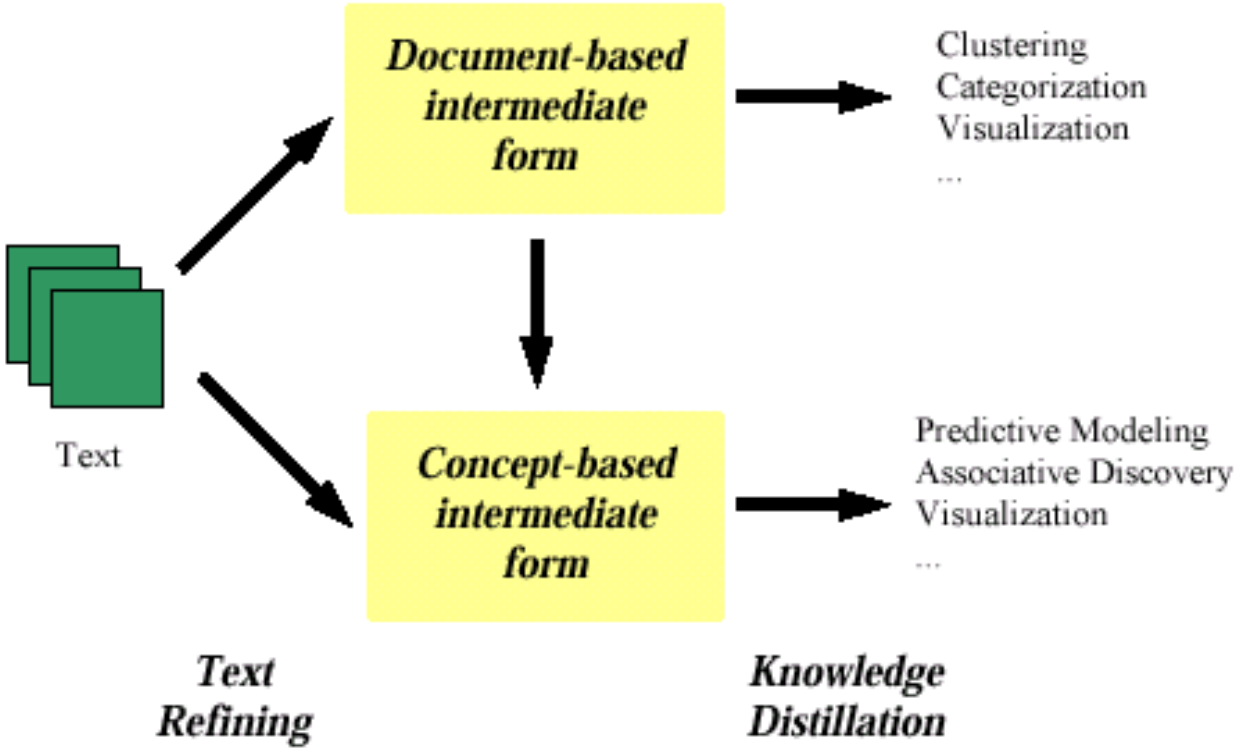
# Advanced IR Research Topics

Text Mining(TM) and Information Extraction(IE) from Scientific Texts

# General Text Mining Framework



# Automatic Text Categorization

- Categorization problems:
  - On-line documents categorization
    - Web pages categorization for search engines
    - E-mail, News group, BBS filtering
  - Information extraction
    - Data extraction for internet agents
    - Keyword, key phrase extraction
    - Summarization
  - Learning about users, detecting intrusion, etc.
- Labor-intensive, in need of solutions from AI

# Text Categorization

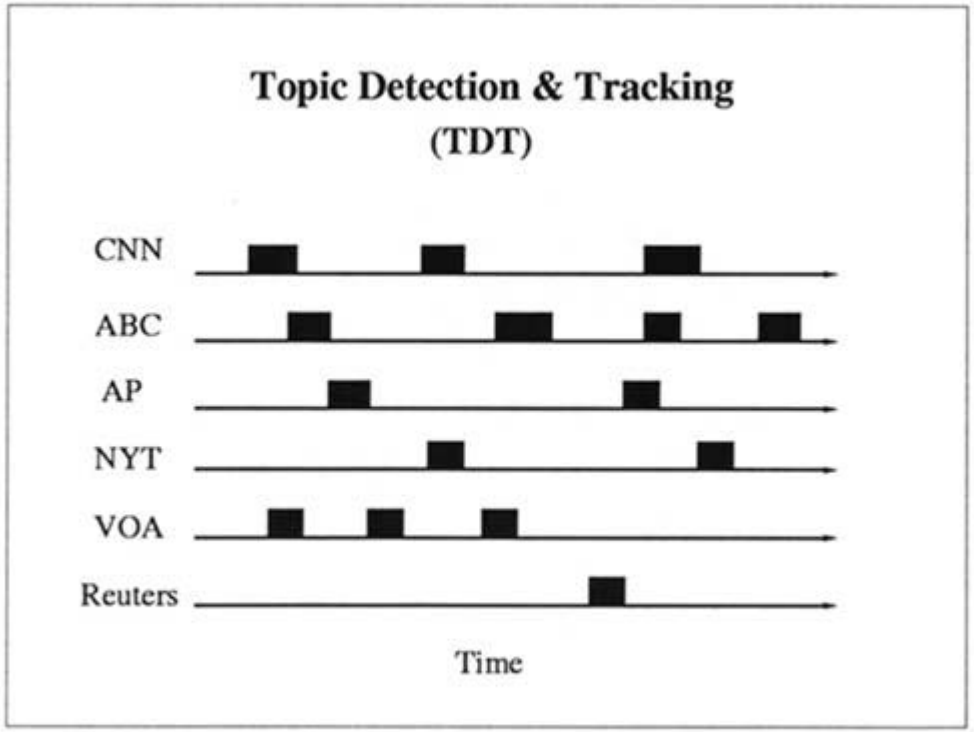
- Categorizing web pages
- assigning Computing Reviews categories to abstracts of articles
- analyzing customer data
- detecting credit card frauds
- routing e-mail questions to staff members
- cataloging e-book for personal use

# Available Techniques for Text Categorization

- K-NN
- Regression Models
- Expert Systems
- Decision Trees
- Rule Induction in FOL
- Support Vector Machines
- Neural Networks
- ...

# Topic Detection and Tracking (TDT)

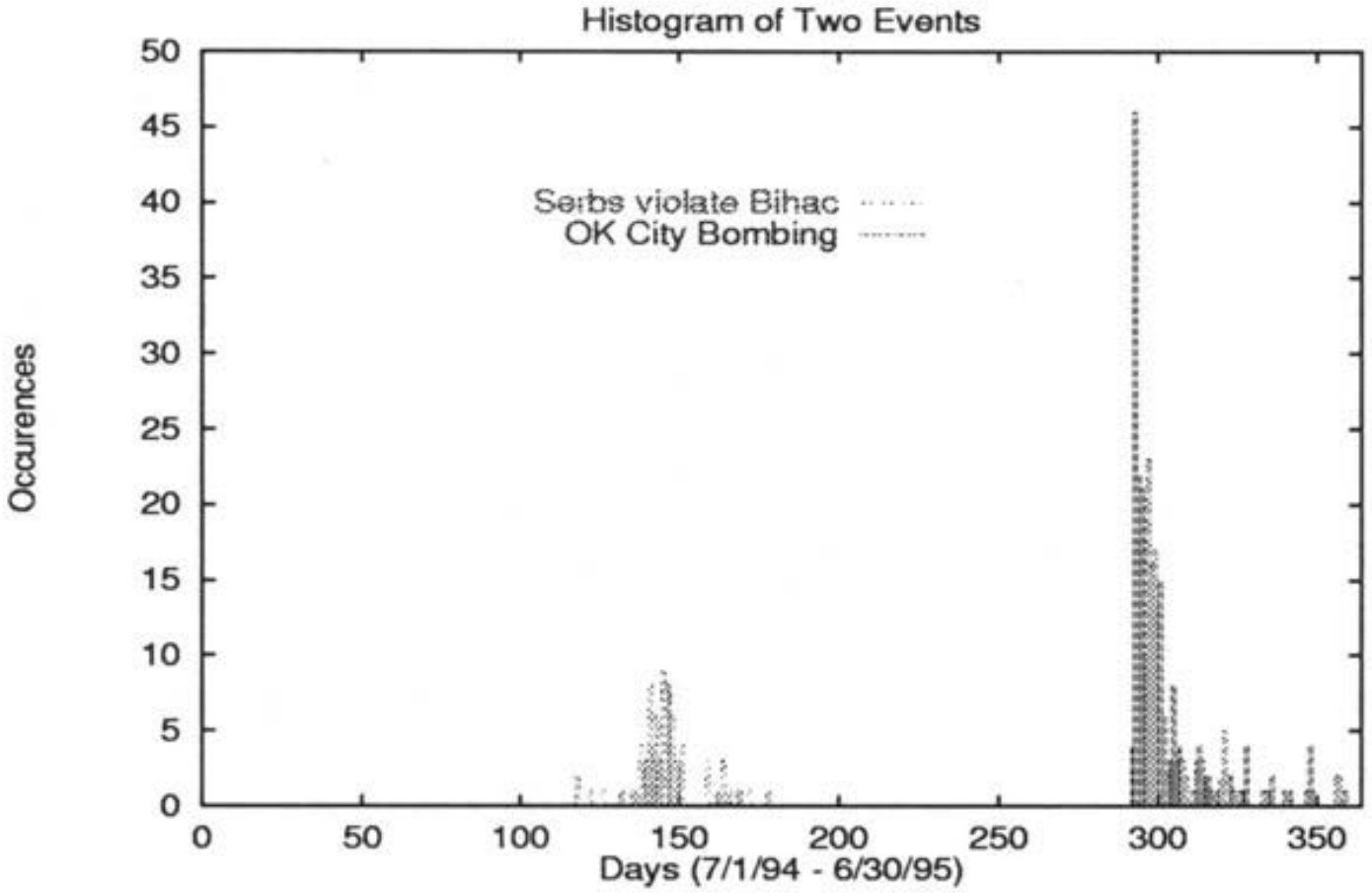
- Dealing with live data (text/video/audio)



