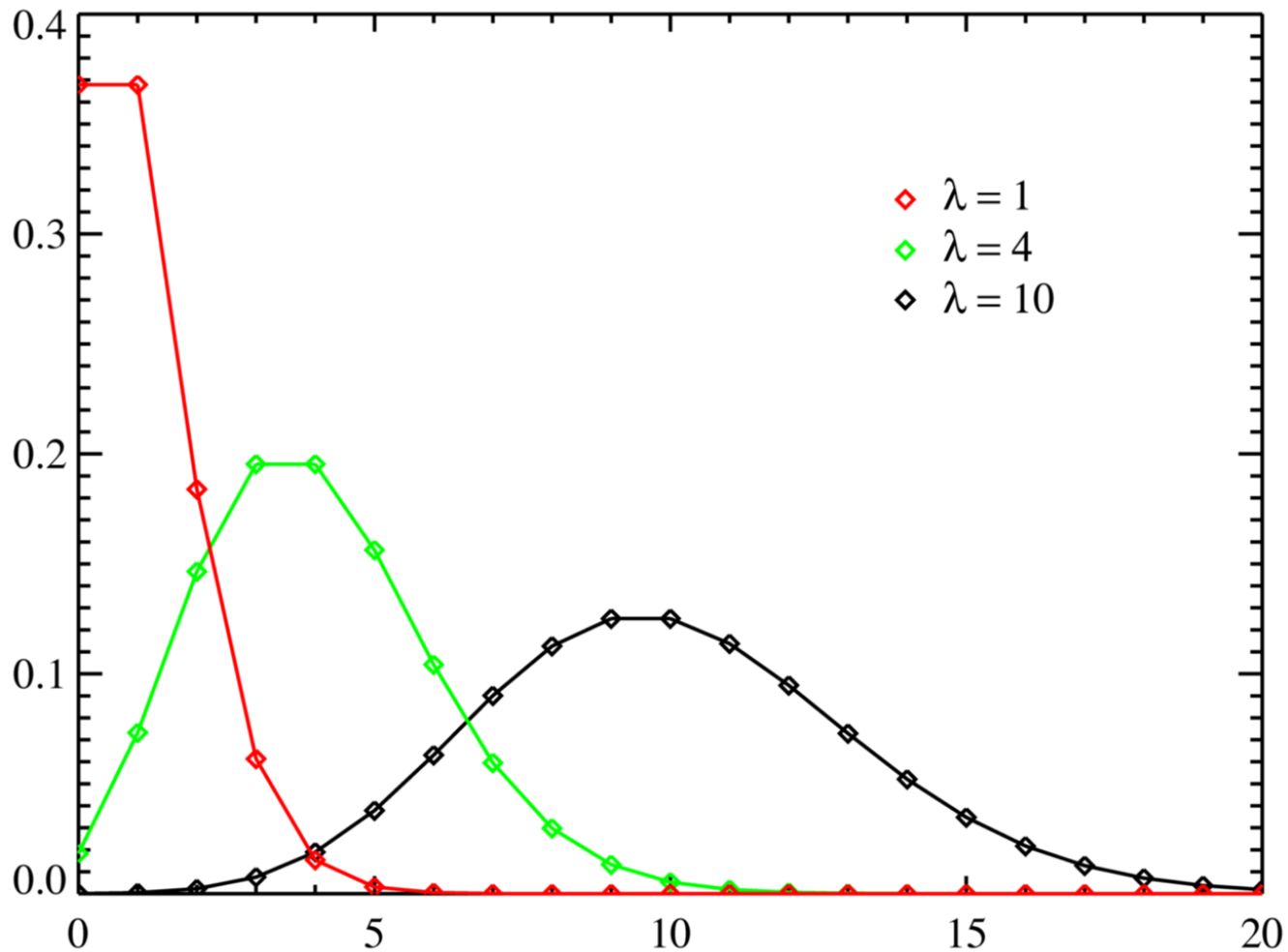
# The Exponential Distribution



$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

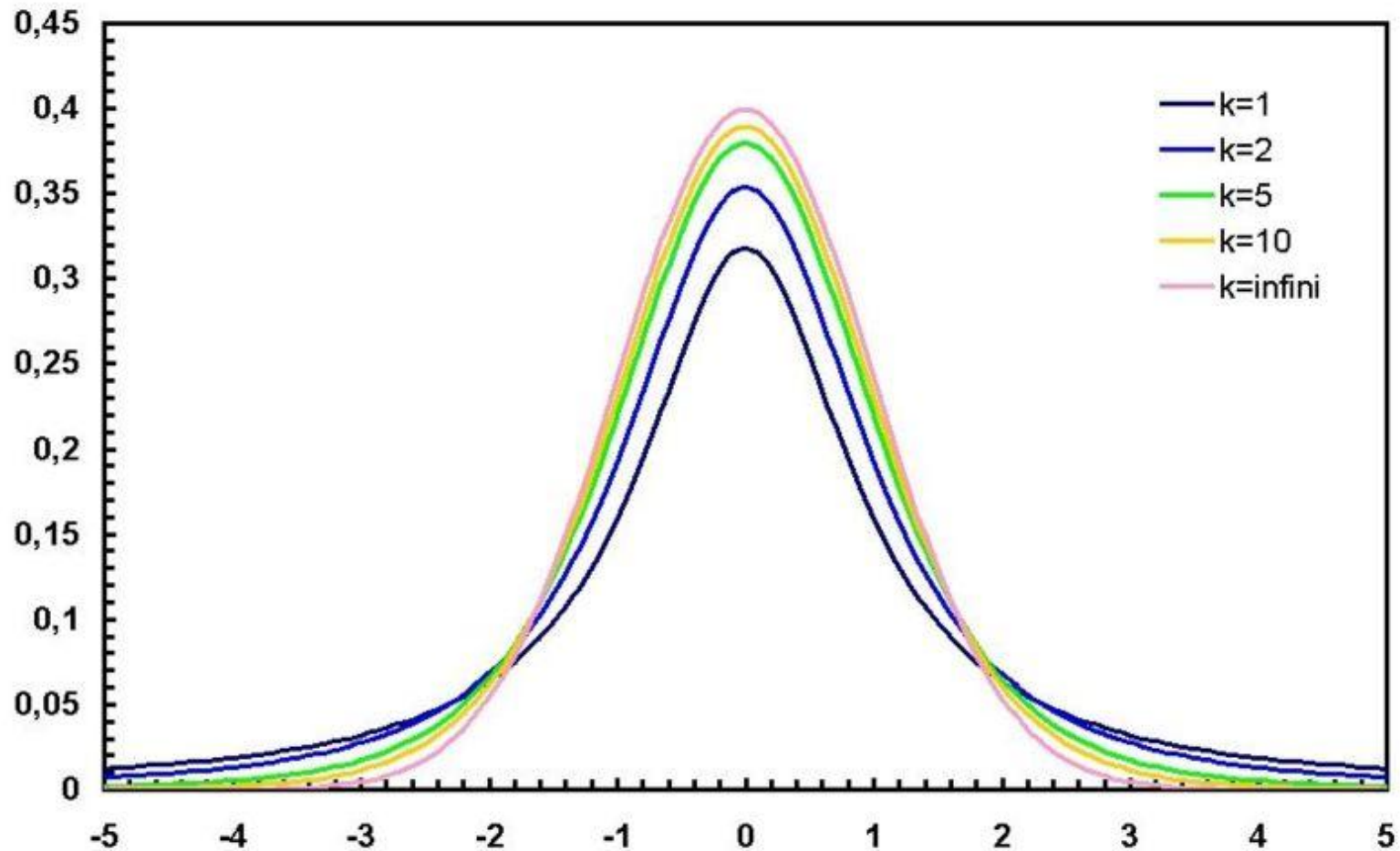


# The Poisson Distribution



$$f(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

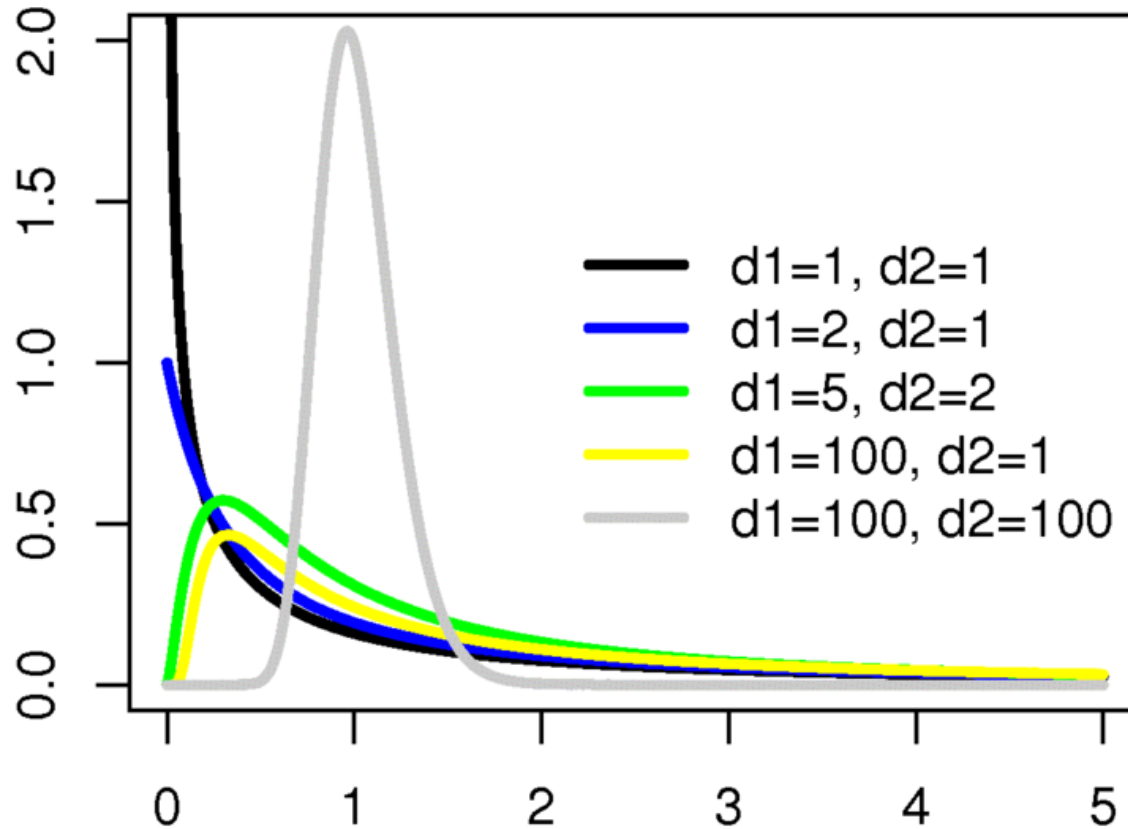
# The T Distribution



$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

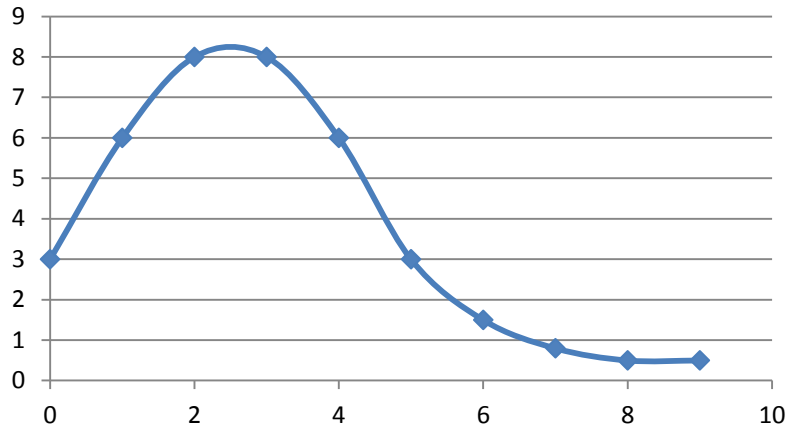
*t*-distribution arises in the problem of estimating the mean of a normally distributed population when the sample size is small

# The F Distribution



$$f(x) = \frac{\sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

# Fitting Chi-Square



Vector  
a

15  
14  
11  
11  
6  
5  
5

$$\max \chi^2 = \sum_{i=1}^n \frac{(a_i - E_i)^2}{E_i}$$

$$E_{ij} = (15 + 14 + 11 + 11 + 6 + 5 + 5) / 7 = 9.57$$

$$\chi^2 = (1/9.57) * ((15 - 9.57)^2 + (14 - 9.57)^2 + (11 - 9.57)^2 + (11 - 9.57)^2 + (6 - 9.57)^2 + (5 - 9.57)^2 + (5 - 9.57)^2) = 107.71 / 9.57 = 11.26$$

# Measuring Term-Category Correlation

$$\chi^2(t_k, c_i) = \frac{[P(t_k, c_i)P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i)P(\bar{t}_k, c_i)]^2}{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)}$$

$P(t_k, c_i)$  → probability document x contains term t and belongs to category c.