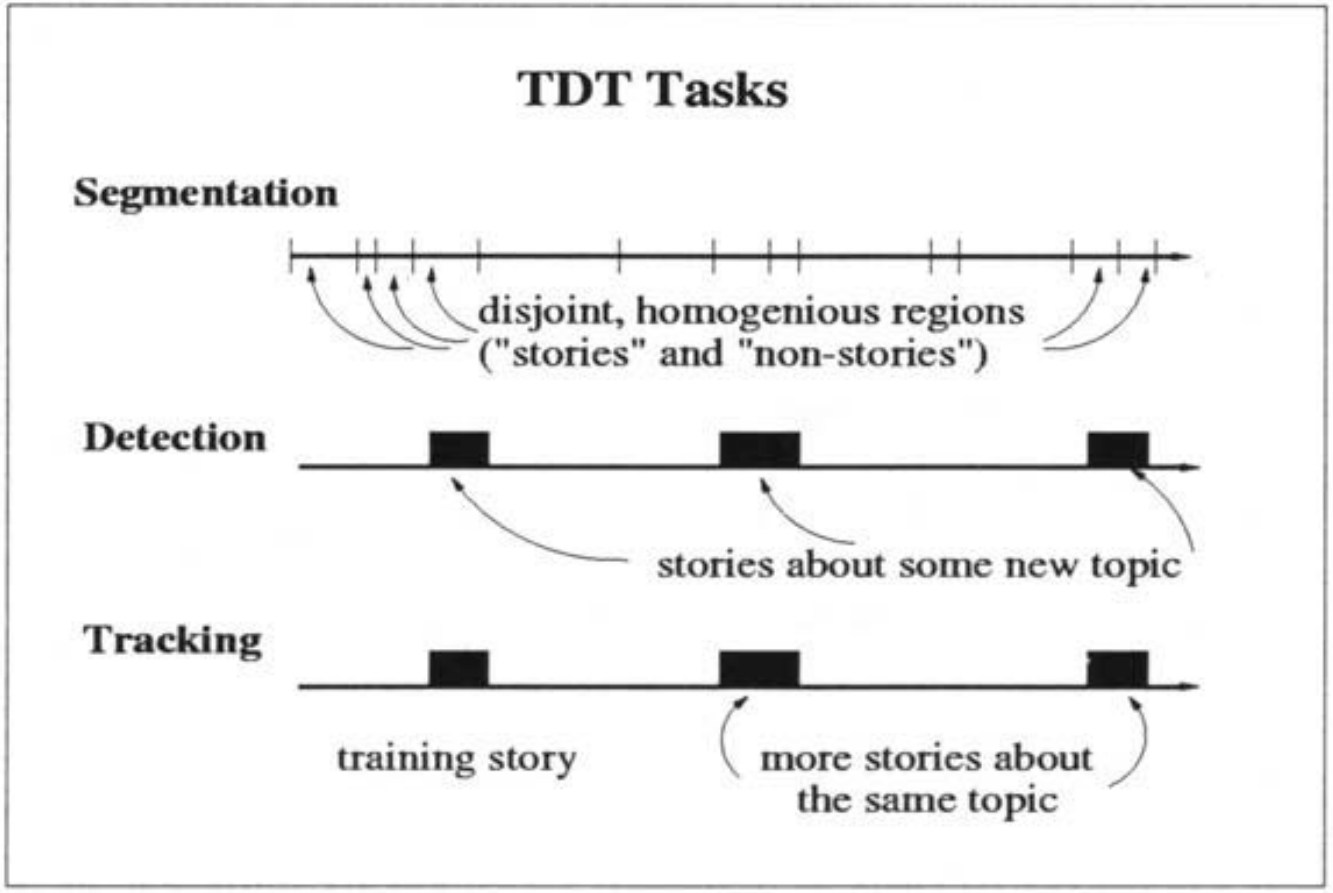
# Topic Detection and Tracking

- News streams from multiple sources (TV, radio, and newswires)
- Stories arrive in chronological order
- News bursts -- i.e. important events
- Events are typically short lasting in duration
- ...

# Topic Detection and Tracking



# Topic Detection and Tracking



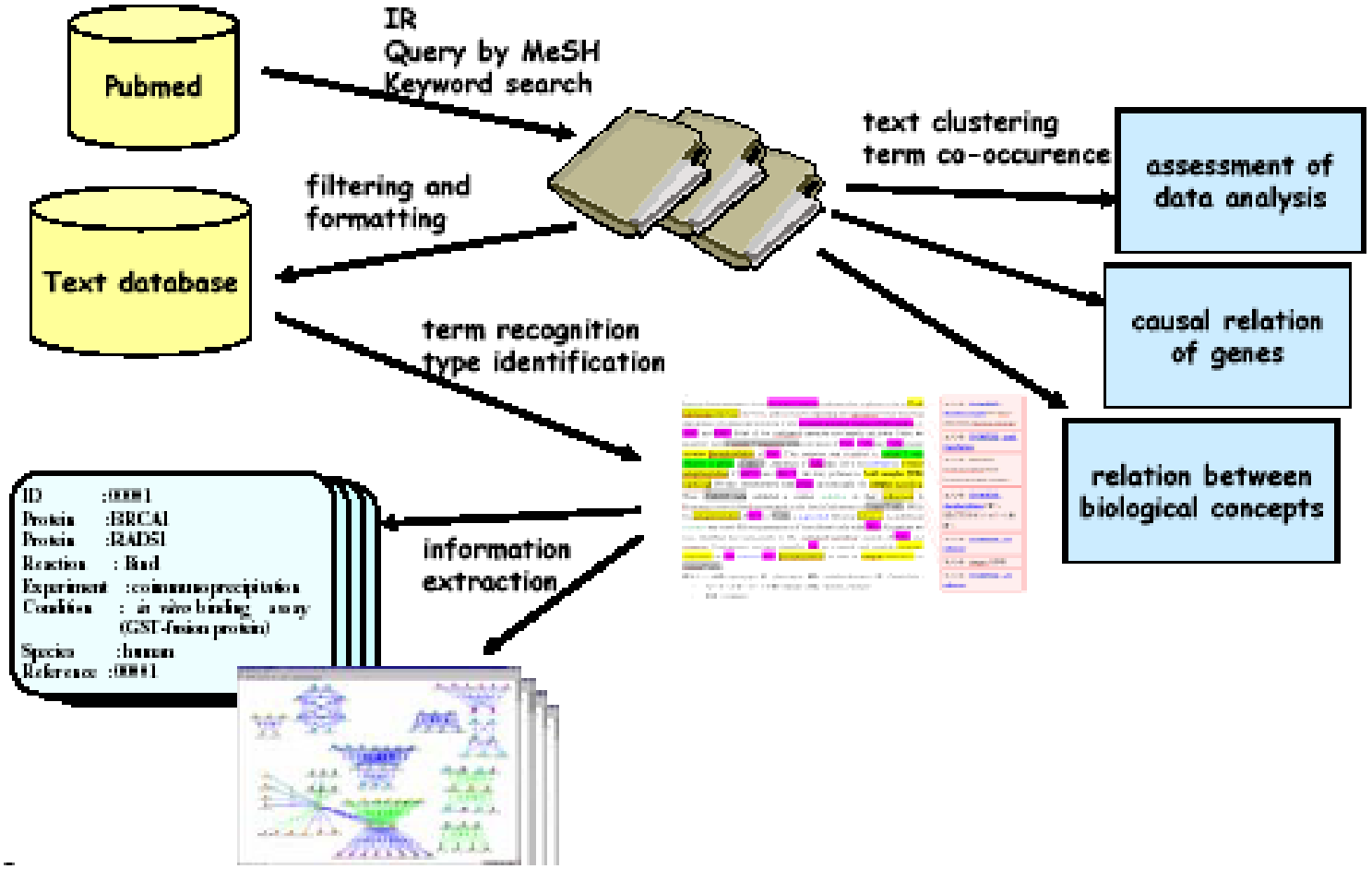
# New Information Detection

- New information:
  - information which we have not seen before in any stories on this topic
- Topics:
  - specific events with all relevant information
- Extract new information from stories as they arrive
- Filtering news stream on a particular topic

# New Information Detection

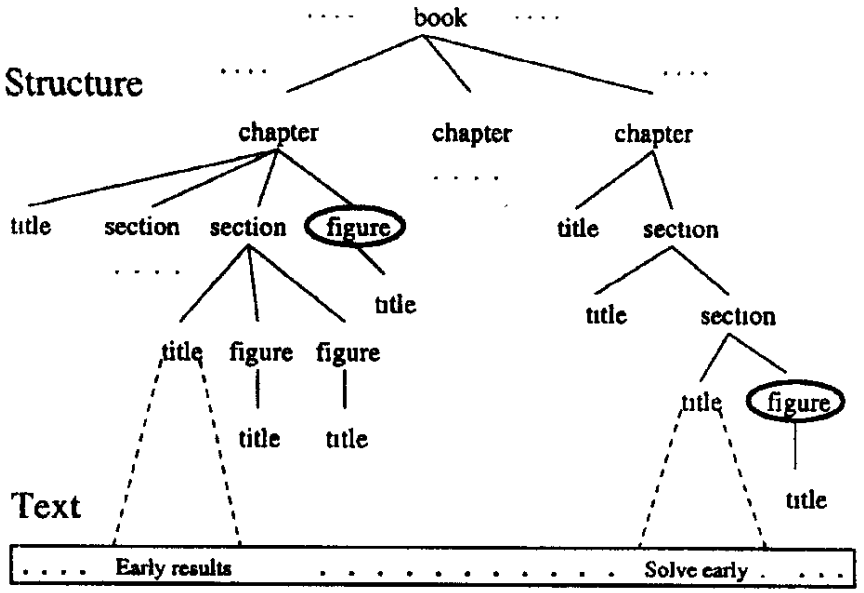
- Niche: news reporting is redundant!!
- News summarization
- New info. / Interesting info.
- Sentence-based approach
- Cluster Analysis approach
  - augmenting the sentence with related concepts

# Overview of Text Processing in Biology



# Text Databases

- Data Model
  - Text
  - Structure



# Text Databases

- Examples
  - Discovering relationships between features
  - the Company-Person name recognizer
  - Words Windows

# Application Tasks of NLP

(1)Information Retrieval/Detection

To search and retrieve documents in response to queries for information

(2)Passage Retrieval

To search and retrieve part of documents in response to queries for information

(3)Information Extraction

To extract information that fits pre-defined database schemas or templates, specifying the output formats

(4) Question/Answering Tasks

To answer general questions by using texts as knowledge base: Fact retrieval, combination of IR and IE

(5)Text Understanding

To understand texts as people do: Artificial Intelligence

# Example #1: FASTUS(1993)

Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

TIE-UP-1
Relationship: TIE-UP
Entities: "Bridgestone Sport Co."
"a local concern"
"a Japanese trading house"
Joint Venture Company:
"Bridgestone Sports Taiwan Co."
Activity: ACTIVITY-1
Amount: NT$200000000

ACTIVITY-1
Activity: PRODUCTION
Company:
"Bridgestone Sports Taiwan Co."
Product:
"iron and 'metal wood' clubs"
Start Date:
DURING: January 1990

…….

Jurgen Pfrang, 51, reportedly stumbled upon the robbers on the second floor of his Nanjing home early on Sunday.

The deputy general manager of Yaxing Benz, a Sino-German joint venture that makes buses and bus chassis in nearby Yangzhou, was hacked to death with 45 cm watermelon knives.

…….

Name of the Venture: Yaxing Benz
Products: buses and bus chassis
Location: Yangzhou,China
Companies involved: (1)Name: X?
Country: German
(2)Name: Y?
Country: China

# Information Extraction

***A German vehicle-firm executive*** was stabbed to death ….

…….

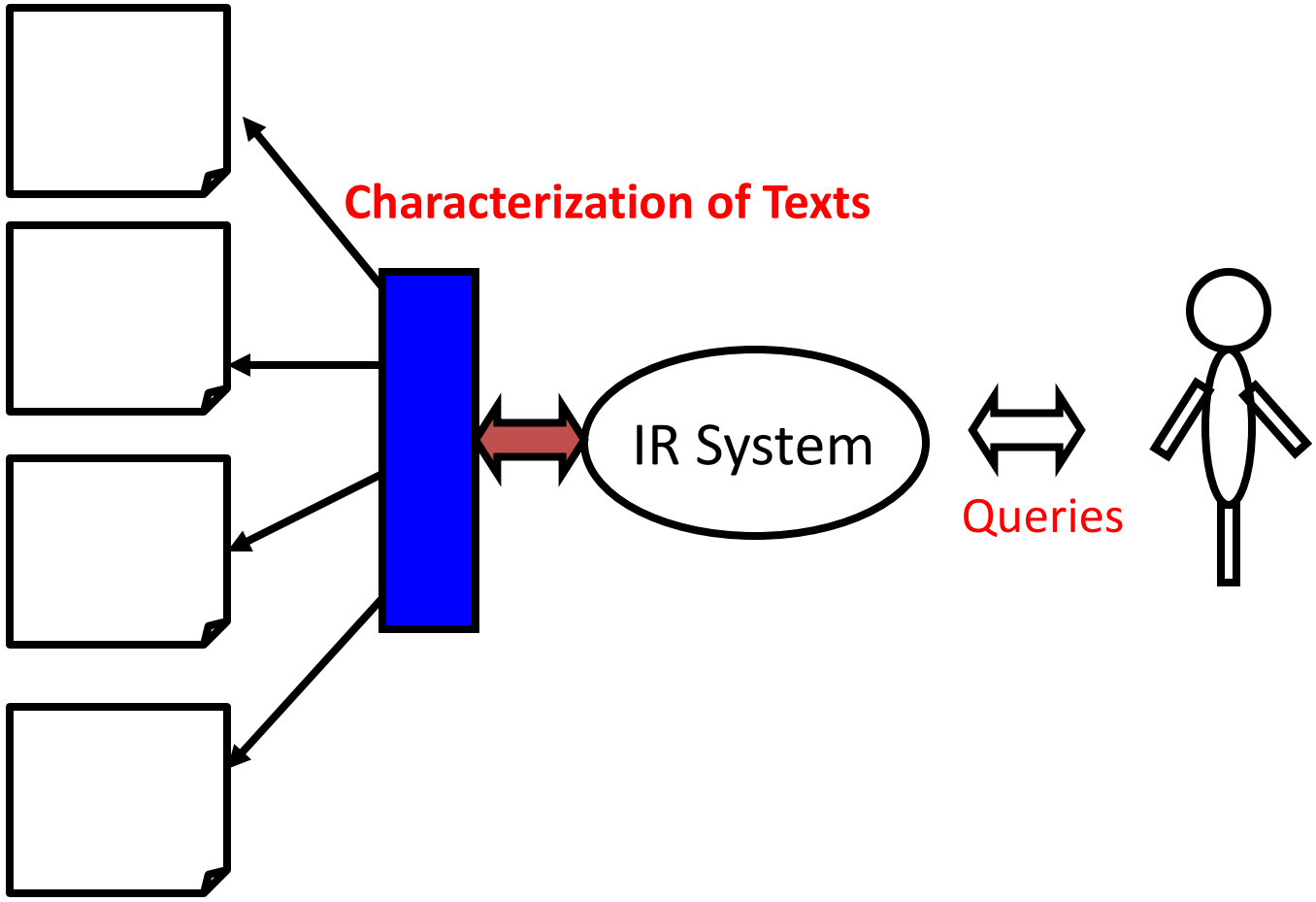
***Jurgen Pfrang***, 51, reportedly stumbled upon the robbers on the second floor of his Nanjing home early on Sunday.

***The deputy general manager*** of Yaxing Benz, a Sino-German joint venture that makes buses and bus chassis in nearby Yangzhou, was hacked to death with 45 cm watermelon knives.

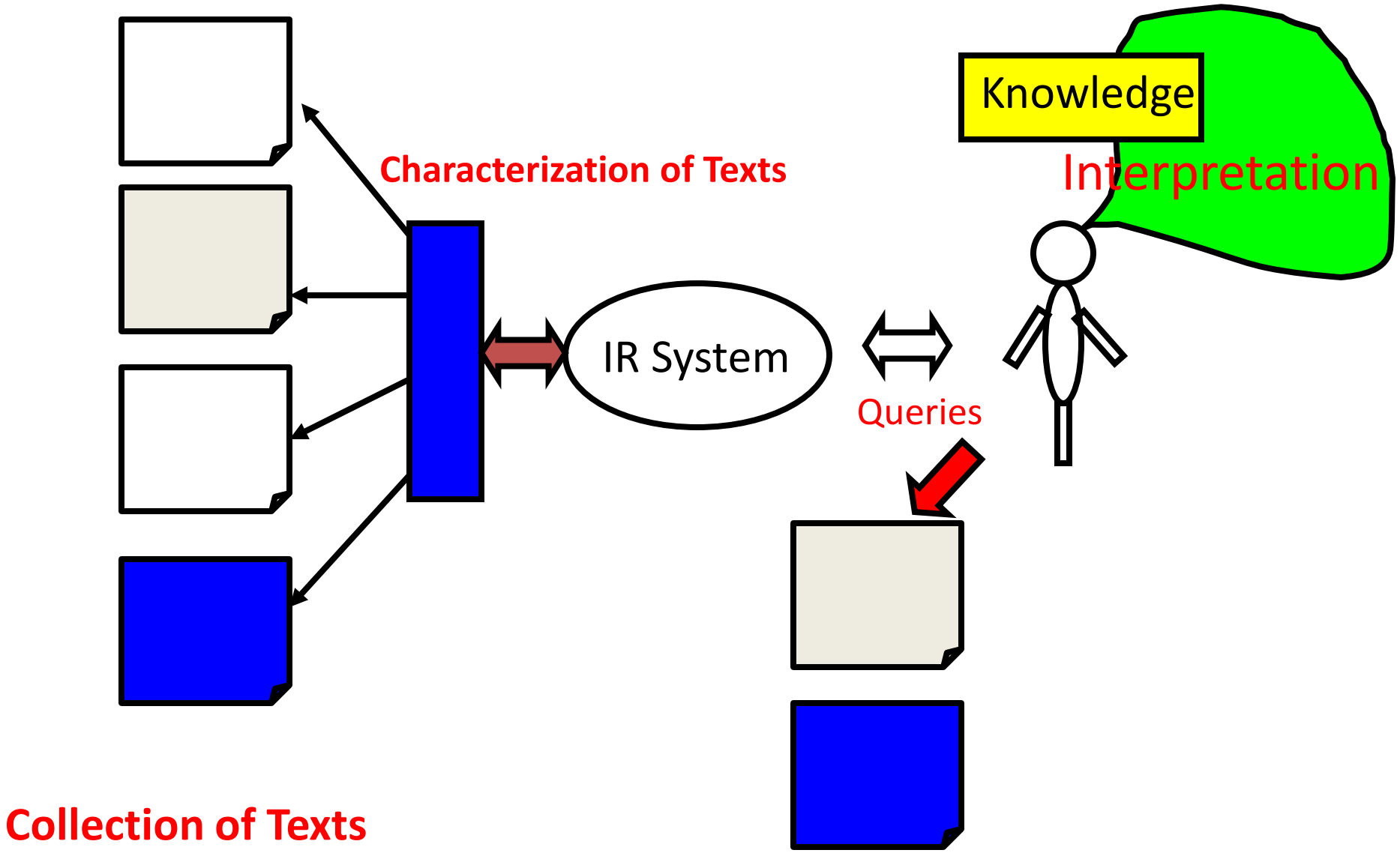
…….

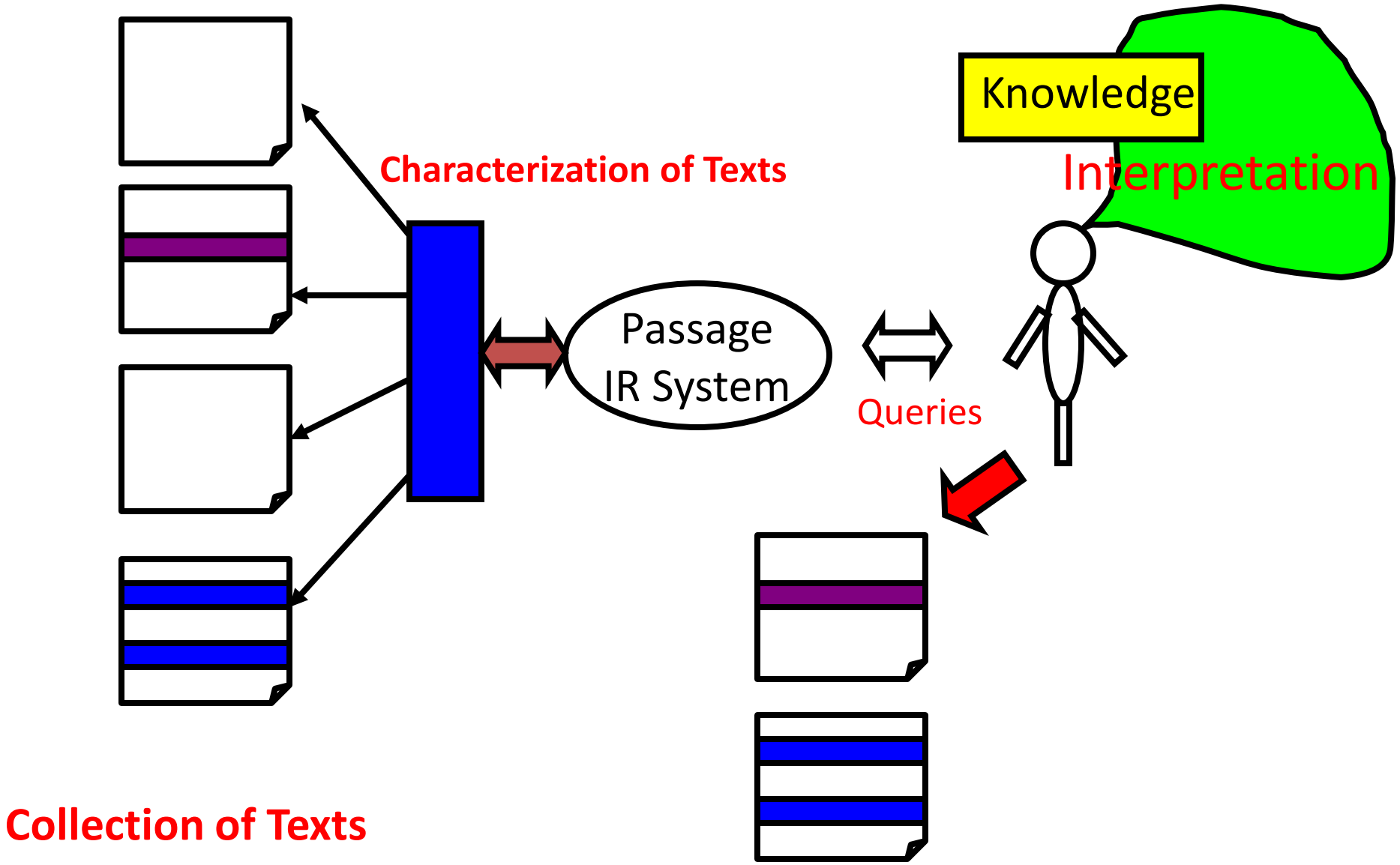
Crime-Type: Murder
Type: Stabbing
The killed: Name: Jurgen Pfrang
Age: 51
Profession: Deputy general manager
Location: Nanjing, China