$P(\bar{t}_k, c_i)$  → probability document x does not contain term t and belongs to category c.

$P(t_k, \bar{c}_i)$  → probability document x contains term t and does not belong to category c.

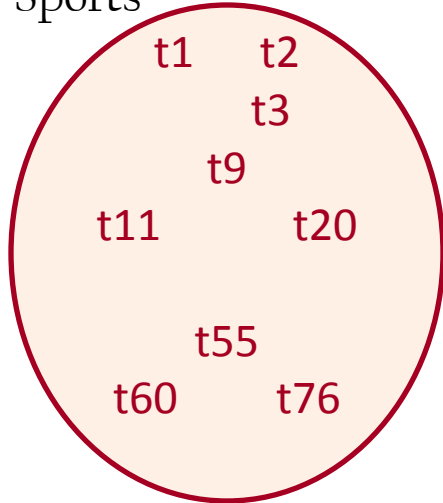
$P(\bar{t}_k, \bar{c}_i)$  → probability document x does not contain term t and does not belong to category c.

$P(t)$  → probability of term t

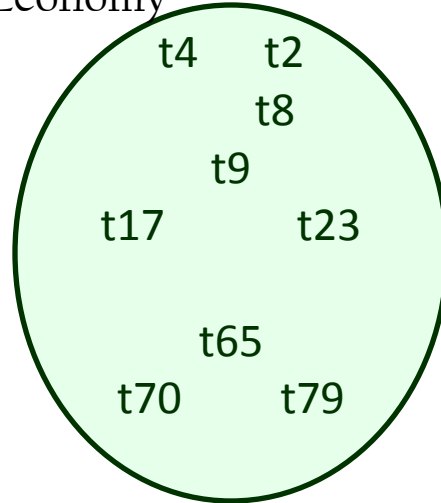
$P(c)$  → probability of category c

# Testing The Membership

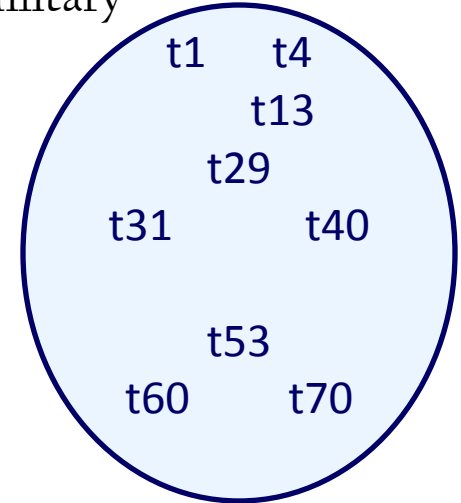
Sports



Economy



Military



$$\chi^2(t_k, c_i) = \frac{[P(t_k, c_i)P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i)P(\bar{t}_k, c_i)]^2}{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)}$$

$$\chi^2(t_1, Sports) = \frac{\left[ \frac{1}{9} * \frac{17}{18} - \frac{1}{18} * \frac{8}{9} \right]^2}{\frac{2}{27} * \frac{25}{27} * \frac{9}{27} * \frac{18}{27}}$$

# Using Chi-Square for Categorization

Another Example:

Term	Frequency per Category				Total
	Communication	Phone	Business	Army	
Link	15	6	2	12	35
Wire	10	12	0	8	30
<b>Total</b>	25	18	2	20	<b>65</b>

$$\chi^2(\text{link}, \text{phone}) = \frac{[6/65 * (18/65) - (29/65) * (12/65)]^2}{(35/65) * (30/65) * (18/65) * (47/65)}$$

## *Using Chi-Square for Multiple sets of Terms*

Group 1	Category		Total
	News	Sports	
Term 1	3	2	5
Term 2	0	4	4
Term 3	2	3	5
Total	5	9	14

Group 2	Category		Total
	News	Sports	
Term 5	1	3	4
Term 7	4	6	10
Total	5	9	14

$$\chi^2 = \sum_{i=1}^n \sum_{j=1}^m \frac{(a_{ij} - E_{ij})^2}{E_{ij}}$$

$$E_{ij} = \frac{(T_{ci} * T_{vj})}{T}$$

$$\begin{aligned} \chi^2(\text{Group 1}) &= (3-1.78)^2 / 1.78 + (2-3.21)^2 / 3.21 + (0-1.42)^2 / 1.42 \\ &\quad + (4-2.57)^2 / 2.57 + (2-1.78)^2 / 1.78 + (3-3.21)^2 / 3.21 = 3.62 \end{aligned}$$

$$\begin{aligned} \chi^2(\text{Group 2}) &= (1-1.42)^2 / 1.42 + (3-2.57)^2 / 2.57 + (4-3.57)^2 / 3.57 \\ &\quad + (6-6.43)^2 / 6.43 = \end{aligned}$$



# Attribute Selection Criteria: Chi-Square

## Example

- T2 is quantized into two intervals 21 ( $T2 \leq 21$ ) and ( $T2 > 21$ )
- T3 is quantized into two intervals 15 ( $T3 \leq 15$ ) and ( $T3 > 15$ )

T2	Decision D		Total
	0	1	
$\leq 21$	1	3	4
$> 21$	4	6	10
Total	5	9	14

T1	Decision D		Total
	0	1	
1	3	2	5
2	0	4	4
3	2	3	5
Total	5	9	14

T3	Decision D		Total
	0	1	
$\leq 15$	1	4	5
$> 15$	4	5	9
Total	5	9	14

T4	Decision D		Total
	0	1	
A	3	3	6
B	2	6	8
Total	5	9	14

T1	T2	T3	T4	D
1	25	10	A	1
1	30	30	A	0
1	35	25	B	0
1	22	35	B	0
1	19	10	B	1
2	22	30	A	1
2	33	18	B	1
2	14	5	A	1
2	31	15	B	1
3	21	20	A	0
3	15	10	A	0
3	25	20	B	1
3	18	20	B	1
3	20	36	B	1

# Attribute Selection Criteria: Chi-Square

$$\chi^2(A) = \sum_{i=1}^n \sum_{j=1}^m \frac{(a_{ij} - E_{ij})^2}{E_{ij}}$$

where A is the attribute to be evaluated against the decision attribute, n is the number of distinct values of A, m is the number of distinct values of the decision attribute,  $a_{ij}$  is the correlation frequency of value number i from A and value number j from the decision attribute;

$$E_{ij} = \frac{(T_{ci} * T_{vj})}{T}$$

where  $T_{ci}$  is the total number of examples belonging to class  $c_i$ ,  $T_{vj}$  is the number of examples containing the value  $v_j$  of the given attribute

$$\begin{aligned} \chi^2(T1) &= (3-1.78)^2 / 1.78 + (2-3.21)^2 / 3.21 + (0-1.42)^2 / 1.42 \\ &+ (4-2.57)^2 / 2.57 + (2-1.78)^2 / 1.78 + (3-3.21)^2 / 3.21 = 3.62 \end{aligned}$$

$$\begin{aligned} \chi^2(T4) &= (3-3.9)^2 / 3.9 + (3-2.1)^2 / 2.1 + (6-5.1)^2 / 5.1 \\ &+ (2-2.9)^2 / 2.9 = 1.1 \end{aligned}$$

T1	D		Total
	0	1	
1	3	2	5
2	0	4	4
3	2	3	5
Total	5	9	14

T2	D		Total
	0	1	
<=21	1	3	4
>21	4	6	10
Total	5	9	14

T3	D		Total
	0	1	
<=15	1	4	5
>15	4	5	9
Total	5	9	14

T4	D		Total
	0	1	
A	3	3	6
B	2	6	8
Total	5	9	14

# *STATISTICS*

## *Part 7*

### *Regression*

# Linear Regression

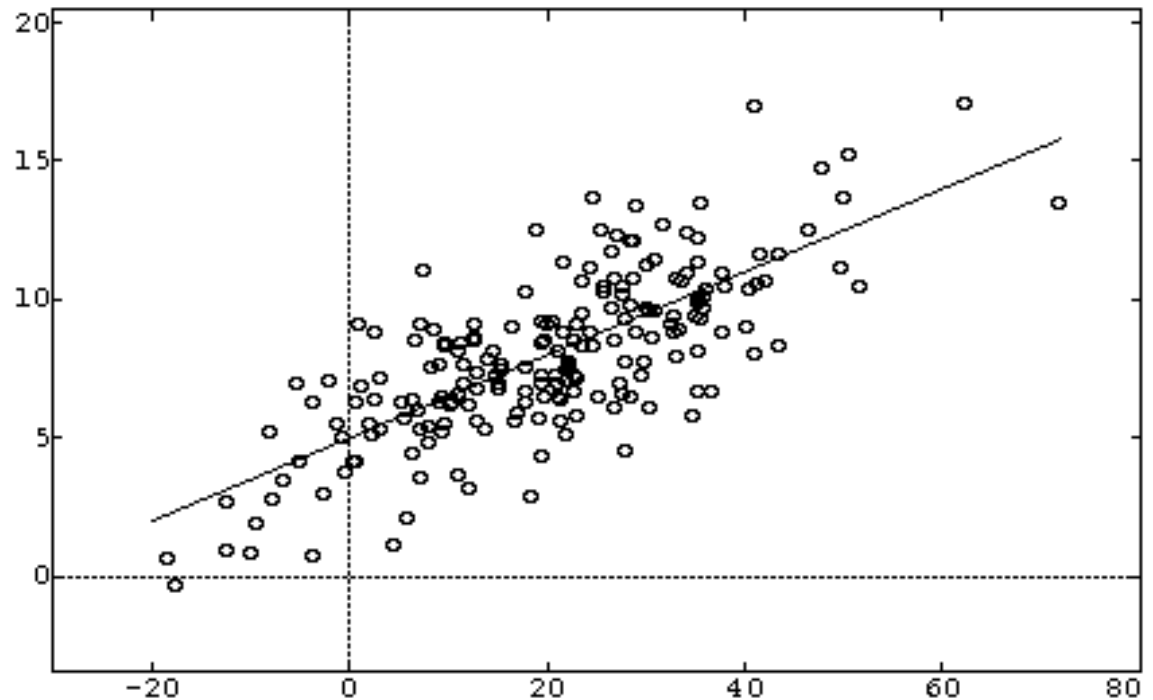
- The linear model states that the dependent variable is directly proportional to the value of the independent variable
- Thus if a theory implies that Y increases in direct proportion to an increase in X, it implies a specific mathematical model of behavior