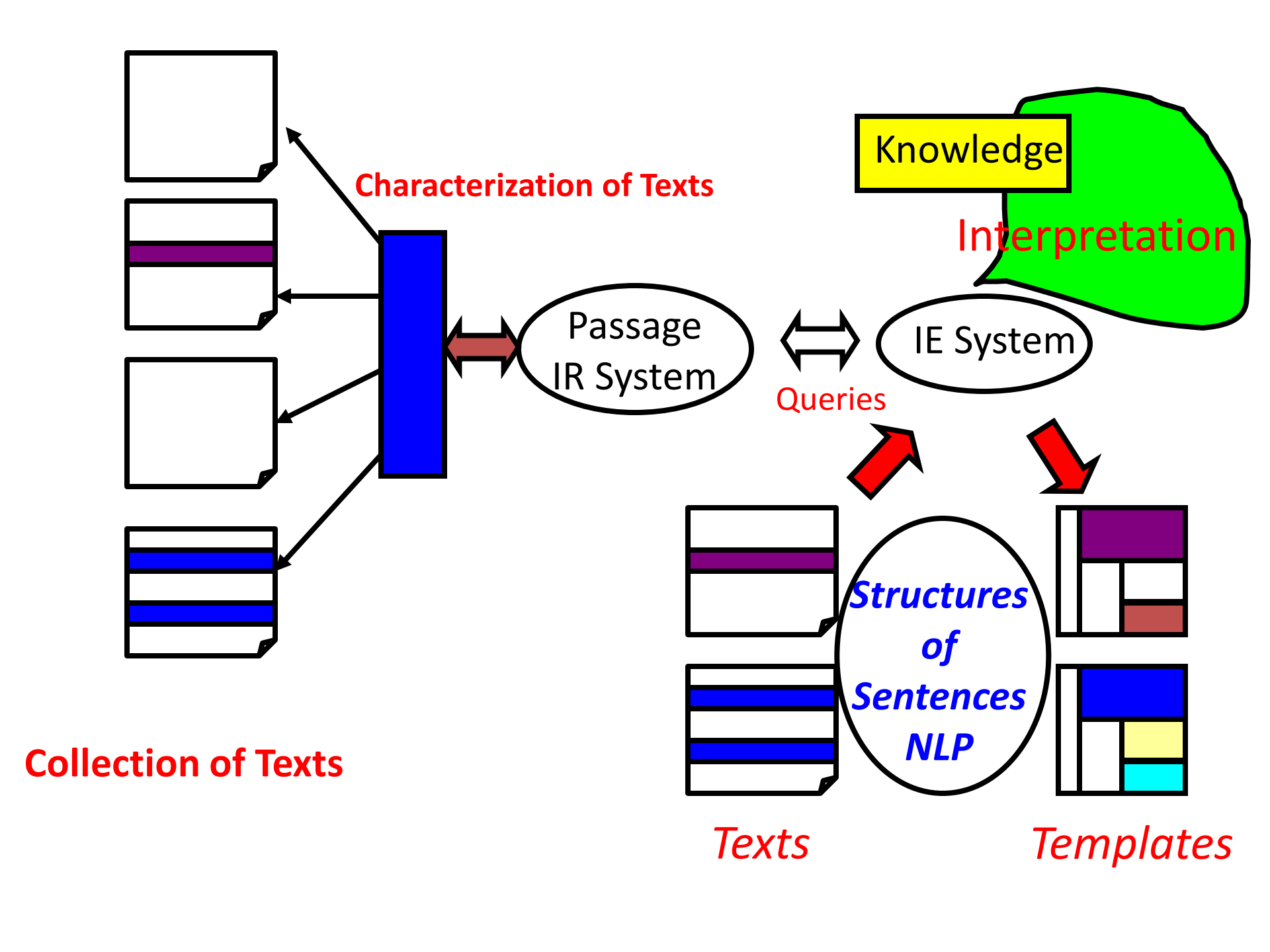
Different template for crimes

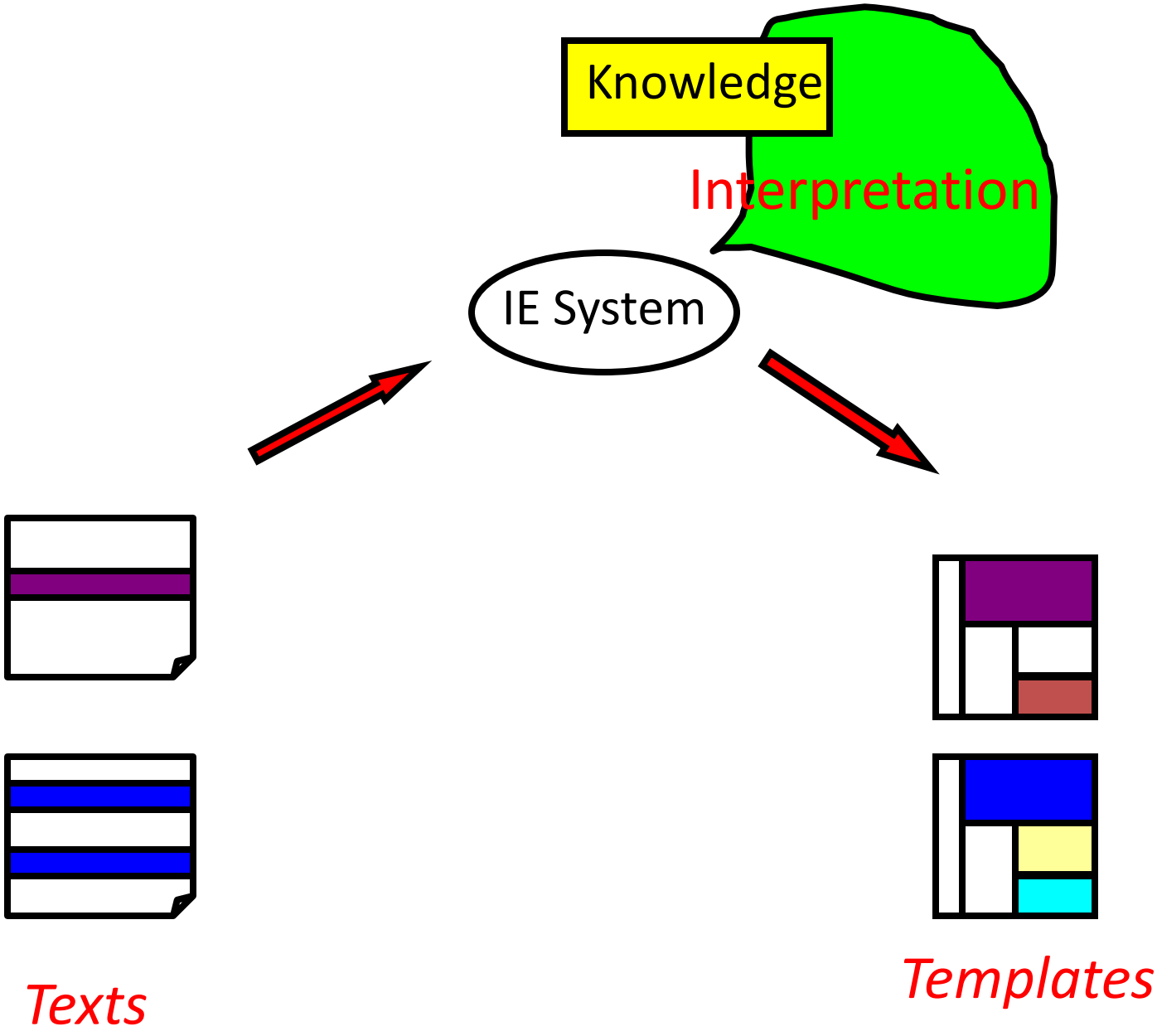



Collection of Texts

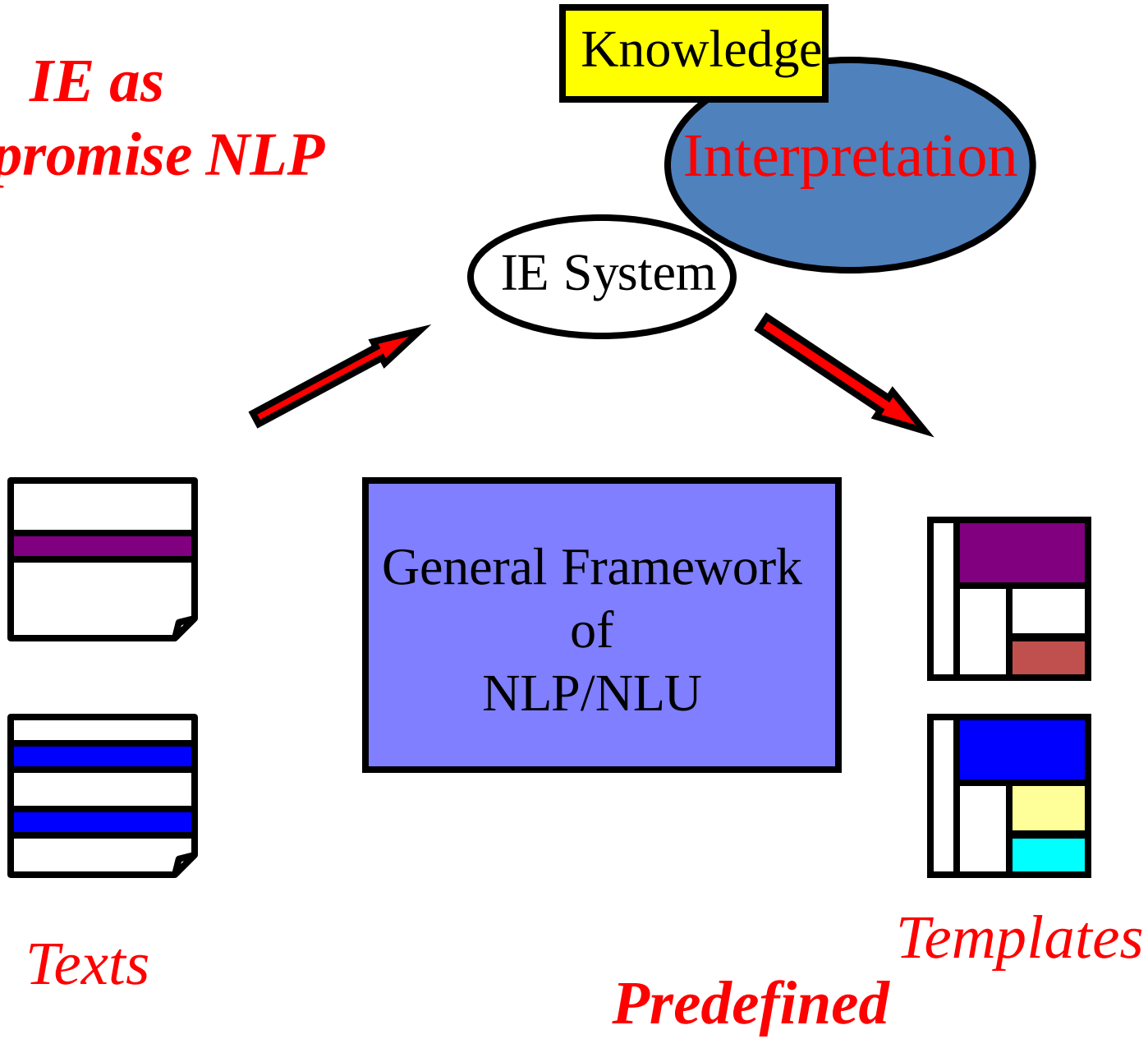

Knowledge
Interpretation
Characterization of Texts
IR System
Queries
Collection of Texts


Knowledge
Interpretation
Characterization of Texts
Passage
IR System
Queries
Collection of Texts


Characterization of Texts
Knowledge
Interpretation
Passage
IR System
IE System
Queries
Structures
of
Sentences
NLP
Collection of Texts
Texts
Templates


Knowledge
Interpretation
IE System
Texts
Templates

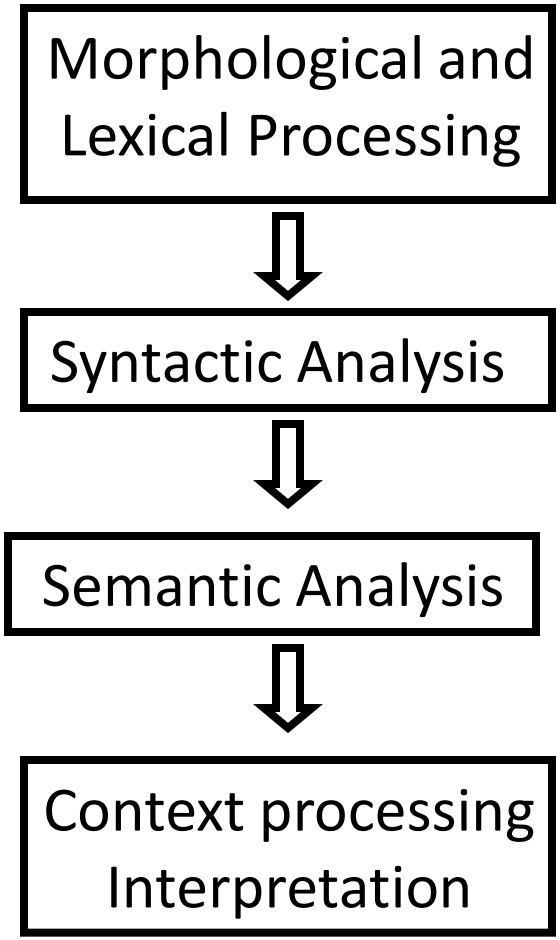
***IE as compromise NLP***



# *General Framework of NLP*

# General Framework of NLP

John runs.

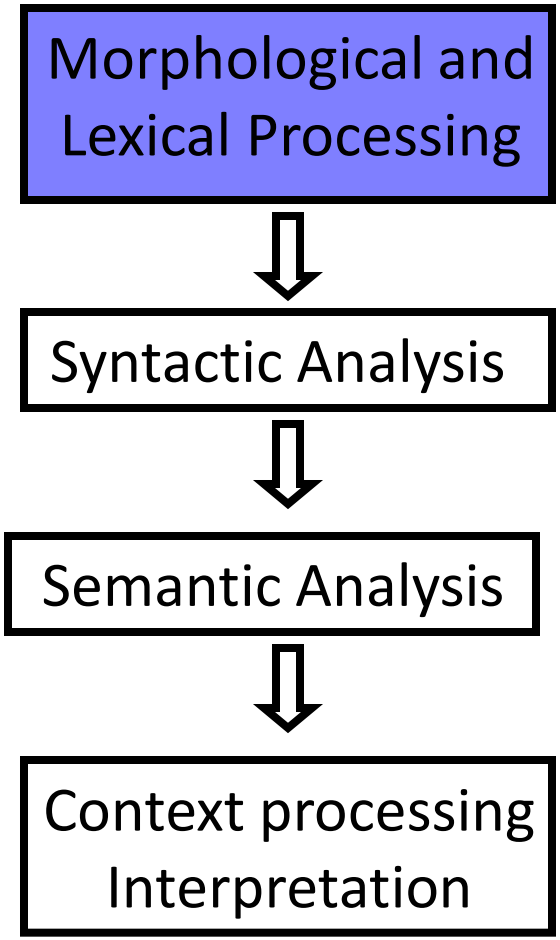


# General Framework of NLP

John runs.

John run+s.

P-N V 3-pre
N plu

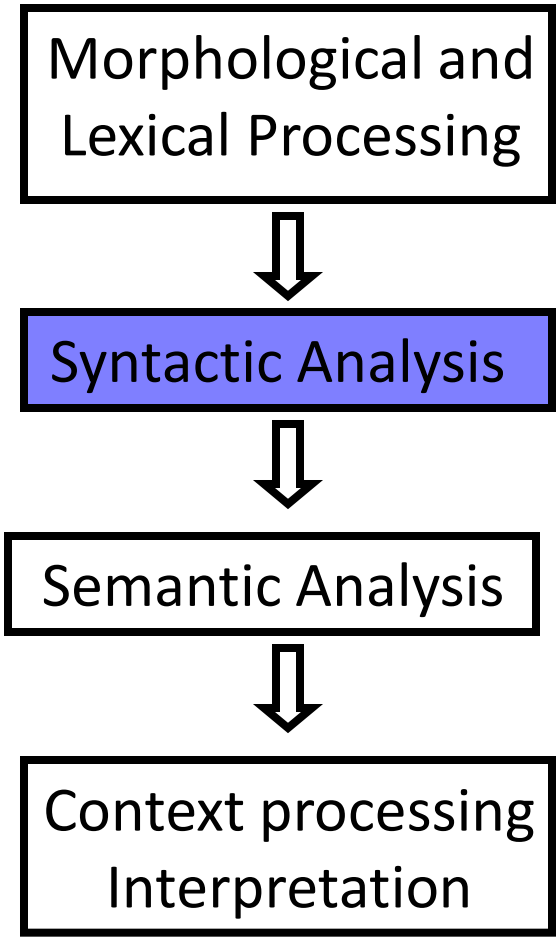


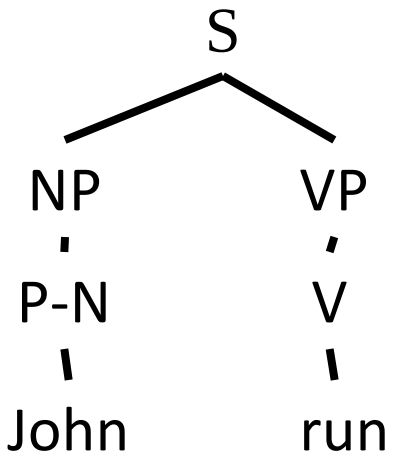
# General Framework of NLP

John runs.

John run+s.

P-N V 3-pre
N plu



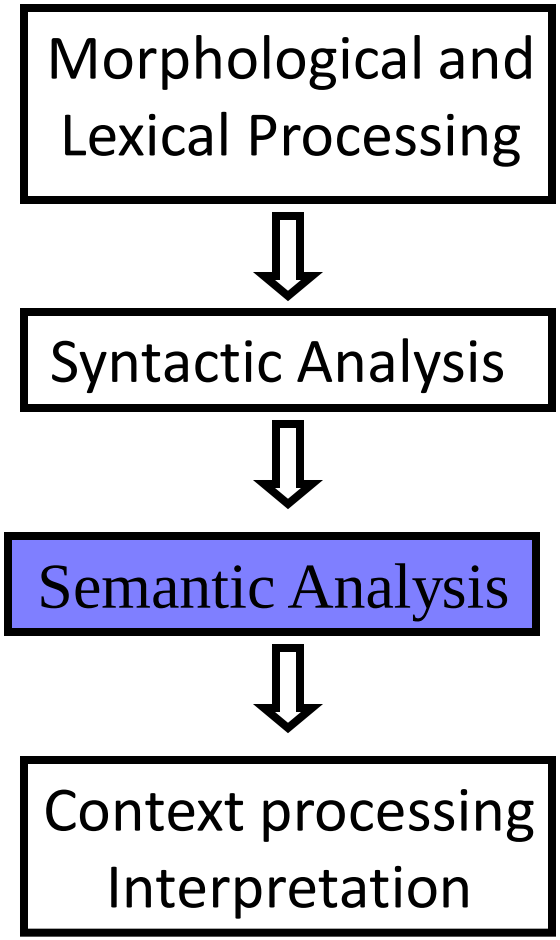



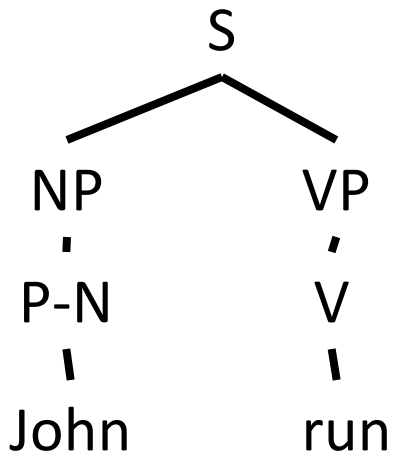
# General Framework of NLP

John runs.

John run+s.

P-N V 3-pre
N plu




[ Pred: RUN
Agent:John ]

# General Framework of NLP

John runs.

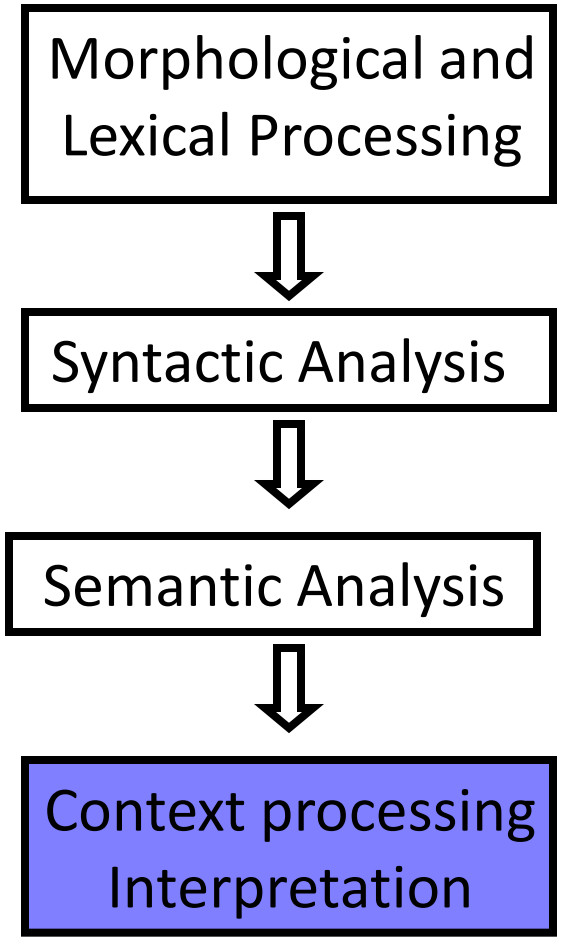
John run+s.

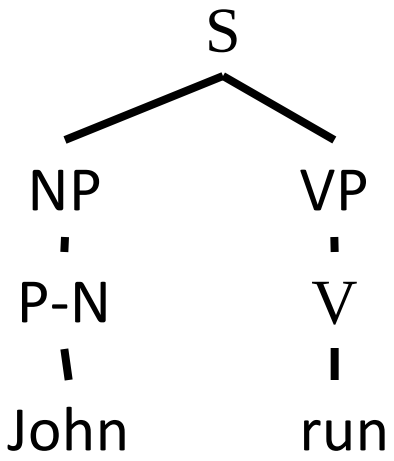
| P-N | V | 3-pre |
|---|---|---|
| | N | plu |

[ Pred: RUN
  Agent:John ]