$$y = ax + b$$

In case of two dimensions

$$a = \text{slope} = \frac{(y_2 - y_1)}{(x_2 - x_1)}$$

$$b = y_2 - \text{slope} * x_2$$



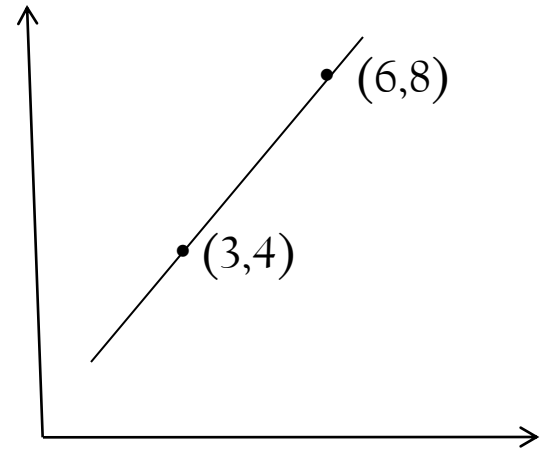
# Linear Regression

$$y = ax + b$$

$$8 = 6a + b \quad \& \quad 4 = 3a + b$$

$$\frac{8-b}{6} = a \quad \& \quad 4 = 3 * \frac{8-b}{6} + b$$

$$b = 0 \quad \& \quad a = \frac{4}{3} = 1.333$$



$$Slope = \frac{8-4}{6-3} = 1.333$$

$$b = 4 - \frac{4}{3} * 3 = 0$$

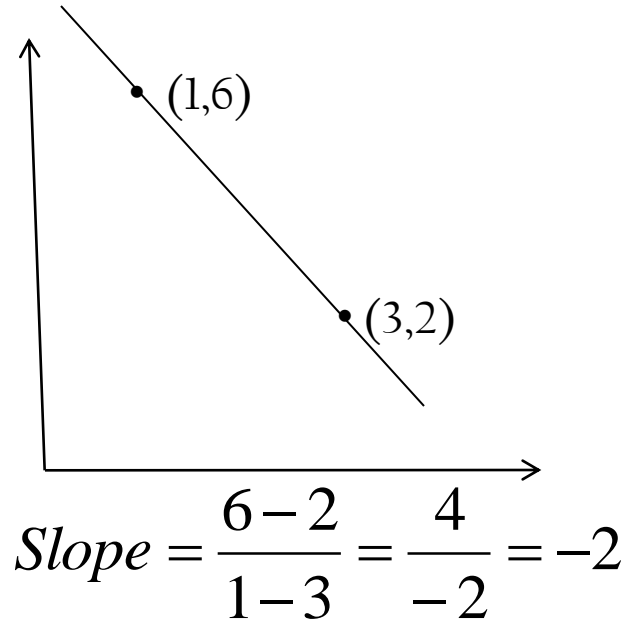
# Linear Regression

$$y = ax + b$$

$$6 = a + b \quad \& \quad 2 = 3a + b$$

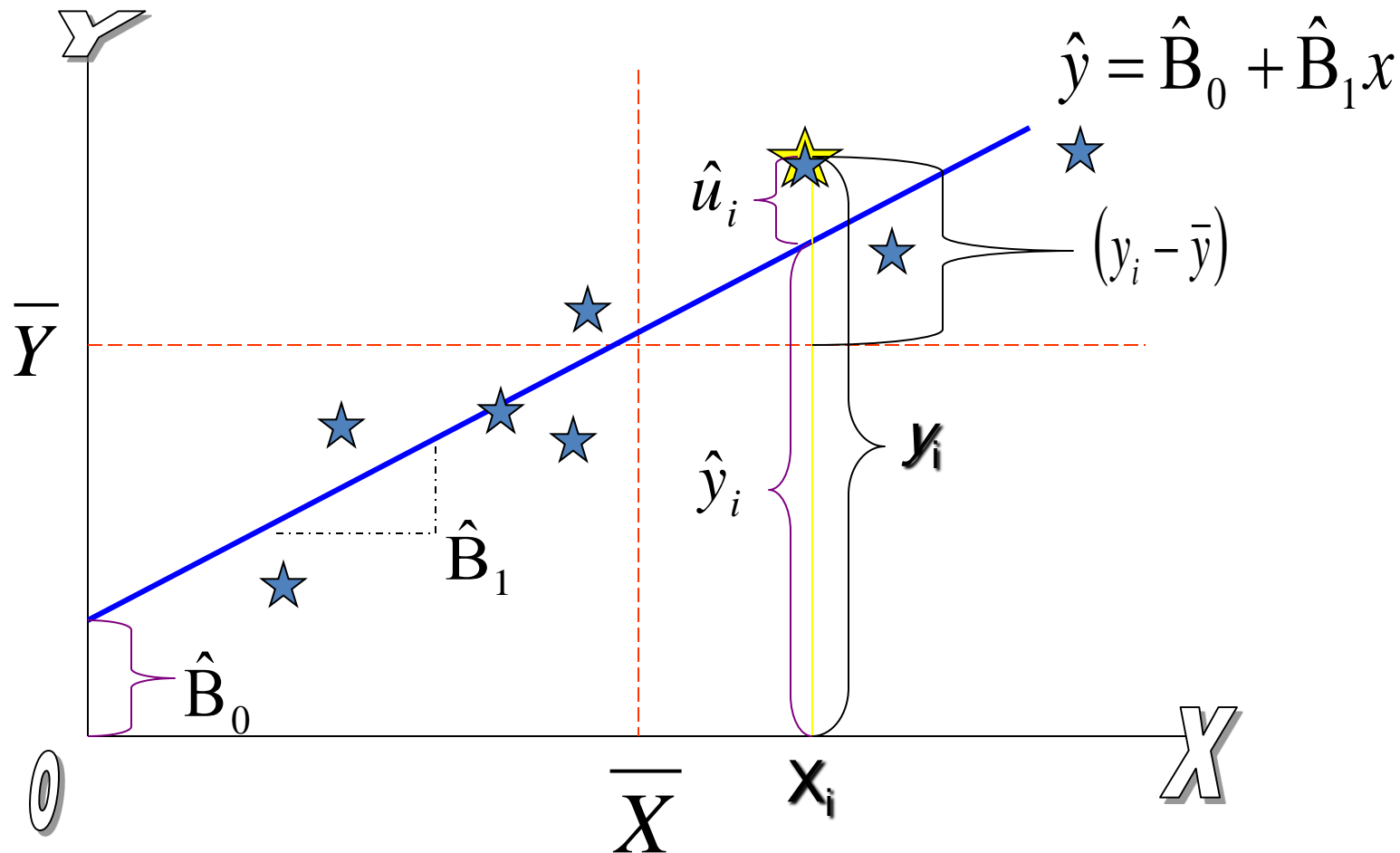
$$6 - b = a \quad \& \quad 2 = 3 * (6 - b) + b$$

$$b = 8 \quad \& \quad a = 6 - 8 = -2$$



$$b = 2 + 2 * 3 = 8$$

# Linear Regression



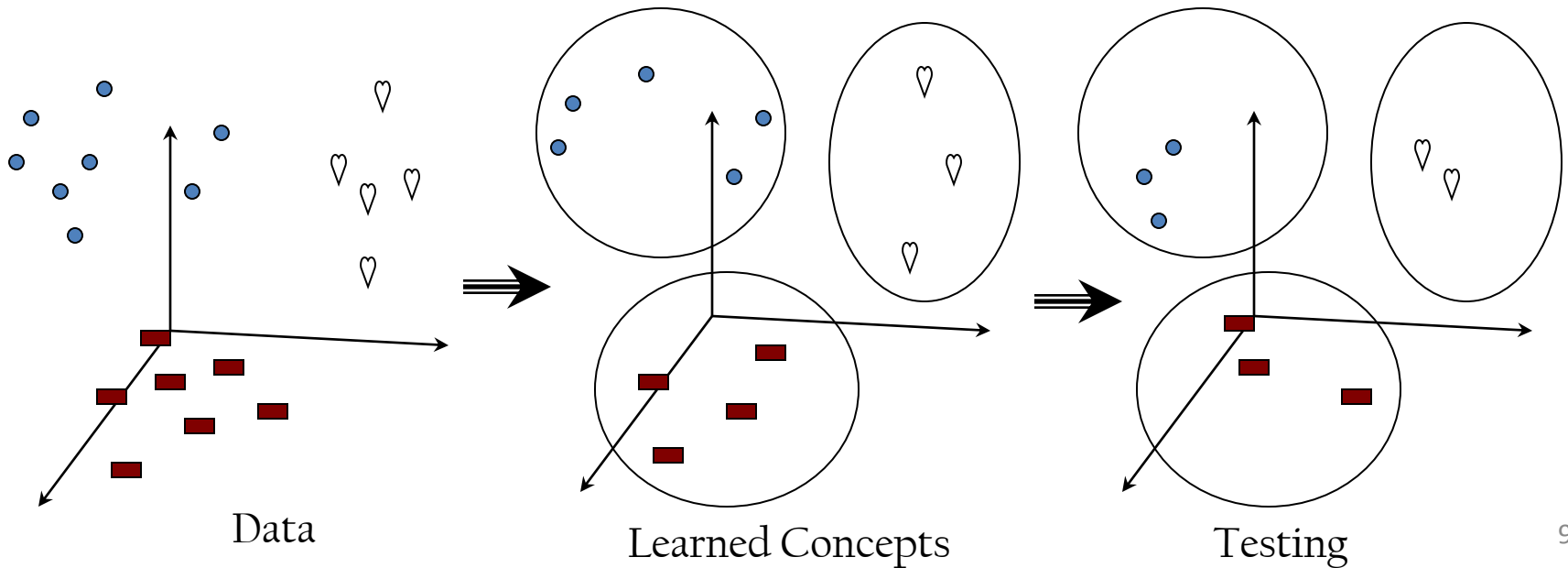
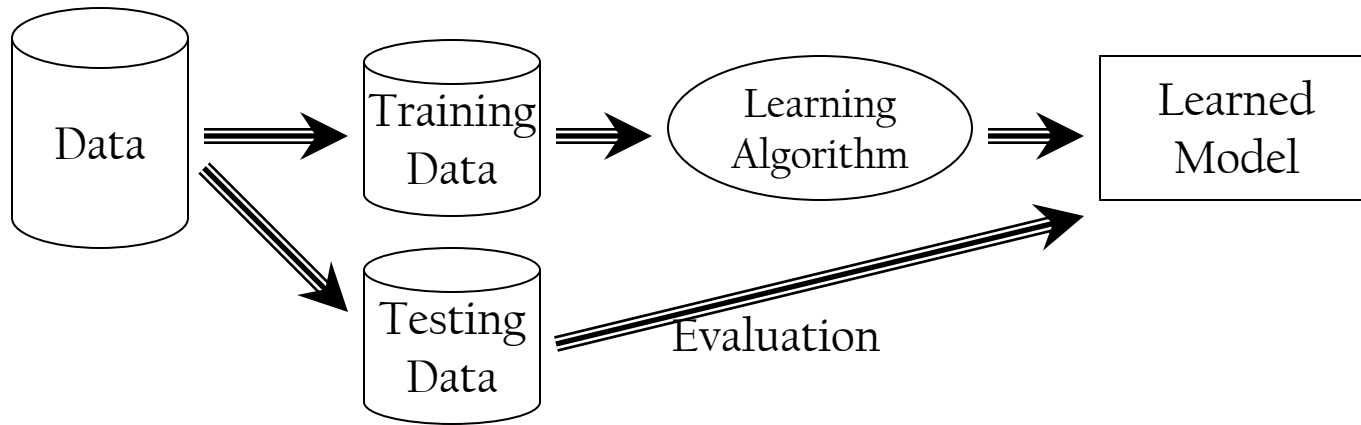
# *Statistics and Testing*

## *Part 8*

### *Testing Samples & Calculating Accuracy*



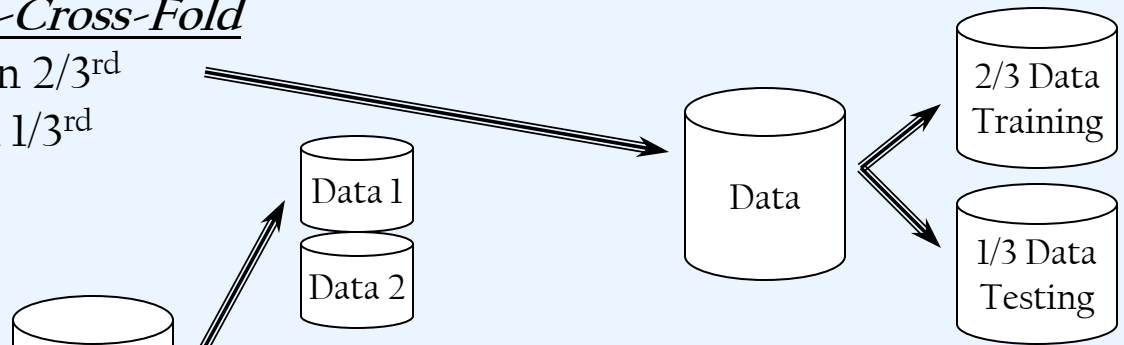
# Training & Testing



# Testing Approaches

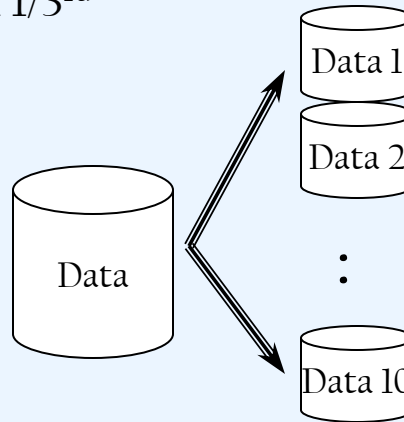
- Two-Cross-Fold

Train on  $2/3^{\text{rd}}$   
Test on  $1/3^{\text{rd}}$



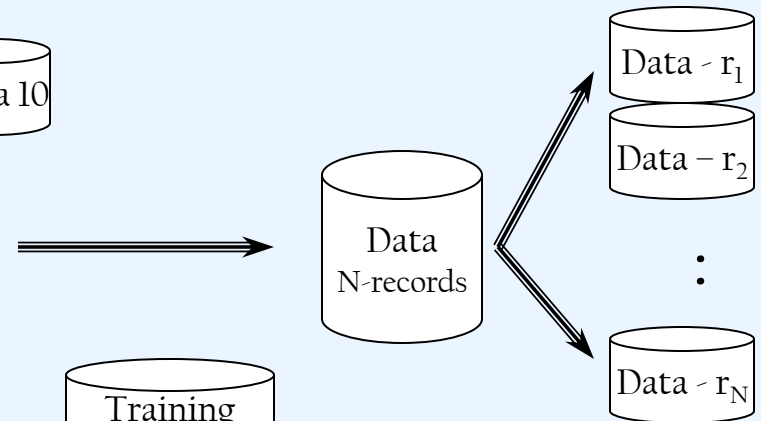
- Ten-Cross-Fold

Train on  $9/10^{\text{th}}$   
Test on  $1/10^{\text{th}}$   
Repeat 10 times



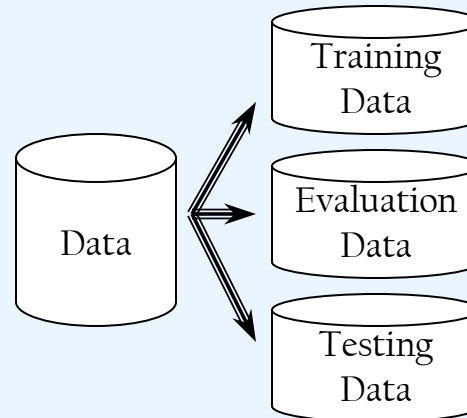
- Hold-One-Out

Train on all data but one  
Test on the selected one



- Learning Evaluation vs. Testing

Train on Training Data  
Evaluate on Evaluation Data  
Test on Testing Data



# Accuracy & Error

Example: Suppose you have a classification model C, and 100 testing records from two classes (P & N). Suppose the following are the classification results:

## ● Accuracy vs. Error Rate

- Accuracy =  $(40+45)/100 = 85\%$
- Error Rate =  $(10+5)/100 = 15\%$

		Actual	
		P	N
Obtained	P	TP	FP
	N	FN	TN

## ● True vs. False Classification

- True Positive: = 88.88%
- True Negative: = 81.82%
- False Positive: = 11.12%
- False Negative: = 18.18%

		Actual	
		P	N
Obtained	P	40	10
	N	5	45

## ● Flexible Matching

- *Using Nearest Neighbors (e.g., majority of nearest 3 neighbors)*
- Using Fuzzy rules (assigning probability for each decision and taking it into consideration when calculating the accuracy)
- Assigning small weights for the false positive and false negative results (not zero)

## ● Testing for Multiple Classes ????

# *Precision, Recall, and F-Measure*

Accuracy: is the percentage of correct results

Error: is the percentage of wrong results

Accuracy only reacts to real errors, and doesn't show how many correct results have been found as such

Precision:

Precision shows the percentage of correct results within an answer:

$$\text{Precision} = (tp) / (tp + fp)$$

Recall:

Recall is the percentage of the correct system results over all correct results:

$$\text{Recall} = (tp) / (tp + fn)$$

*Makhoul, John; Francis Kubala; Richard Schwartz; Ralph Weischedel: [Performance measures for information extraction](#). In: Proceedings of DARPA Broadcast News Workshop, Herndon, VA, February 1999*

# *Precision, Recall, and F-Measure*

Precision and Recall can be defined differently for different tasks

For example: In Information Retrieval,

- Recall =  $|\{\text{relevant documents}\} \cap \{\text{documents retrieved}\}| / |\{\text{relevant documents}\}|$
- Precision =  $|\{\text{relevant documents}\} \cap \{\text{documents retrieved}\}| / |\{\text{documents retrieved}\}|$

# *Precision, Recall, and F-Measure*

## *F-Measure (harmonic mean):*

$F_\beta$  “measures the effectiveness of  $\beta$  times as much importance to recall as precision”. The general form of F-Measure:

$$F_\beta = (1 + \beta^2) * (\text{precision} * \text{recall}) / (\beta^2 * \text{precision} + \text{recall})$$

when  $\beta=1$ ,

$$F_1 = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$$

# *STATISTICS*

## *Part 9*

### *Test of Significance*

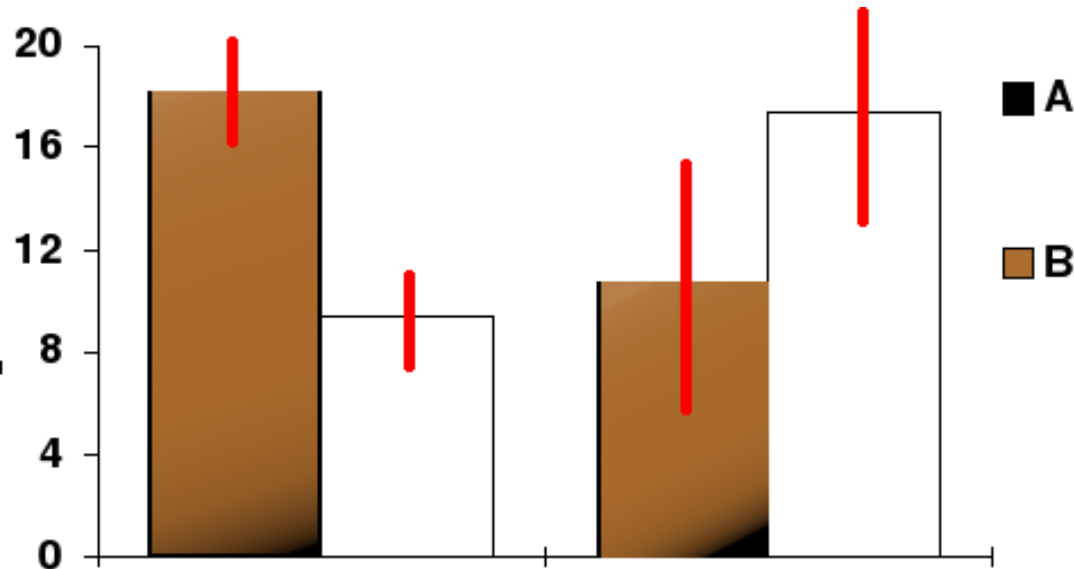
# Test of Significance (1/5)

- The probability that a result is not due to chance; or Is the observed value differs enough from a hypothesized value?
- The hypothesized value is called the null hypothesis
- If this probability is sufficiently low, then the difference between the parameter and the statistic is said to be "statistically significant"
- Just how low is sufficiently low? The choice of 0.05 and 0.01 are most commonly used

- Suppose your algorithm produced error rate of 1.5 and another algorithm produced an error of 2.1 on the same data set; are the two algorithms similar?



## Test of Significance (2/5)



- The top ends of the bars indicate observation means
- The red line segments represent the confidence intervals surrounding them
- The difference between the two populations on the left is significant
- However, it is a common misconception to suppose that two parameters whose 95% confidence intervals fail to overlap are significantly different at the 5% level

## Test of Significance (3/5)

- The system you are comparing against reported results of 250; the value reported is considered as a random variable  $X$ ; the distribution of  $X$  is assumed as normal distribution with unknown mean and standard deviation  $\sigma=2.5$ ; You ran your system 25 times; it reported values ( $x_1, x_2, \dots, x_{25}$ ); the average of these values is 250.2.

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^{25} x_i = 250.2$$

Sample Mean

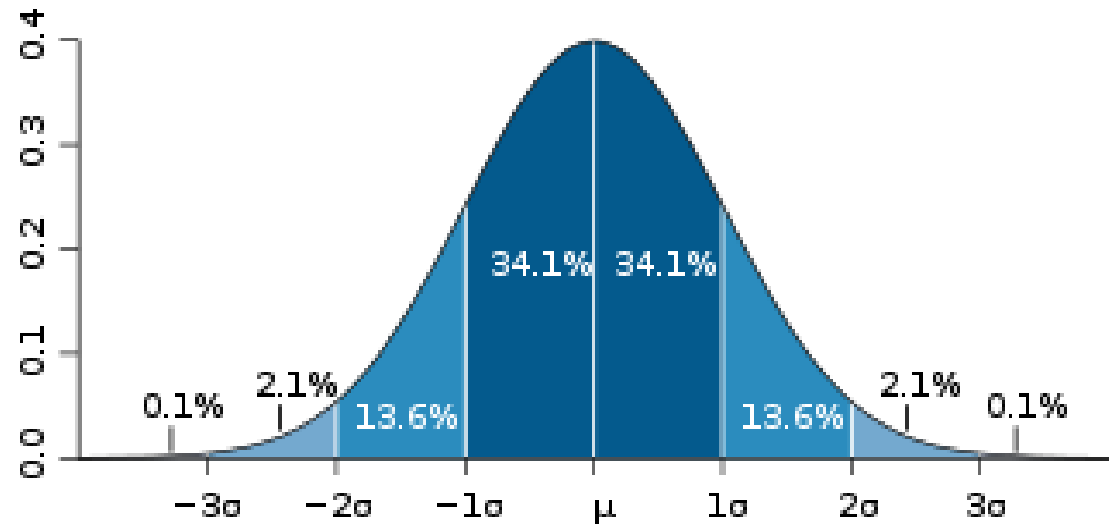
$$\text{Standard Error} = \sigma / \sqrt{n} = 2.5 / \sqrt{25} = 0.5$$

$n$  is the sample size

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{\bar{X} - \mu}{0.5}$$

$\mu$  is not known

# Test of Significance (4/5)



$$P(-z \leq Z \leq z) = 1 - \alpha = 0.95$$

$$\Phi(z) = P(Z \leq z) = 1 - \frac{\alpha}{2} = 0.975$$

From Tables

$$z = \Phi^{-1}(\Phi(z)) = \Phi^{-1}(0.975) = 1.96$$

$$0.95 = 1 - \alpha = P(-z \leq Z \leq z) = P(-1.96 \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq 1.96)$$

## Test of Significance (5/5)

$$P(-z \leq Z \leq z) = P(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}})$$

$$P(-z \leq Z \leq z) = P(\bar{X} - 1.96 * 0.5 \leq \mu \leq \bar{X} + 1.96 * 0.5)$$

$$P(-z \leq Z \leq z) = P(\bar{X} - 0.98 \leq \mu \leq \bar{X} + 0.98)$$

$$\text{Our Interval} = (250.2 - 0.98; 250.2 + 0.98)$$

$$\text{Our Interval} = (249.22; 251.0)$$

- Any value within this interval is not significant

# *The Information Theory*

## *Part 9*

*Introduction*  
*Entropy*

# The Information Theory

The information conveyed by a message can be measured in bits by its probability

# The Information Theory: Given Data

*Attributes:*

*D1, D2, D3, D4*

*Domain(D1)={1,2,3}*

*Domain(D2)={1,2}*

*Domain(D3)={1,2}*

*Domain(D4)={A,B}*

D1	D2	D3	D4	D5
1	2	1	A	1
1	2	2	A	0
1	2	2	B	0
1	2	2	B	0
1	1	1	B	1
2	2	2	A	1
2	2	2	B	1
2	1	1	A	1
2	2	1	B	1
3	1	2	A	0
3	1	1	A	0
3	2	2	B	1
3	1	2	B	1
3	1	2	B	1

*Decision Attributes: D5*

*Domain(D5)={0,1}*

*Two Decisions: 0, 1*

# The Information Theory: Given Data

		D1		2		3	
		1	2	1	2	1	2
A	D4		1	1		0	
	D3\D2		0		1	0	
B	D4	1	1		1	1	
	D3\D2		0		1	1	1

D1	D2	D3	D4	D5
1	2	1	A	1
1	2	2	A	0
1	2	1	B	0
1	2	2	B	0
1	1	1	B	1
2	2	2	A	1
2	2	2	B	1
2	1	1	A	1
2	2	1	B	1
3	1	2	A	0
3	1	1	A	0
3	2	2	B	1
3	1	1	B	1
3	1	2	B	1



# *The Information Theory: Entropy*

THE INFORMATION THEORY: information conveyed by a message depends on its probability and can be measured in bits as minus the logarithm (base 2) of that probability

suppose  $D_1, \dots, D_m$  are  $m$  attributes and  $C_1, \dots, C_n$  are  $n$  decision classes in a given data. Suppose  $S$  is any set of cases, and  $T$  is the initial set of training cases  $S \subset T$ . The frequency of class  $C_i$  in the set  $S$  is:

$$\text{freq}(C_i, S) = \text{Number of examples in } S \text{ belonging to } C_i$$

If  $|S|$  is the total number of examples in  $S$ , the probability that an example selected at random from  $S$  belongs to class  $C_i$  is

$$\text{freq}(C_i, S) / |S|$$

The information conveyed by the message that “a selected example belongs to a given decision class,  $C_i$ ”, is determined by

$$-\log_2(\text{freq}(C_i, S) / |S|) \quad \text{bits}$$

# *The Information Theory: Entropy*

The information conveyed by the message “*a selected example belongs to a given decision class,  $C_i$* ”

$$-\log_2(\text{freq}(C_i, S) / |S|) \quad \text{bits}$$

*The Entropy:* The expected information from a message stating class membership is given by

$$\text{Info}(S) = -\sum_{i=1}^k (\text{freq}(C_i, S) / |S|) * \log_2(\text{freq}(C_i, S) / |S|) \quad \text{bits}$$