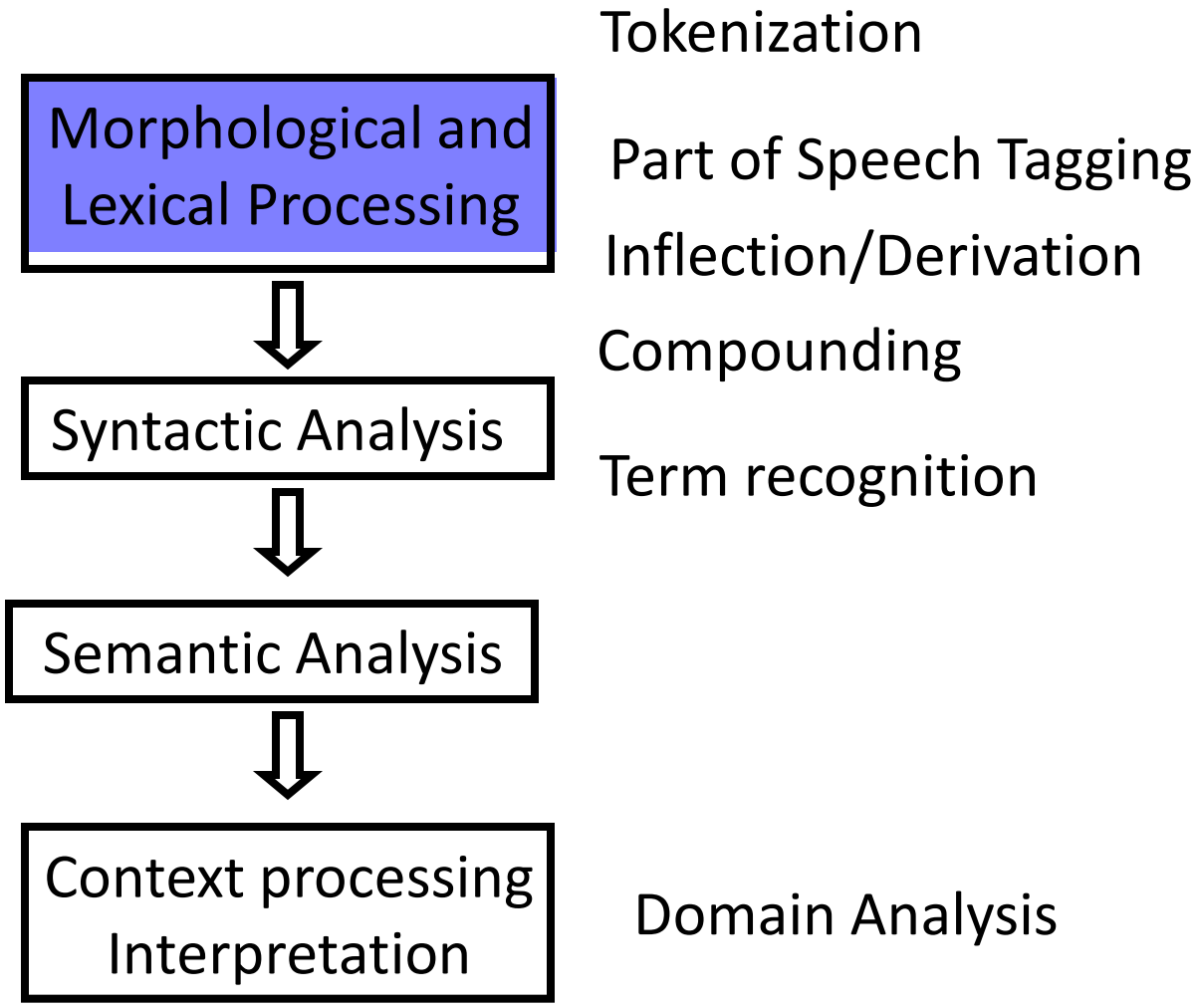
John is a student.

He runs.



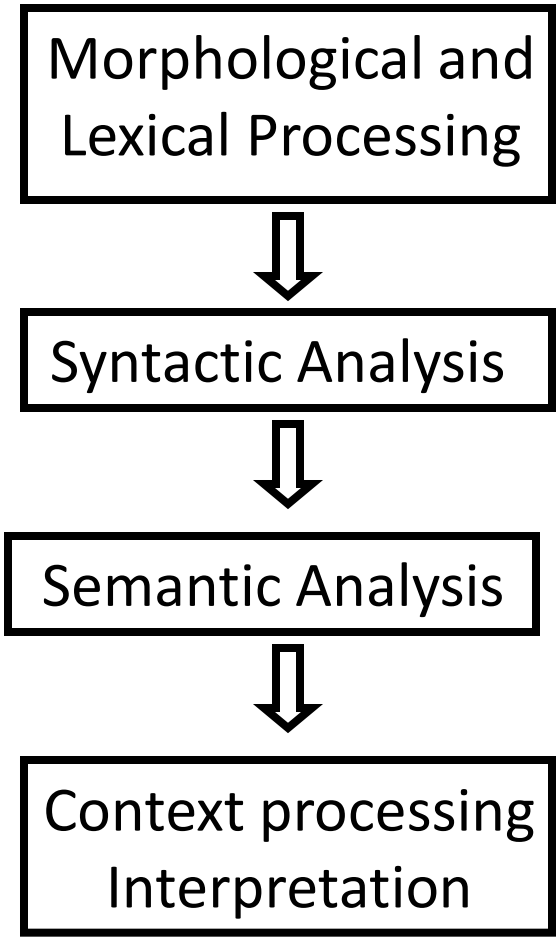
# General Framework of NLP



***Difficulties of NLP***

# General Framework of NLP

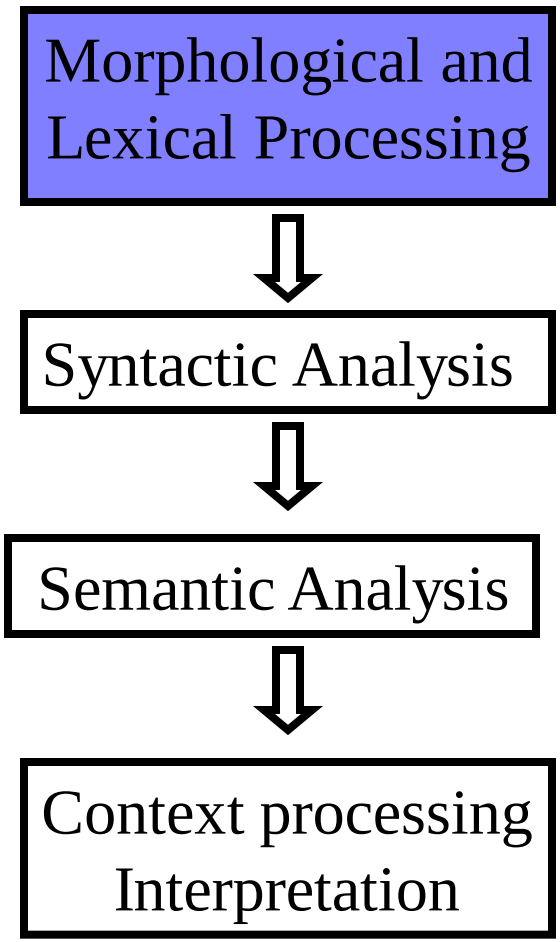
(1) Robustness:
Incomplete Knowledge



***Difficulties of NLP***

# General Framework of NLP

(1) Robustness:
Incomplete Knowledge




Incomplete Lexicons
- Open class words
- Terms

Term recognition

Named Entities
- Company names
- Locations
- Numerical expressions

***Difficulties of NLP***

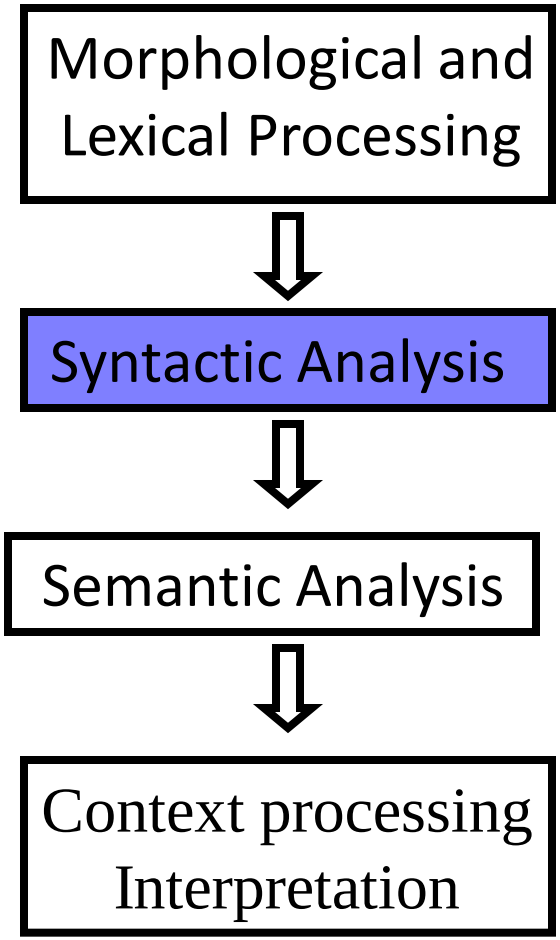
# General Framework of NLP

(1) Robustness:

Incomplete Knowledge

Incomplete Grammar

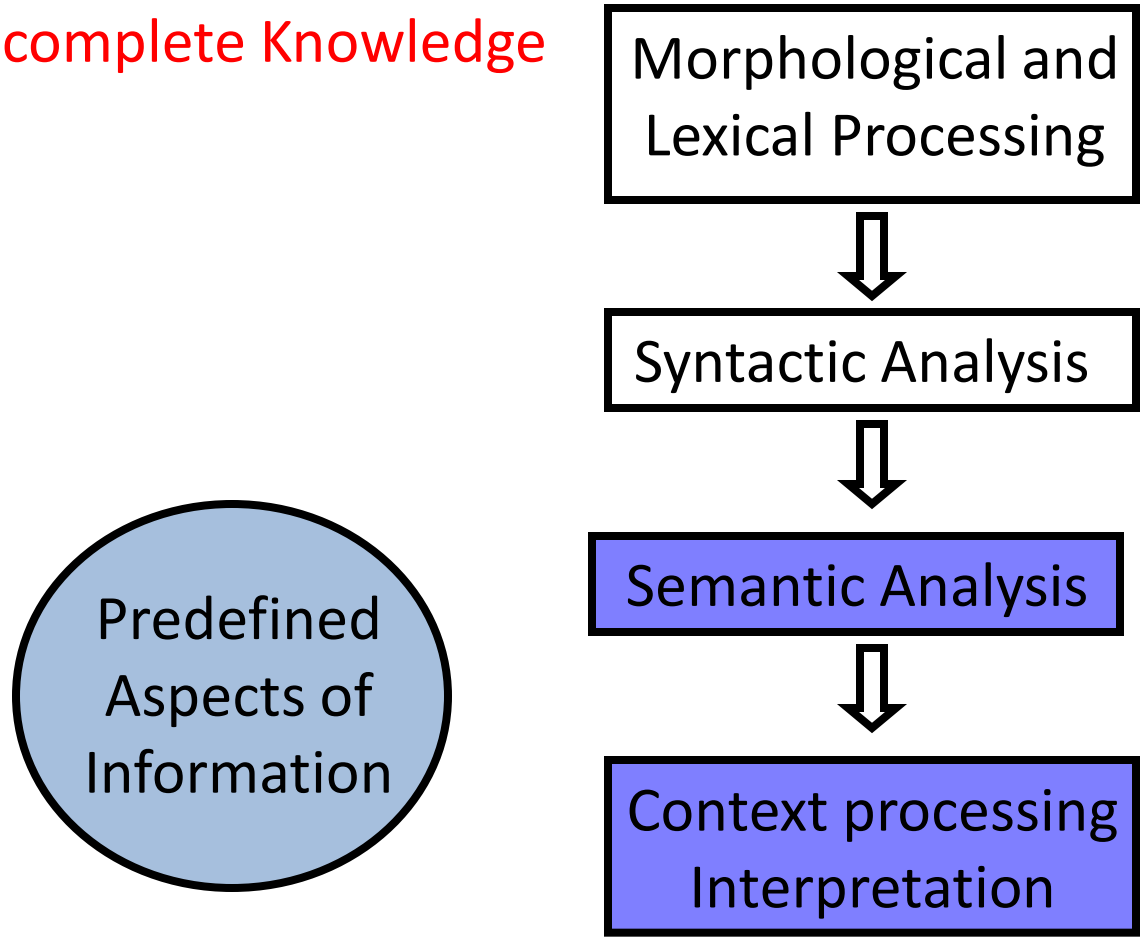
- Syntactic Coverage
- Domain Specific Constructions
- Ungrammatical Constructions



***Difficulties of NLP***

# General Framework of NLP

(1) Robustness:
Incomplete Knowledge

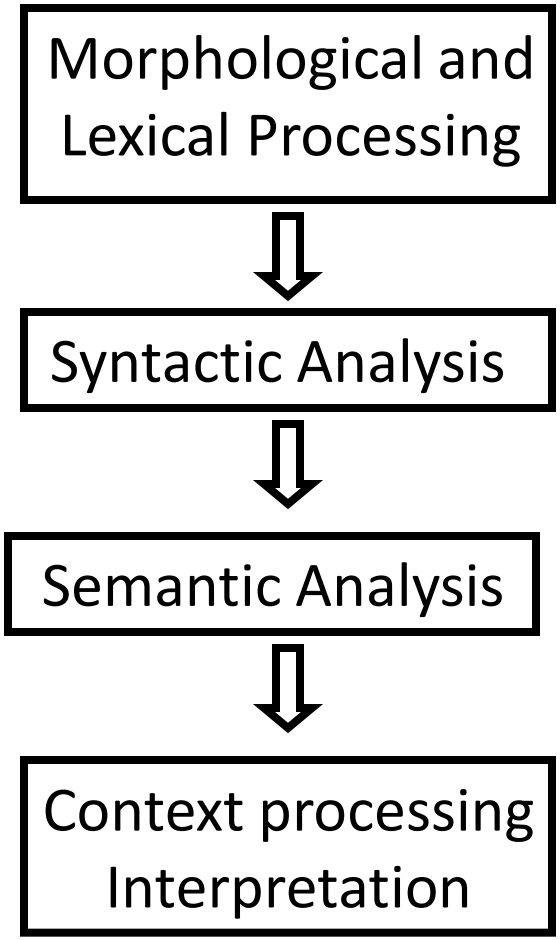


***Difficulties of NLP***

# General Framework of NLP

(1) Robustness:
Incomplete Knowledge

(2) Ambiguities:
Combinatorial
Explosion

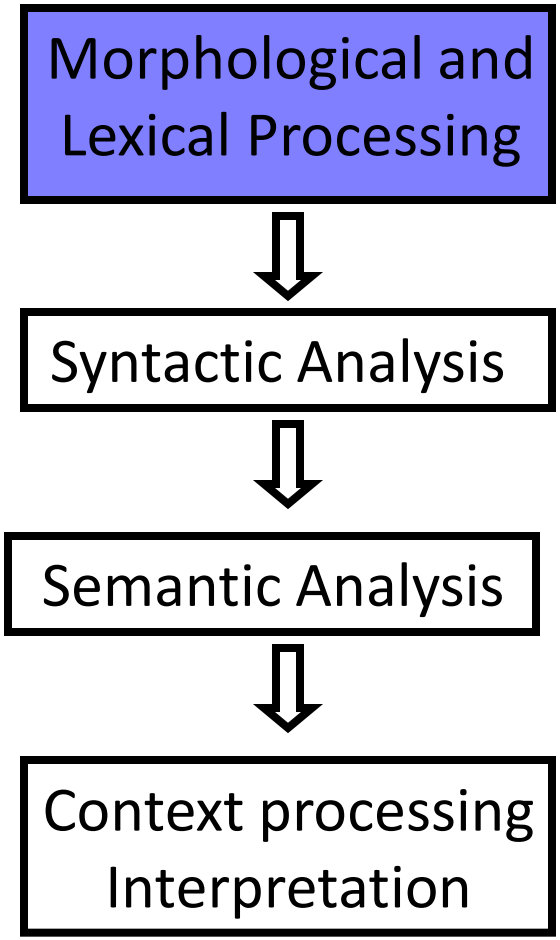


***Difficulties of NLP***

# General Framework of NLP

(1) Robustness:
Incomplete Knowledge

(2) Ambiguities:
Combinatorial
Explosion




Most words in English are ambiguous in terms of their part of speeches.