$\text{info}(S)$  is known as the *entropy* of the set  $S$ . When  $S$  is the initial set of training examples, *info(S) determines the average amount of information needed to identify the class of an example in S.*

# The Information Theory: The Gain Ratio

S

Example

$$freq(0, S) = 5$$

$$freq(1, S) = 9$$

$$freq(0, S) / |S| = 5/14$$

$$freq(1, S) / |S| = 9/14$$

The Entropy: the average amount of information needed to identify the class of an example in S

$$Info(S) = -9/14 * \log_2(9/14) - 5/14 * \log_2(5/14) = 0.94bits$$

Using D<sub>1</sub> to Split the data provide 3 subsets of data

$$Info_{D_1}(S_1) = -3/5 * \log_2(3/5) - 2/5 * \log_2(2/5) = 0.94$$

$$Info_{D_1}(S_2) = -4/4 * \log_2(4/4) = 0.94$$

$$Info_{D_1}(S_3) = -2/5 * \log_2(2/5) - 3/5 * \log_2(3/5) = 0.94$$

$$Info_{D_1}(S) = (5/14) * Info_{D_1}(S_1) + (4/14) * Info_{D_1}(S_2) + (5/14) * Info_{D_1}(S_3) = 0.694$$

D1	D2	D3	D4	D5
1	2	1	A	1
1	2	2	A	0
1	2	2	B	0
1	2	2	B	0
1	1	1	B	1
2	2	2	A	1
2	2	2	B	1
2	1	1	A	1
2	2	1	B	1
3	1	2	A	0
3	1	1	A	0
3	2	2	B	1
3	1	2	B	1
3	1	2	B	1

## *The Information Theory: The Gain Ratio*

Suppose attribute  $D_i$  is selected to be the root and it has  $k$  possible values. The expected information of selecting  $D$  to partition the training set  $S$ ,  $\text{info}_{D_i}(S)$ , can be calculated as follows:

$$\text{Info}_{D_i}(S) = \sum_{i=1}^k \left( \frac{|S_i|}{|S|} \right) * \text{Info}(S_i)$$

$S_i$  is the subset number  $i$  of the data;  $k$  is the number of values of  $D_i$

The information gained by partitioning the training examples  $S$  into subset using the attribute  $D_1$  is given by

$$\text{Gain}(D_i) = \text{Info}(S) - \text{Info}_{D_i}(S)$$

## *The Information Theory: The Gain Ratio*

The attribute to be selected is the attribute with maximum gain value. Quinlan found out that a key attribute will have the maximum gain. This is not good!

$$\textit{Split\_Info}(S) = -\sum_{i=1}^k (|S_i| / |S|) * \log_2(|S_i| / |S|)$$

The gain ratio is given by:

$$\textit{Gain\_Ratio}(D_i) = \textit{Gain}(D_i) / \textit{Split\_Info}(D_i)$$

# The Information Theory: The Gain Ratio

Example Cont.

$$Info_{D_1}(S) = \left(\frac{5}{14}\right) * Info_{D_1}(S_1) + \left(\frac{4}{14}\right) * Info_{D_1}(S_2) + \left(\frac{5}{14}\right) * Info_{D_1}(S_3) = 0.694$$

$$Gain(D_1) = 0.94 - 0.694 = 0.246$$

$$Split\_Info(S) = -5/14 * \log_2(5/14) - 4/14 * \log_2(4/14) - 5/14 \log_2(5/14) = 1.577 \text{ bits}$$

$$Gain\_Ratio(D_1) = 0.246 / 1.577 = 0.156$$

S

D1	D2	D3	D4	D5
1	2	1	A	1
1	2	2	A	0
1	2	2	B	0
1	2	2	B	0
1	1	1	B	1
2	2	2	A	1
2	2	2	B	1
2	1	1	A	1
2	2	1	B	1
3	1	2	A	0
3	1	1	A	0
3	2	2	B	1
3	1	2	B	1
3	1	2	B	1

# Information Gain: Term vs. Category

It measures the classification power of a term

$$IG(t_k, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} P(t, c) \log_2 \frac{P(t, c)}{P(t)P(c)}$$

$P(t_k, c_i)$  → probability document x contains term t and belongs to category c.

$P(\bar{t}_k, c_i)$  → probability document x does not contain term t and belongs to category c.

$P(t_k, \bar{c}_i)$  → probability document x contains term t and does not belong to category c.

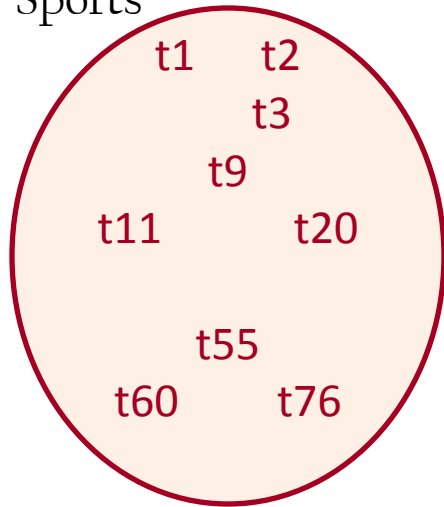
$P(\bar{t}_k, \bar{c}_i)$  → probability document x does not contain term t and does not belong to category c.

$P(t)$  → probability of term t.

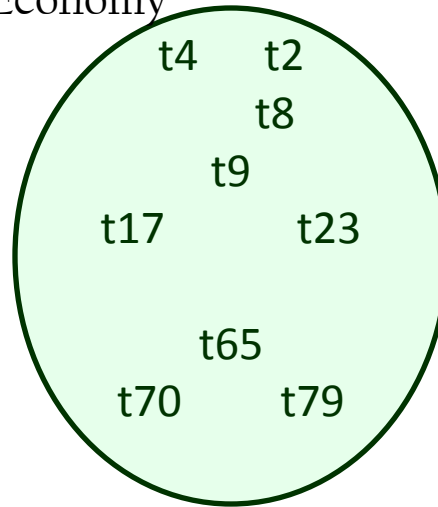
$P(c)$  → probability of category c.

# Testing The Membership

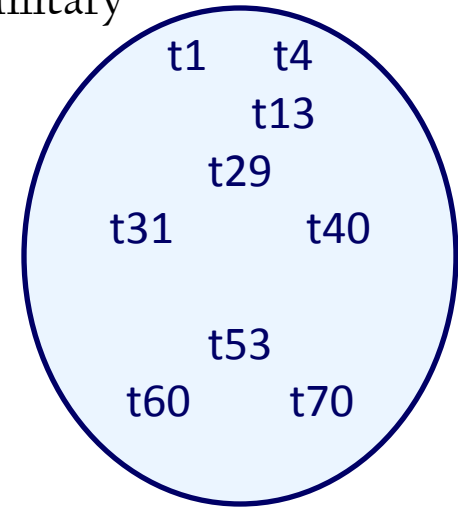
Sports



Economy



Military



$$IG(t_k, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} P(t, c) \log_2 \frac{P(t, c)}{P(t)P(c)}$$

$$IG(t_1, sport) = \frac{1}{9} * \log_2 \frac{1/9}{(2/27) * (9/27)} + \frac{8}{9} * \log_2 \frac{8/9}{(25/27) * (9/27)}$$

$$+ \frac{1}{18} * \log_2 \frac{1/18}{(2/27) * (18/27)} + \frac{17}{27} * \log_2 \frac{17/27}{(25/27) * (18/27)}$$



# The Gain Ratio

$$GR(t_k, c_i) = \frac{\sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} P(t, c) \log_2 \frac{P(t, c)}{P(t)P(c)}}{- \sum_{c \in \{c_i, \bar{c}_i\}} P(c) \log_2 P(c)}$$

$P(t_k, c_i)$  → probability document  $x$  contains term  $t$  and belongs to category  $c$ .

$P(\bar{t}_k, c_i)$  → probability document  $x$  does not contain term  $t$  and belongs to category  $c$ .

$P(t_k, \bar{c}_i)$  → probability document  $x$  contains term  $t$  and does not belong to category  $c$ .

$P(\bar{t}_k, \bar{c}_i)$  → probability document  $x$  does not contain term  $t$  and does not belong to category  $c$ .

$P(t)$  → probability of term  $t$ .

$P(c)$  → probability of category  $c$ .

# *STATISTICAL ASSOCIATIONS*

## *Part II*

*Association Rules*  
*<http://giwebb.com/>*

# The Magnum Opus System

Magnum Opus - Tutorial.data

File Edit Modes Action Preferences View Help

Tutorial.data: 500 cases / 500 holdout cases / 39 values

Search for: RULES Maximum no.: 100 Maximum size: 4

Search by: LEVERAGE

	Proportion	Count
Minimum leverage:	-1.0	-2147483647
Minimum coverage:	0.0	1
Minimum support:	0.0	0

Filter out: INSIGNIFICANT

Minimum strength: 0.0  
Minimum lift: 0.0  
 Use m-estimate

Values allowed on LHS:

- Profitability99<438
- 438<=Profitability99<=931
- Profitability99>931
- Profitability98<368
- 368<=Profitability98<=754
- Profitability98>754
- Spend99<2200
- 2200<=Spend99<=4464
- Spend99>4464
- Spend98<1927
- 1927<=Spend98<=4088
- Spend98>4088
- NoVisits99<37
- 37<=NoVisits99<=69
- NoVisits99>69
- NoVisits98<33
- 33<=NoVisits98<=66
- NoVisits98>66
- Dairy<250

Values allowed on RHS:

- Profitability99<438
- 438<=Profitability99<=931
- Profitability99>931
- Profitability98<368
- 368<=Profitability98<=754
- Profitability98>754
- Spend99<2200
- 2200<=Spend99<=4464
- Spend99>4464
- Spend98<1927
- 1927<=Spend98<=4088
- Spend98>4088
- NoVisits99<37
- 37<=NoVisits99<=69
- NoVisits99>69
- NoVisits98<33
- 33<=NoVisits98<=66
- NoVisits98>66
- Dairy<250

Ready

Attributes and their values for the Tutorial database

- Profitability99: numeric 3
- Profitability98: numeric 3
- Spend99: numeric 3
- Spend98: numeric 3
- NoVisits99: numeric 3
- NoVisits98: numeric 3
- Dairy: numeric 3
- Deli: numeric 3
- Bakery: numeric 3
- Grocery: numeric 3
- SocioEconomicGroup: categorical
- Promotion1: t, f
- Promotion2: t, f

# *The Magnum Opus System: Example*

bananas  
plums, lettuce, tomatoes  
celery, confectionery  
confectionery  
apples, carrots, tomatoes, potatoes  
potatoes  
confectionery  
carrots  
confectionery  
apples, oranges, lettuce, tomatoes  
peaches, oranges, celery, potatoes, confectionery  
beans  
oranges, lettuce, carrots, tomatoes  
apples, bananas, plums, carrots, tomatoes, onions,  
confectionery  
apples, potatoes  
lettuce, peas, beans  
carrots, tomatoes  
grapes, plums, lettuce, beans, potatoes, onions  
confectionery  
confectionery  
carrots, peas, potatoes, onions, confectionery  
tomatoes  
confectionery  
carrots, potatoes  
peaches, apples, bananas  
lettuce, beans, tomatoes, potatoes, confectionery  
grapes, lettuce, tomatoes, confectionery  
oranges

oranges, lettuce, confectionery  
tomatoes  
lettuce, carrots, tomatoes, confectionery  
celery, potatoes, confectionery  
oranges, carrots, beans, potatoes  
peaches, oranges, bananas  
lettuce, carrots, tomatoes, potatoes, onions  
onions  
peaches, apples, lettuce, peas, potatoes, onions  
oranges, carrots, confectionery  
bananas  
lettuce, carrots, tomatoes, potatoes  
carrots, confectionery  
oranges, plums  
peaches, oranges, lettuce, peas  
lettuce, carrots, beans, tomatoes  
plums, lettuce, peas, tomatoes, potatoes  
carrots, tomatoes  
bananas, lettuce, onions, confectionery  
oranges, tomatoes  
oranges, potatoes  
confectionery  
oranges, plums, potatoes  
bananas, lettuce, carrots, tomatoes, potatoes  
potatoes  
lettuce, tomatoes, onions  
lettuce, onions  
apples, oranges, beans  
corn

# The Magnum Opus System

carrots -> tomatoes

[Coverage=0.175 (175); Support=0.085 (85);  
Strength=0.486; Lift=1.85; Leverage=0.0390 (39.0);  
p=1.83E-012]

bananas -> peaches

[Coverage=0.127 (127); Support=0.040 (40);  
Strength=0.315; Lift=2.42; Leverage=0.0235 (23.5);  
p=2.74E-009]

carrots -> potatoes

[Coverage=0.175 (175); Support=0.068 (68);  
Strength=0.389; Lift=1.37; Leverage=0.0185 (18.5);  
p=0.000575]

apples -> peaches

[Coverage=0.221 (221); Support=0.044 (44);  
Strength=0.199; Lift=1.53; Leverage=0.0153 (15.3);  
p=0.000635]

bananas & apples -> peaches

[Coverage=0.029 (29); Support=0.017 (17); Strength=0.586; Lift=4.51; Leverage=0.0132 (13.2); p=0.000540]

apples -> lettuce

[Coverage=0.221 (221); Support=0.058 (58); Strength=0.262; Lift=1.21; Leverage=0.0100 (10.0); p=0.0404]

carrots & beans -> potatoes

[Coverage=0.010 (10); Support=0.007 (7); Strength=0.700; Lift=2.47; Leverage=0.0042 (4.2); p=0.0420]

The screenshot shows the 'Magnum Opus Demo - Tutorial.itl' window. The interface includes a menu bar (File, Edit, Modes, Action, Preferences, View, Help) and a toolbar with icons for file operations and search. The main area displays search parameters for 'RULES' with a maximum of 100 items and a size of 4. Search criteria are set to 'LEVERAGE' and 'INSIGNIFICANT'. A table shows search results with columns for 'Proportion' and 'Count'. Below the table are two lists: 'Values allowed on LHS' and 'Values allowed on RHS', both containing a list of fruits and vegetables.