runs: v/3pre, n/plu

clubs: v/3pre, n/plu

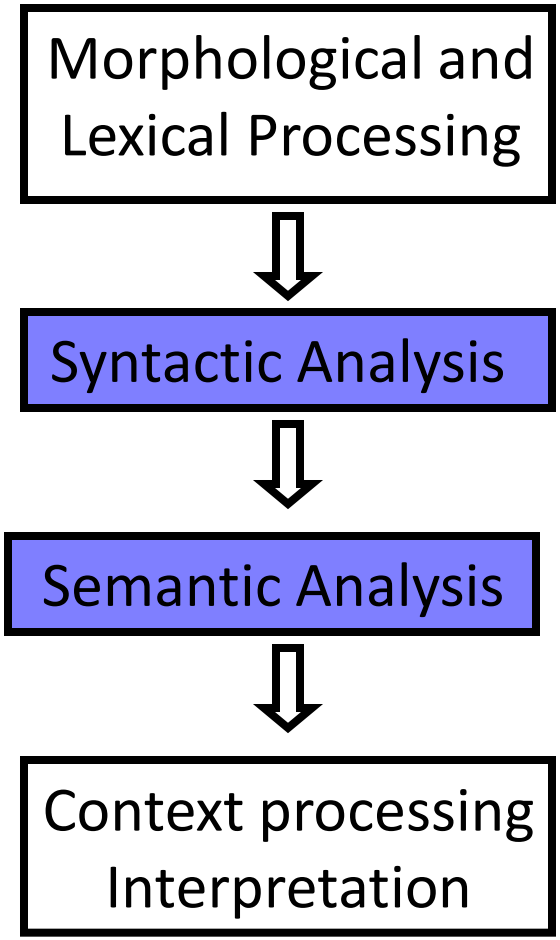
and two meanings

***Difficulties of NLP***

# General Framework of NLP

(1) Robustness:
Incomplete Knowledge

(2) Ambiguities:
Combinatorial
Explosion




Structural Ambiguities

Predicate-argument Ambiguities

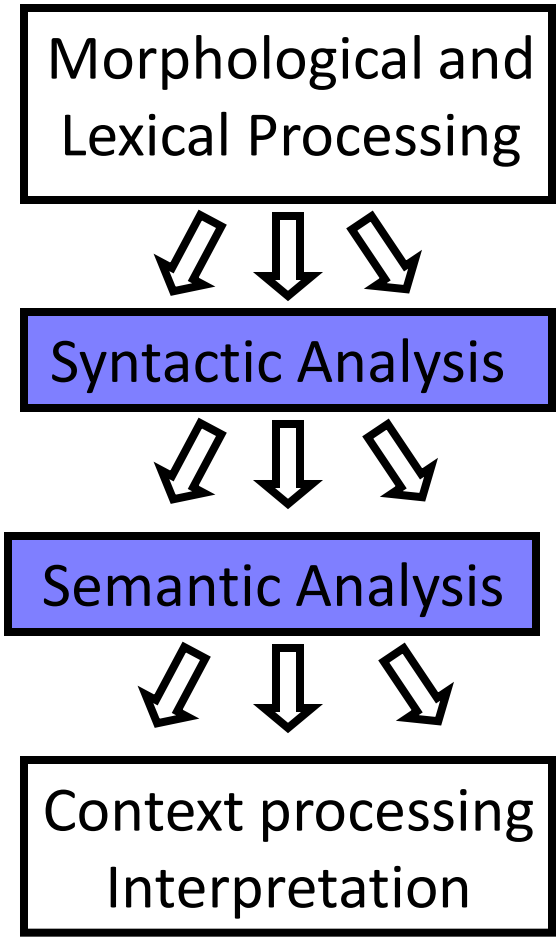
***Difficulties of NLP***

# General Framework of NLP

(1) Robustness:
Incomplete Knowledge

(2) Ambiguities:
Combinatorial
Explosion

***Combinatorial Explosion***




Structural Ambiguities

Predicate-argument Ambiguities

# Note:

# Ambiguities vs Robustness

More comprehensive knowledge: More Robust

big dictionaries

comprehensive grammar

More comprehensive knowledge: More ambiguities

Adaptability: Tuning, Learning

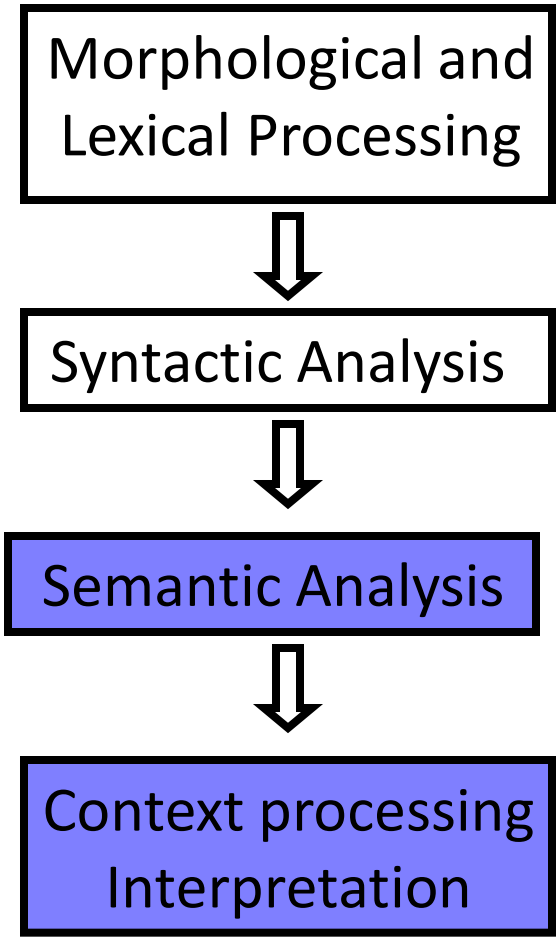
# *Framework of IE*

IE as compromise NLP

***Difficulties of NLP***

# General Framework of NLP

(1) Robustness:
Incomplete Knowledge



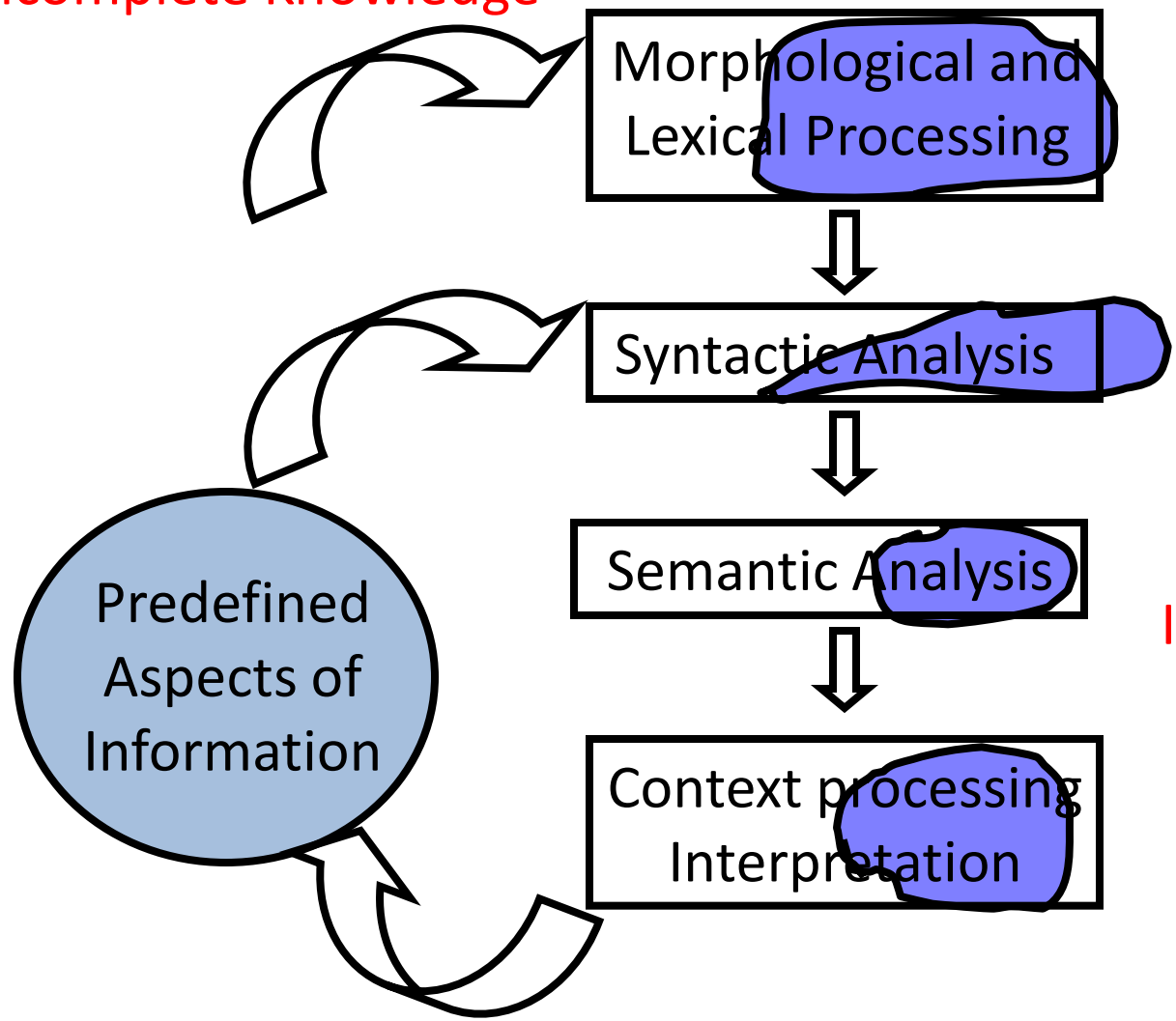

Predefined Aspects of Information

Incomplete
Domain Knowledge
Interpretation Rules

***Difficulties of NLP***

# General Framework of NLP

(1) Robustness:
Incomplete Knowledge




Incomplete
Domain Knowledge
Interpretation Rules

# Techniques in IE

(1) Domain Specific Partial Knowledge:

Knowledge relevant to information to be extracted

(2) Ambiguities:

Ignoring irrelevant ambiguities

Simpler NLP techniques

(3) Robustness:

Coping with Incomplete dictionaries

(open class words)

Ignoring irrelevant parts of sentences

(4) Adaptation Techniques:

Machine Learning, Trainable systems

General Framework of NLP

Morphological and Lexical Processing

⇩

Syntactic Analysis

⇩

Semantic Anaysis

⇩

Context processing Interpretation

Part of Speech Tagger

***95 %***

FSA rules

Statistic taggers

Local Context

Statistical Bias

Open class words:

Named entity recognition

(ex) Locations

Persons

Companies

Organizations

Position names

***F-Value 90***

⇩

***Domain Dependent***

Domain specific rules:

<Word><Word>, Inc.

Mr. <Cpt-L>. <Word>

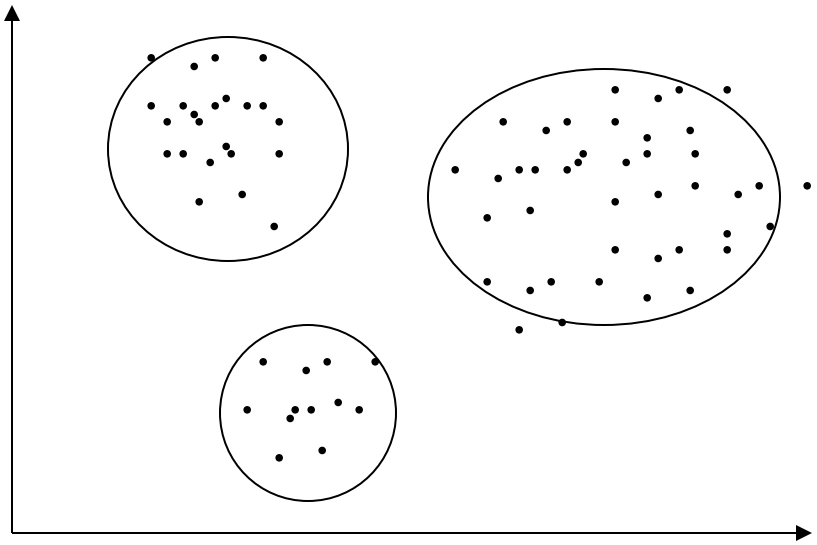
Machine Learning:

HMM, Decision Trees

Rules + Machine Learning

# Text Clustering

- K-Nearest Neighbor

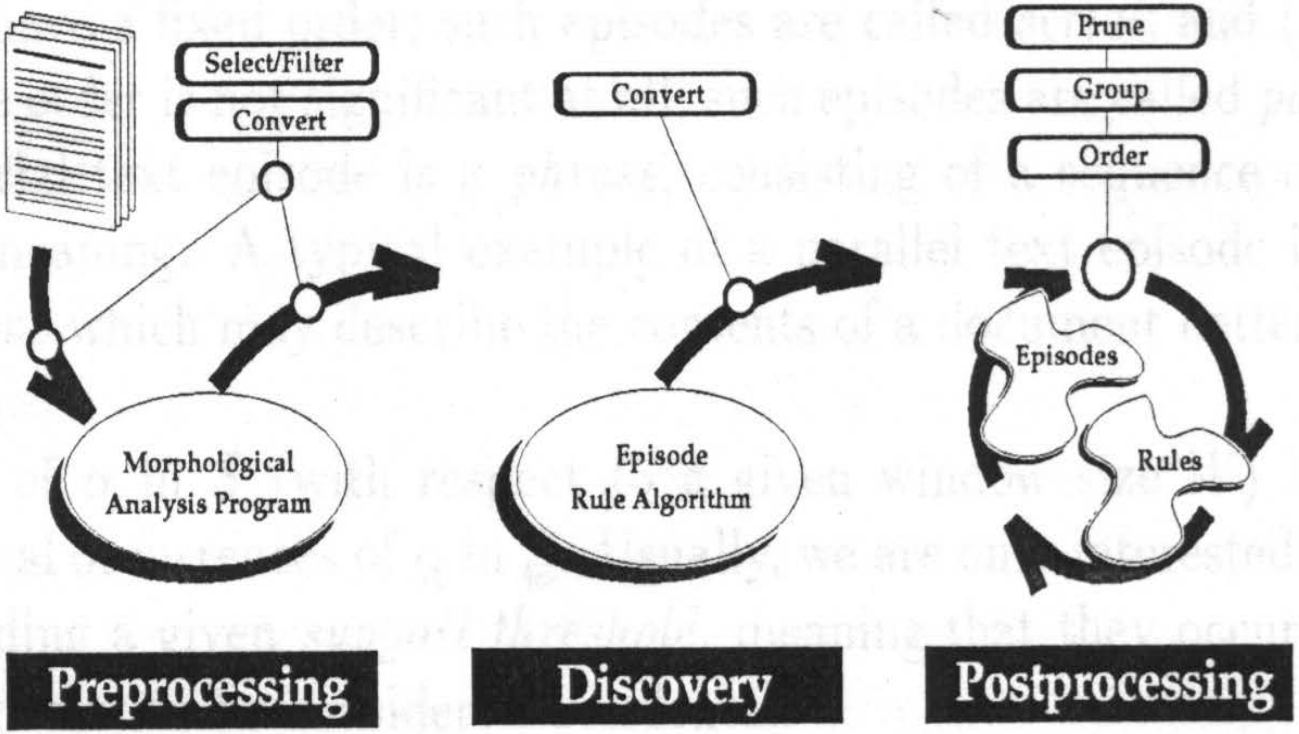


$$Dice(d, d') = \frac{2 \times |d \cap d'|}{|d| + |d'|}$$

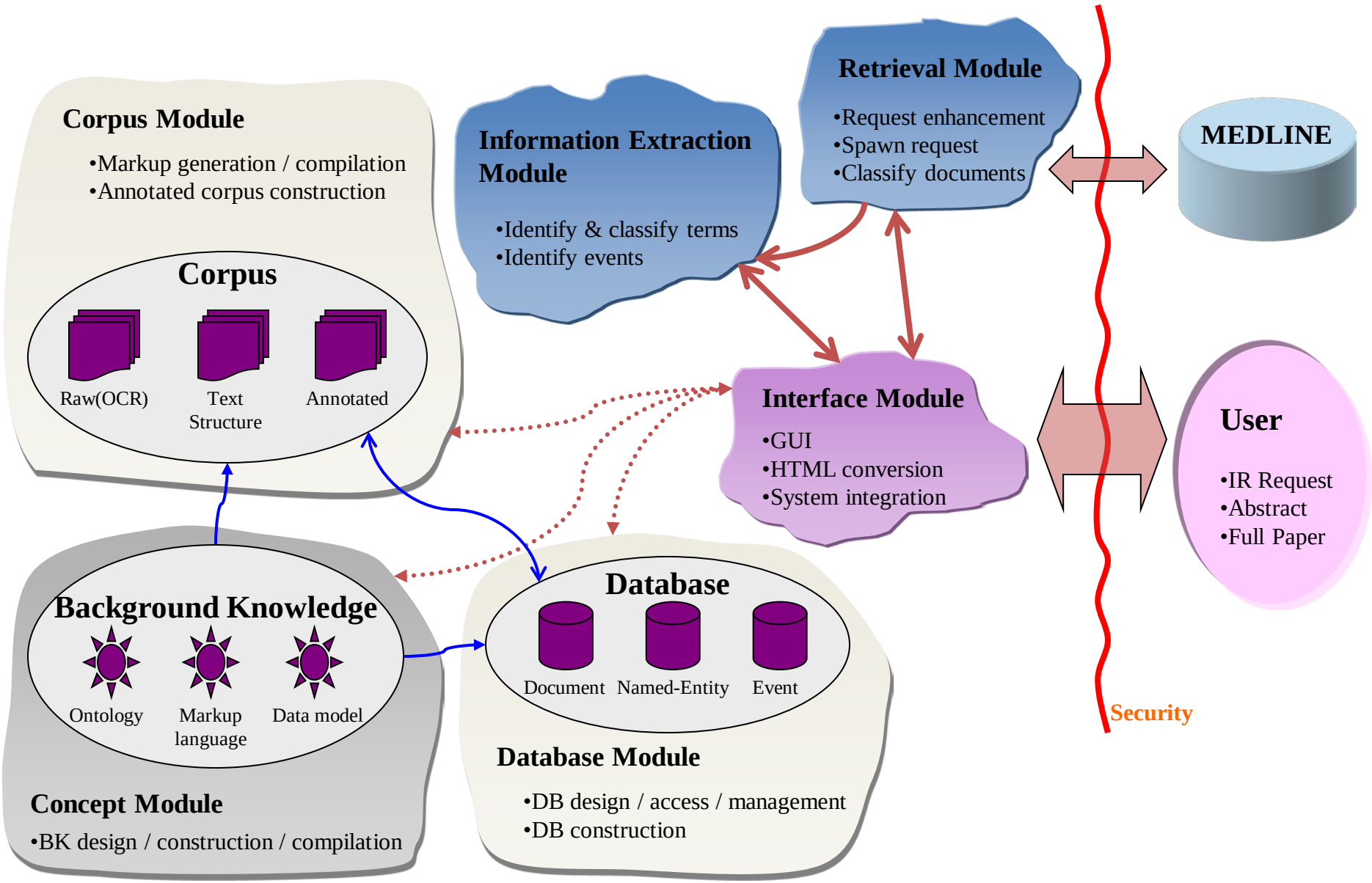
# Text Categorization

- Application
  - Query Adjusted
  - Query Extended
  - Document Indexing
  - Information Filtering

# KDD of Text Mining



# Overview of GENIA System



# Other Applications

- Customer Profile Analysis
- Patent Analysis
- Information Dissemination
- Company Resource Planning