Search for:	Maximum no.:	Maximum size:
RULES	100	4

Search by:	Minimum leverage:	Count	Minimum strength:
LEVERAGE	-1.0	-2147483647	0.0
Filter out:	Minimum coverage:	Minimum lift:	0.0
INSIGNIFICANT	0.0	1	0.0
	Minimum support:	0	<input type="checkbox"/> Use m-estimate
	0.0	0	

Values allowed on LHS:

- apples
- bananas
- beans
- carrots
- celery
- confectionery
- corn
- grapes
- lettuce
- onions
- oranges
- peaches
- peas
- plums
- potatoes
- tomatoes

Values allowed on RHS:

- apples
- bananas
- beans
- carrots
- celery
- confectionery
- corn
- grapes
- lettuce
- onions
- oranges
- peaches
- peas
- plums
- potatoes
- tomatoes

For Help, press F1

# *The Magnum Opus System: Example*

ID001, bananas  
ID002, plums  
ID002, lettuce  
ID002, tomatoes  
ID003, celery  
ID003, confectionery  
ID004, confectionery  
ID005, apples  
ID005, carrots  
ID005, tomatoes  
ID005, potatoes  
ID006, potatoes  
ID007, confectionery  
ID008, carrots  
ID009, confectionery  
ID00a, apples  
ID00a, oranges  
ID00a, lettuce  
ID00a, tomatoes  
ID00b, peaches  
ID00b, oranges  
ID00b, celery  
ID00b, potatoes  
ID00b, confectionery  
ID00c, beans  
ID00d, oranges  
ID00d, lettuce  
ID00d, carrots  
ID00d, tomatoes

ID00e, apples  
ID00e, bananas  
ID00e, plums  
ID00e, carrots  
ID00e, tomatoes  
ID00e, onions  
ID00e, confectionery  
ID00f, apples  
ID00f, potatoes  
ID010, lettuce  
ID010, peas  
ID010, beans  
ID011, carrots  
ID011, tomatoes  
ID012, grapes  
ID012, plums  
ID012, lettuce  
ID012, beans  
ID012, potatoes  
ID012, onions  
ID013, confectionery  
ID014, confectionery  
ID015, carrots  
ID015, peas  
ID015, potatoes  
ID015, onions  
ID015, confectionery  
ID016, tomatoes  
ID017, confectionery

# The Magnum Opus System

carrots -> tomatoes

[Coverage=0.175 (175); Support=0.085 (85);  
Strength=0.486; Lift=1.85; Leverage=0.0390 (39.0);  
p=1.83E-012]

bananas -> peaches

[Coverage=0.127 (127); Support=0.040 (40);  
Strength=0.315; Lift=2.42; Leverage=0.0235 (23.5);  
p=2.74E-009]

carrots -> potatoes

[Coverage=0.175 (175); Support=0.068 (68);  
Strength=0.389; Lift=1.37; Leverage=0.0185 (18.5);  
p=0.000575]

apples -> peaches

[Coverage=0.221 (221); Support=0.044 (44);  
Strength=0.199; Lift=1.53; Leverage=0.0153 (15.3);  
p=0.000635]

bananas & apples -> peaches

[Coverage=0.029 (29); Support=0.017 (17); Strength=0.586; Lift=4.51; Leverage=0.0132 (13.2); p=0.000540]

apples -> lettuce

[Coverage=0.221 (221); Support=0.058 (58); Strength=0.262; Lift=1.21; Leverage=0.0100 (10.0); p=0.0404]

carrots & beans -> potatoes

[Coverage=0.010 (10); Support=0.007 (7); Strength=0.700; Lift=2.47; Leverage=0.0042 (4.2); p=0.0420]

The screenshot shows the 'Magnum Opus Demo - Tutorial.idi' window. The title bar includes standard window controls. The menu bar contains 'File', 'Edit', 'Modes', 'Action', 'Preferences', 'View', and 'Help'. Below the menu bar is a toolbar with icons for file operations and search. The main area displays search parameters: 'Search for: RULES', 'Search by: LEVERAGE', and 'Filter out: INSIGNIFICANT'. On the right, there are input fields for 'Maximum no.: 100', 'Maximum size: 4', 'Minimum leverage: -1.0', 'Minimum coverage: 0.0', and 'Minimum support: 0.0'. A table shows 'Proportion' and 'Count' for these values. Below the search parameters are two lists: 'Values allowed on LHS' and 'Values allowed on RHS'. The LHS list includes items like apples, bananas, beans, carrots, celery, confectionery, corn, grapes, lettuce, onions, oranges, peaches, peas, plums, potatoes, and tomatoes. The RHS list includes items like apples, bananas, beans, carrots, celery, confectionery, corn, grapes, lettuce, onions, oranges, peaches, peas, plums, potatoes, and tomatoes. The status bar at the bottom says 'For Help, press F1' and has a 'NUM' button.

# *The Magnum Opus System: Example*

829, 709, 5250, 6560, 70, 82, 1074, 390, 878, 1995, C, f, f  
141, 118, 722, 928, 19, 16, 15, 155, 139, 404, C, f, f  
1044, 783, 3591, 4026, 63, 61, 81, 218, 232, 2908, D2, f, t  
78, 63, 331, 336, 7, 8, 54, 68, 63, 167, D1, t, f  
511, 419, 2142, 1947, 34, 33, 59, 106, 239, 1477, C, f, f  
987, 1402, 4032, 5376, 56, 64, 891, 681, 995, 1411, C, f, f  
313, 286, 1137, 1008, 22, 18, 153, 63, 146, 762, D1, t, f  
1800, 859, 7350, 3159, 75, 81, 441, 2315, 1433, 1837, D1, f, f  
226, 126, 1034, 612, 11, 6, 351, 377, 259, 196, C, f, f  
58, 28, 343, 140, 24, 14, 24, 18, 35, 248, A, t, f  
1136, 597, 4602, 3068, 59, 59, 554, 870, 949, 2623, D1, f, f  
376, 274, 1980, 1675, 22, 25, 356, 261, 344, 792, C, f, f  
223, 172, 1656, 1400, 18, 14, 355, 430, 323, 579, C, f, f  
1808, 976, 7600, 7396, 80, 86, 501, 718, 852, 5928, C, f, f  
114, 180, 462, 1008, 14, 16, 4, 28, 27, 364, D2, f, f  
1169, 1125, 4356, 3723, 45, 51, 359, 427, 134, 2107, D1, t, f  
226, 235, 1230, 1575, 15, 15, 414, 284, 267, 418, D1, f, f  
493, 189, 2408, 1035, 28, 23, 318, 503, 344, 1083, D1, f, f  
915, 842, 4260, 5487, 71, 59, 1265, 796, 1148, 1917, C, f, t  
1263, 739, 6136, 4277, 52, 47, 903, 1060, 589, 2208, B, f, f  
668, 429, 4992, 5841, 78, 59, 988, 955, 593, 1697, B, f, f  
259, 187, 1069, 930, 12, 10, 329, 182, 76, 481, B, t, f  
1021, 778, 4118, 3127, 58, 53, 432, 467, 432, 2388, D1, f, f  
751, 425, 3159, 1896, 27, 24, 262, 147, 542, 1516, C, f, f  
1397, 929, 6210, 5162, 54, 58, 1630, 2329, 1676, 1552, C, f, t  
336, 526, 1620, 3534, 60, 57, 211, 272, 183, 939, B, f, f  
38, 52, 182, 518, 14, 14, 16, 17, 9, 131, C, f, t  
578, 869, 1960, 3555, 70, 79, 219, 185, 212, 1274, D2, f, t

Profitability99: numeric 3  
Profitability98: numeric 3  
Spend99: numeric 3  
Spend98: numeric 3  
NoVisits99: numeric 3  
NoVisits98: numeric 3  
Dairy: numeric 3  
Deli: numeric 3  
Bakery: numeric 3  
Grocery: numeric 3  
SocioEconomicGroup: categorical  
Promotion1: t, f  
Promotion2: t, f



# The Magnum Opus System

Spend98<1782 -> NoVisits98<31  
 [Coverage=0.331 (331); Support=0.277 (277);  
 Strength=0.837; Lift=2.57; Leverage=0.1694 (169.4);  
 p=1.64E-136]

Spend99<2030 -> Grocery<873  
 [Coverage=0.333 (333); Support=0.278 (278);  
 Strength=0.835; Lift=2.51; Leverage=0.1671 (167.1);  
 p=1.13E-130]

Profitability99<419 -> Grocery<873  
 [Coverage=0.333 (333); Support=0.277 (277);  
 Strength=0.832; Lift=2.50; Leverage=0.1661 (166.1);  
 p=6.14E-129]

Profitability99<419 & Spend99<2030 -> Grocery<873  
 [Coverage=0.302 (302); Support=0.265 (265);  
 Strength=0.877; Lift=2.64; Leverage=0.1644 (164.4);  
 p=2.52E-008]

Spend99<2030 -> NoVisits99<35  
 [Coverage=0.333 (333); Support=0.272 (272); Strength=0.817; Lift=2.48; Leverage=0.1624 (162.4); p=2.42E-123]

Spend98<1782 -> NoVisits99<35  
 [Coverage=0.331 (331); Support=0.271 (271); Strength=0.819; Lift=2.49; Leverage=0.1621 (162.1); p=4.58E-123]

Spend99<2030 & Spend98<1782 -> NoVisits99<35  
 [Coverage=0.259 (259); Support=0.246 (246); Strength=0.950; Lift=2.89; Leverage=0.1608 (160.8); p=7.04E-027]

Magnum Opus Demo - Tutorial.data

File Edit Modes Action Preferences View Help

Tutorial.data: 1000 cases / 0 holdout cases / 39 values

Search for: RULES Maximum no.: 100 Maximum size: 4

Search by: LEVERAGE

Filter out: INSIGNIFICANT

	Proportion	Count	
Minimum leverage:	-1.0	-2147483647	Minimum strength: 0.0
Minimum coverage:	0.0	1	Minimum lift: 0.0
Minimum support:	0.0	0	<input type="checkbox"/> Use m-estimate

Values allowed on LHS:

- Profitability99<419
- 419<=Profitability99<=897
- Profitability99>897
- Profitability98<327
- 327<=Profitability98<=713
- Profitability98>713
- Spend99<2030
- 2030<=Spend99<=4278
- Spend99>4278
- Spend98<1782
- 1782<=Spend98<=4029
- Spend98>4029
- NoVisits99<35
- 35<=NoVisits99<=68
- NoVisits99>68
- NoVisits98<31
- 31<=NoVisits98<=64
- NoVisits98>64
- Dairy<215

Values allowed on RHS:

- Profitability99<419
- 419<=Profitability99<=897
- Profitability99>897
- Profitability98<327
- 327<=Profitability98<=713
- Profitability98>713
- Spend99<2030
- 2030<=Spend99<=4278
- Spend99>4278
- Spend98<1782
- 1782<=Spend98<=4029
- Spend98>4029
- NoVisits99<35
- 35<=NoVisits99<=68
- NoVisits99>68
- NoVisits98<31
- 31<=NoVisits98<=64
- NoVisits98>64
- Dairy<215

For Help, press F1

NUM

*Statistical Association*

*Magnum Opus*

*DEMO*

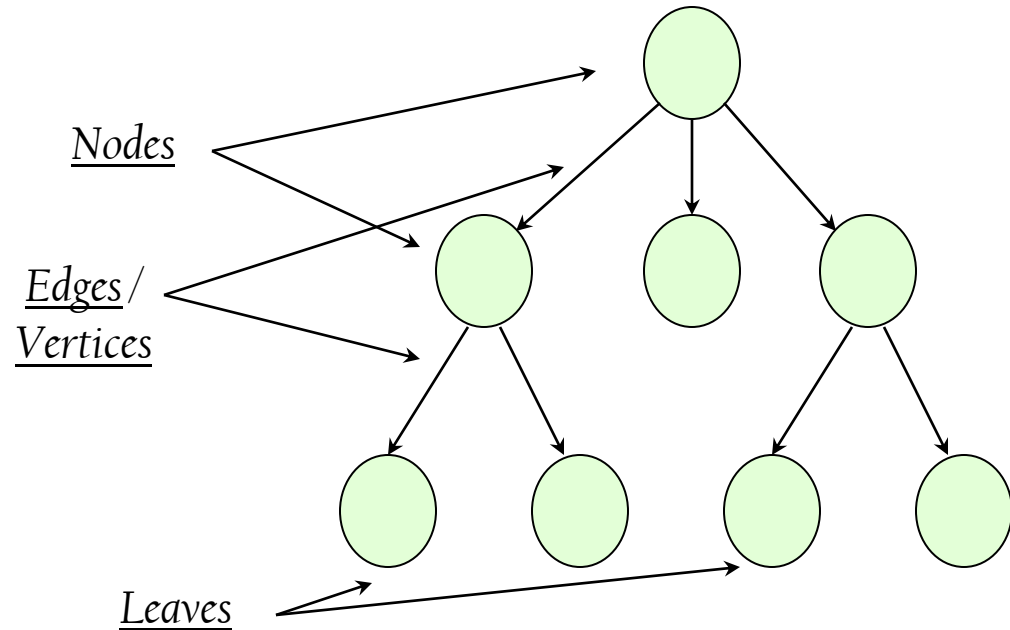
# *DECISION TREES*

## *Part 12*

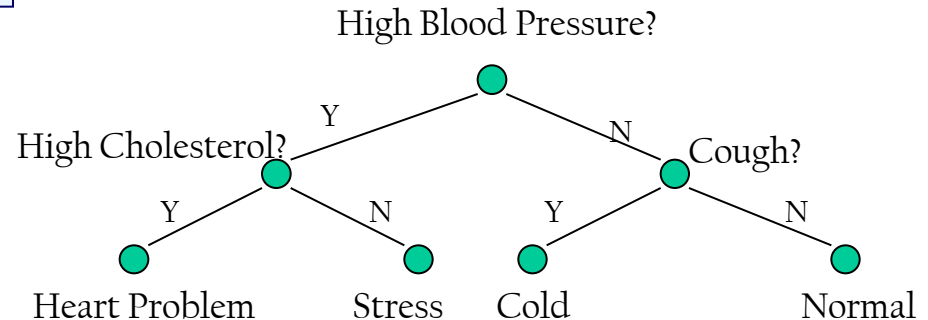
*Using Statistical &  
Information Theory  
<http://rulequest.com/>*

# Learning Decision Trees

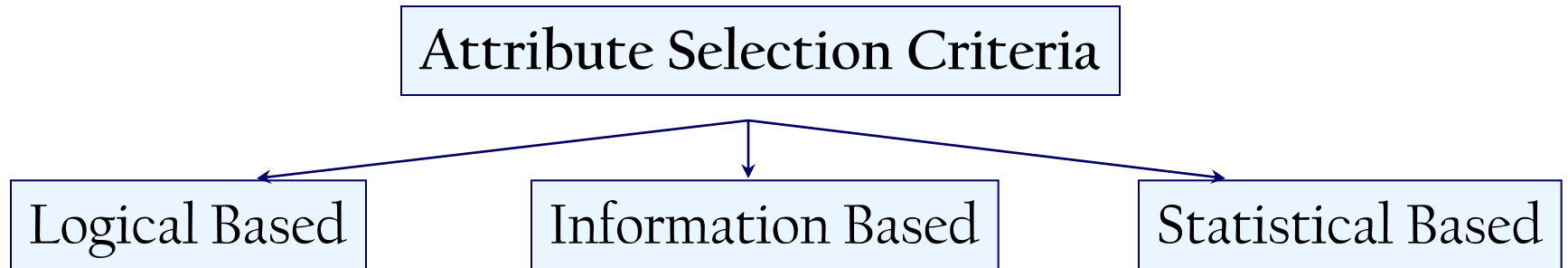
- A Tree is a **D**irected **A**cyclic **G**raph (**DAG**) + each node has one parent at most
- A Decision Tree is a tree where nodes associated with attributes, edges associated with attribute values, and leaves associated with decisions



## Example:



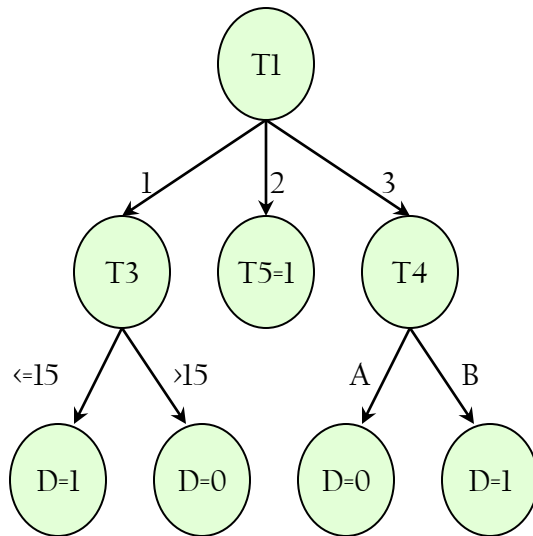
# *Learning Decision Trees*



# Information Theory

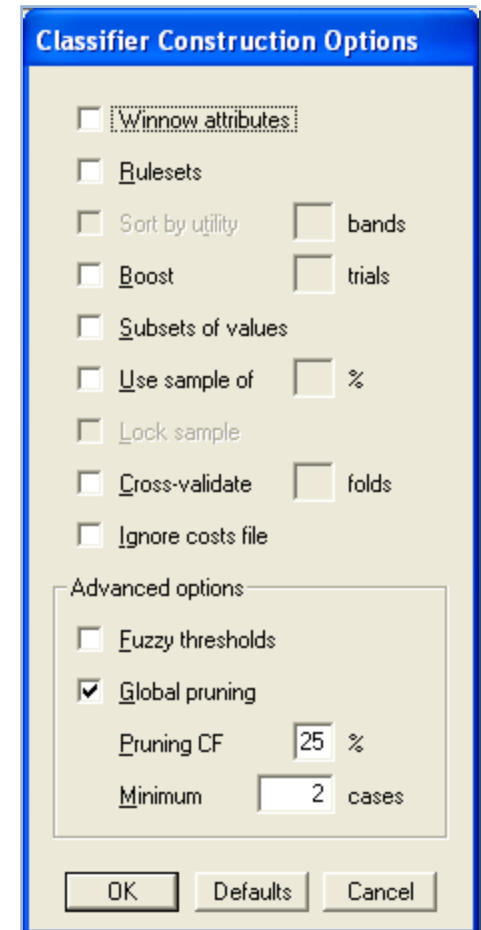
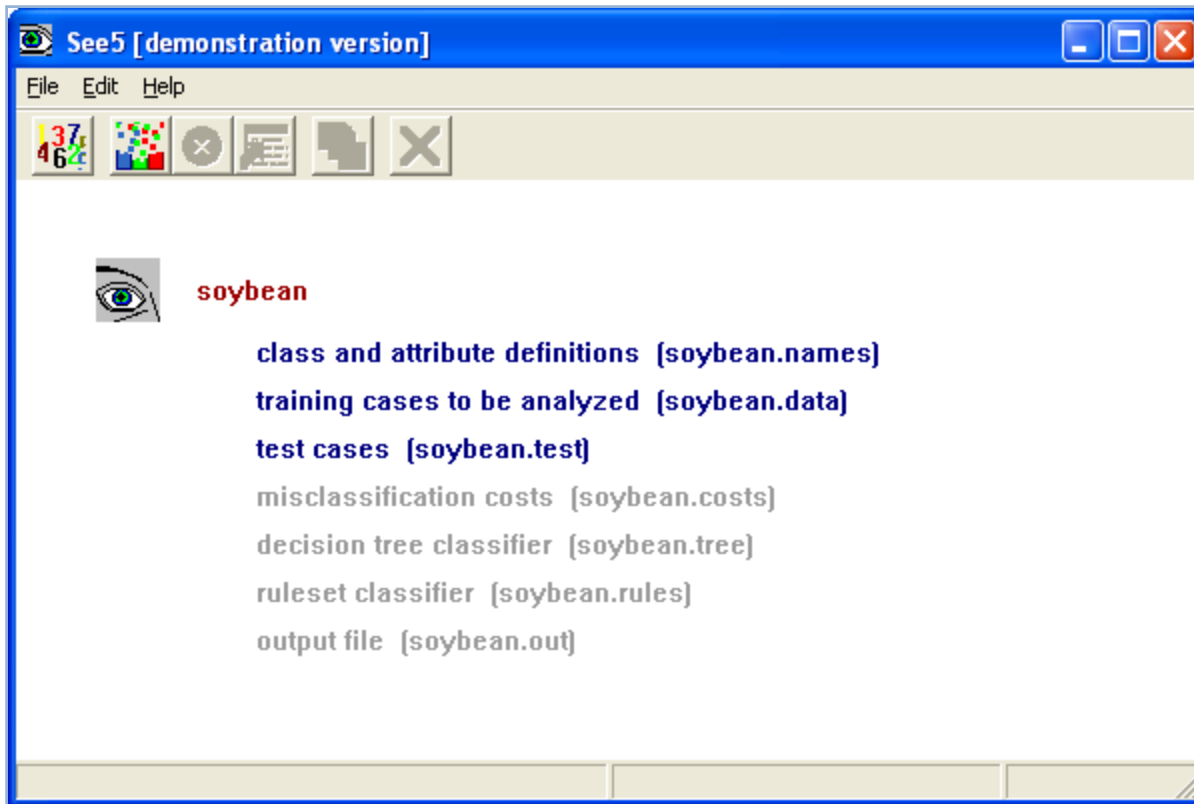
## Example

- T2 is quantized into two intervals at 21 ( $T2 \leq 21$ ) and ( $T2 > 21$ )
- T3 is quantized into two intervals at 15 ( $T3 \leq 15$ ) and ( $T3 > 15$ )



T1	T2	T3	T4	D
1	25	10	A	1
1	30	30	A	0
1	35	25	B	0
1	22	35	B	0
1	19	10	B	1
2	22	30	A	1
2	33	18	B	1
2	14	5	A	1
2	31	15	B	1
3	21	20	A	0
3	15	10	A	0
3	25	20	B	1
3	18	20	B	1
3	20	36	B	1

# C5



*Decision Trees*

*C5*

*DEMO*



# *NEURAL NETWORKS*

## *Part 13*

### *How It Works?*

# Learning Neural Networks

Supervised

Unsupervised

In terms of Design

As Learning Algorithm

In terms of Design

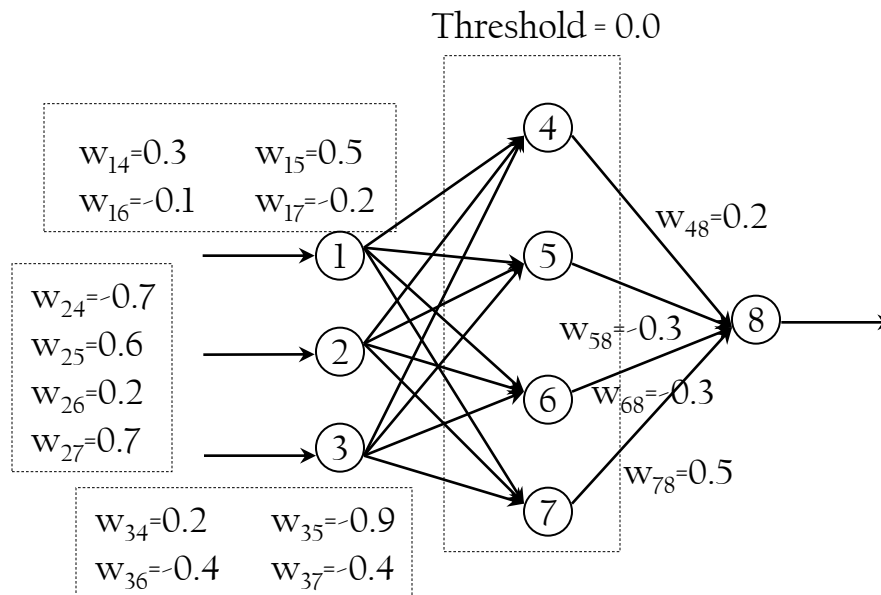
As Learning Algorithm

The user defines the number of nodes and levels in the hidden layer

The data is labeled and both input and output are given to the neural network

No. of nodes and levels in the hidden layer are defined automatically by the algorithm

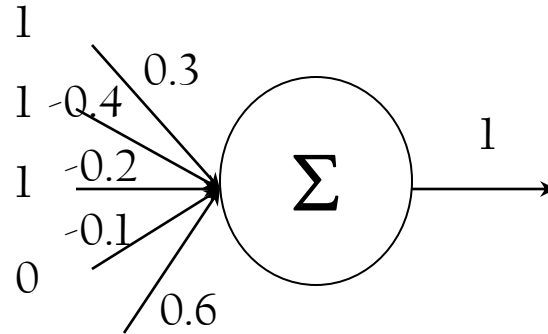
The data is not labeled. Only the input records are given to the neural network



Test Data

A	B	C	Decision
0	0	0	
0	0	1	
0	1	0	
0	1	1	1
1	0	0	
1	0	1	
1	1	0	
1	1	1	

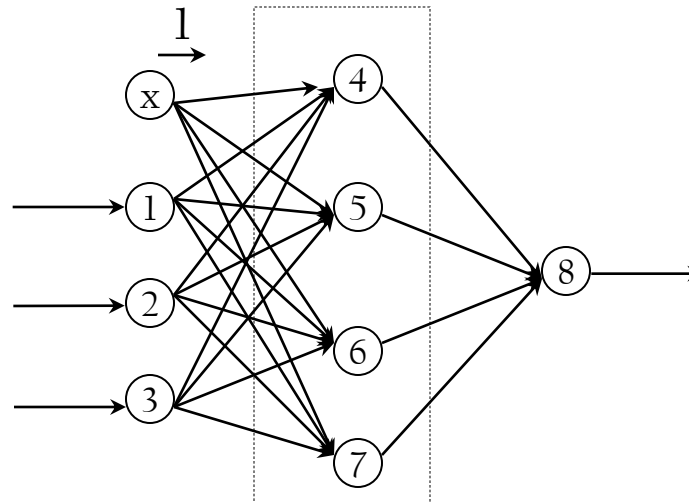
# Learning Neural Networks



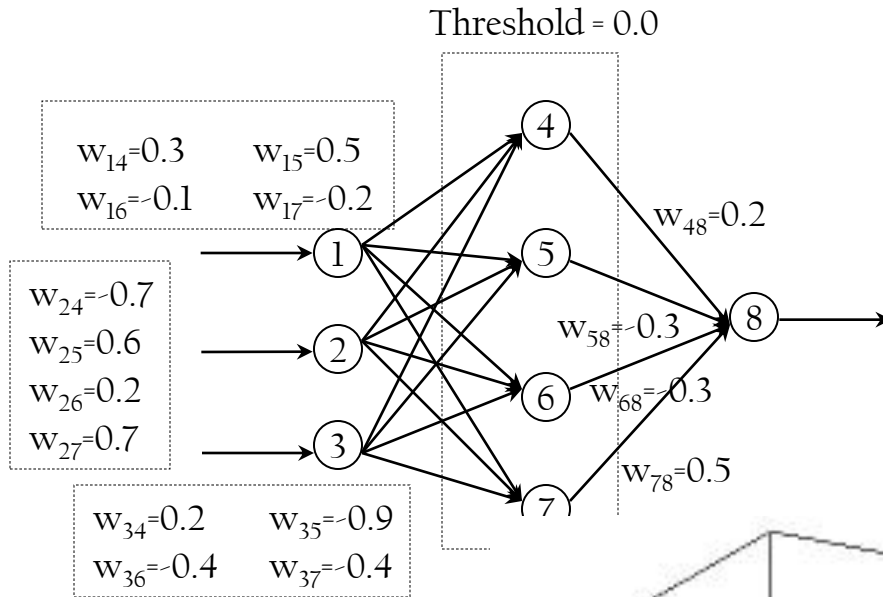
The Sigmoid Function

$$1 = 1 * 0.3 - 1 * 0.4 - 1 * 0.2 - 0 * 0.1 + 1 * 0.6 = 0.3 > 0.0$$

To avoid setting the threshold:

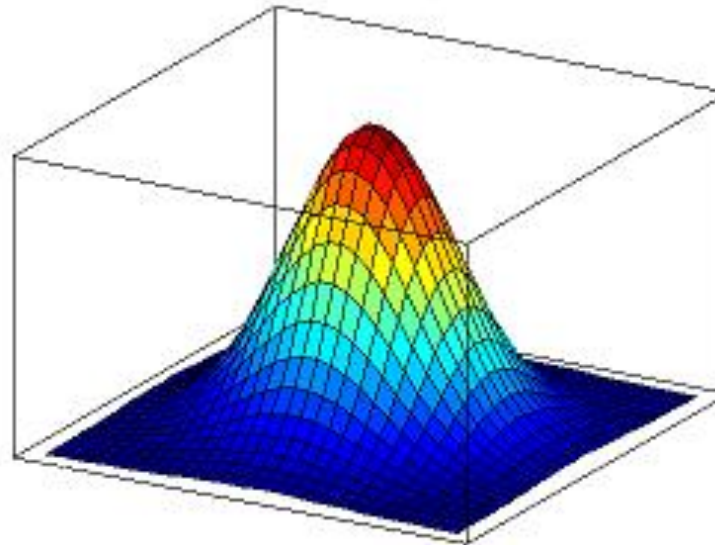


# Learning Neural Networks



Test Data

A	B	C	Decision
0	0	0	
0	0	1	
0	1	0	
0	1	1	
1	0	0	
1	0	1	
1	1	0	
1	1	1	



# *MACHINE TRANSLATION*

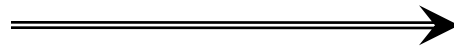
## *Part 14*

### *Statistical Machine Translation*

# Statistical Machine Translation

- For each English sentence “e”, we need the Arabic sentence “a” which maximize  $P(a|e)$   
 $P(a|e) = P(a) * P(e|a) / P(e)$

English  
Document



Arabic  
Document

# Language Model

- A statistical language model assigns a probability to a sequence of  $m$  words by means of a probability distribution
- Record every sentence that anyone ever says in Arabic; Suppose you record a database of one billion utterances; If the sentence “كيف حالك؟” appears 76,413 times in that database, then we say  $P(\text{كيف حالك؟}) = 76,413/1,000,000,000 = 0.000076413$
- One big problem is that many perfectly good sentences will be assigned a  $P(a)$  of zero

Arabic Sentence	Probability
كيف حالك	0.000076413
الولد سعيد	0.000066392

# N-Grams

- An n-word substring is called an n-gram
  - If n=2, we say bigram. If n=3, we say trigram
  - Let  $P(y | x)$  be the probability that word y follows word x
$$P(y | x) = \text{number-of-occurrences}(\text{"xy"}) / \text{number-of-occurrences}(\text{"x"})$$
$$P(z | x y) = \text{number-of-occurrences}(\text{"xyz"}) / \text{number-of-occurrences}(\text{"xy"})$$
- $P(\text{ذهب إلى المدرسة} | \text{start-of-sentence}) * P(\text{الولد} | \text{ذهب}) * P(\text{إلى} | \text{المدرسة}) * P(\text{end-of-sentence} | \text{المدرسة})$
- $P(\text{ذهب إلى المدرسة} | \text{start-of-sentence}) * P(\text{ذهب, الولد} | \text{start-of-sentence, إلى}) * P(\text{إلى, المدرسة} | \text{الولد, إلى}) * P(\text{end-of-sentence} | \text{المدرسة, إلى, end-of-sentence})$



# N-Grams Language Model

$$P(w_1, \dots, w_m) = \prod_{i=1}^m P(w_i | w_1, \dots, w_{i-1}) \approx \prod_{i=1}^m P(w_i | w_{i-(n-1)}, \dots, w_{i-1})$$

$$P(w_i | w_{i-(n-1)}, \dots, w_{i-1}) = \frac{\text{count}(w_{i-(n-1)}, \dots, w_i)}{\text{count}(w_{i-(n-1)}, \dots, w_{i-1})}$$

## Example:

In a bigram ( $n=2$ ) language model, the approximation looks like

$$P(I, \text{saw}, \text{the}, \text{red}, \text{house}) \approx P(I)P(\text{saw} | I)P(\text{the} | \text{saw})P(\text{red} | \text{the})P(\text{house} | \text{red})$$

In a trigram ( $n=3$ ) language model, the approximation looks like

$$P(I, \text{saw}, \text{the}, \text{red}, \text{house}) \approx P(I)P(\text{saw} | I)P(\text{the} | I, \text{saw})P(\text{red} | \text{saw}, \text{the})P(\text{house} | \text{the}, \text{red})$$

# Translation Model

- $P(e | a)$ , the probability of an English string “e” given an Arabic string “a”; This is called a translation model
- $P(e | a)$  will be a module in overall English-to-Arabic machine translation system; When we see an actual English string e, we want to reason backwards ... What Arabic string a is likely to be expressed, and likely to subsequently translate to e? We're looking for the a that maximizes  $P(a) * P(e | a)$

Arabic Sentence	English Sentence	$P(a e)$
ذهب الولد إلى المدرسة	The boy went to School	0.0034
إنخفاض البورصة اليوم	Today, the stock market went down	0.00021
:	:	

- Example, BuckWalter

# Translation Model

- For each word  $a_i$  in an Arabic sentence ( $i = 1 \dots l$ ), we choose a fertility  $\phi_i$ . The choice of fertility depends on the Arabic word in question. It is not dependent on the other Arabic words in the Arabic sentence, or on their fertilities
- For each word  $a_i$ , we generate  $\phi_i$  English words. The choice of English word depends on the Arabic word that generates it. It is not dependent on the Arabic context around the Arabic word. It is not dependent on other English words that have been generated from this or any other Arabic word
- All those English words are permuted. Each English word is assigned an absolute target “position slot.” For example, one word may be assigned position 3, and another word may be assigned position 2 -- the latter word would then precede the former in the final English sentence. The choice of position for a English word is dependent solely on the absolute position of the Arabic word that generates it

# REFERENCES

- W. Weaver (1955). Translation (1949). In: *Machine Translation of Languages*, MIT Press, Cambridge, MA.
- P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2), 263-311.
- S. Vogel, H. Ney and C. Tillmann. 1996. HMM-based Word Alignment in Statistical Translation. In COLING '96: The 16th International Conference on Computational Linguistics, pp. 836-841, Copenhagen, Denmark.
- F. Och and H. Ney. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19-51
- P. Koehn, F.J. Och, and D. Marcu (2003). Statistical phrase based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL)*.
- D. Chiang (2005). A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*.
- F. Och and H. Ney. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19-51
- P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. ACL 2007, Demonstration Session, Prague, Czech Republic
- Q. Gao, S. Vogel, "Parallel Implementations of Word Alignment Tool", *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pp. 49-57, June, 2008
- W. J. Hutchens and H. Somers. (1992). *An Introduction to Machine Translation*, 18.3:322. ISBN 0-12-36280-X

# REFERENCES

- W. The Sage Dictionary of Statistics, pg. 76, Duncan Cramer, Dennis Howitt, 2004, [ISBN 076194138X](#)
- E.L. Lehmann and Joseph P. Romano (2005). *Testing Statistical Hypotheses* (3E ed.). New York, NY: Springer. [ISBN 0387988645](#)
- D.R. Cox and D.V.Hinkley (1974). *Theoretical Statistics*. [ISBN 0412124293](#).
- [Fisher, Sir Ronald A.](#) (1956) [1935]. "[Mathematics of a Lady Tasting Tea](#)". in James Roy Newman. *The World of Mathematics, volume 3*.  
<http://books.google.com/books?id=oKZwtLQmNAC&pg=PA1512&dq=%22mathematics+of+a+lady+tasting+tea%22&sig=8-NQlCLzrh-oV0wjfwa0EgspSNU>
- R.A. Fisher, the Life of a Scientist, Box, 1978, p134
- McCloskey, Deirdre (2008). *The Cult of Statistical Significance*. Ann Arbor: University of Michigan Press. [ISBN 0472050079](#)
- *What If There Were No Significance Tests?*, Harlow, Mulaik & Steiger, 1997, [ISBN 978-0-8058-2634-0](#)
- Rosnow, R.L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44, 1276-1284
- Loftus, G.R. 1991. On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology* 36: 102-105
- [Cohen, J.](#) 1990. Things I have learned (so far). *American Psychologist* 45: 1304-1312. ^ Introductory Statistics, Fifth Edition, 1999, pg. 521, Neil A. Weiss, [ISBN 0-201-59877-9](#)
- Ioannidis JP (July 2005). "Contradicted and initially stronger effects in highly cited clinical research". *JAMA* 294 (2): 218–28.