

# **Internet-based *Information Extraction Technologies (CS438)***

Fang Li (李芳)

*Dept. of Computer Science & Engineering*

[Li-fang@cs.sjtu.edu.cn](mailto:Li-fang@cs.sjtu.edu.cn)

# Contents

---

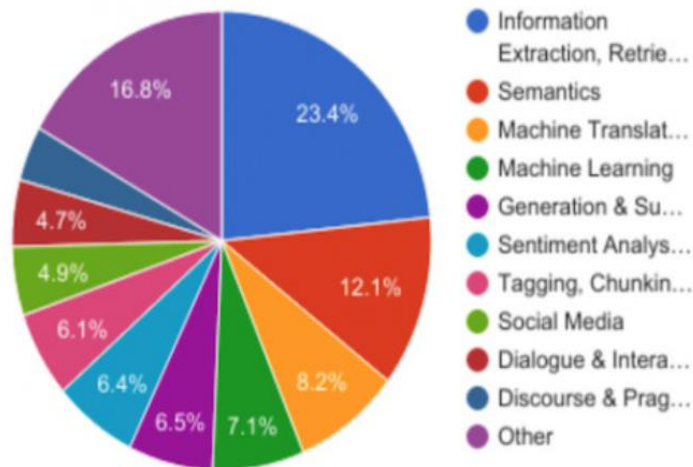
- ◆ Motivation & Aim
- ◆ Related Technologies
- ◆ Application Domain
- ◆ Course Introduction & Syllabus
- ◆ Teaching Methods

# Why we need to know IE

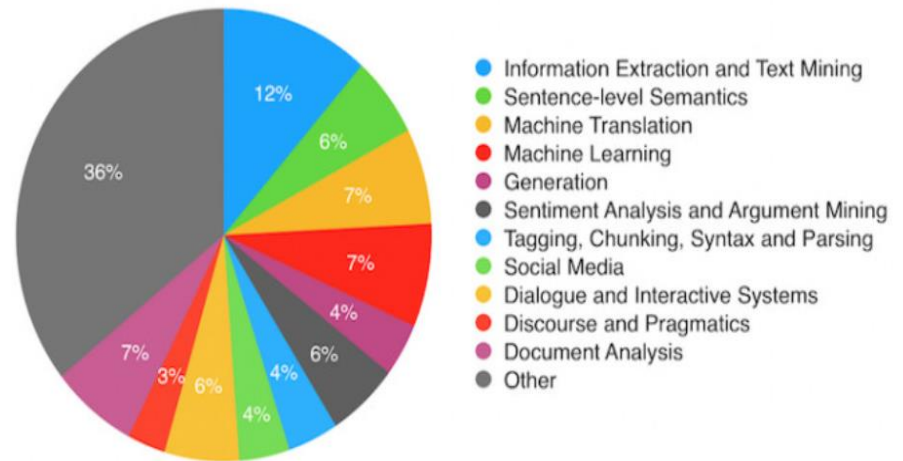
- ◆ Information Extraction is a **hot research topic** in recent years.
- ◆ Information Extraction technologies have been **used in many applications** in the real world.

# Hot research topic in many conferences

ACL 2017 Submissions



ACL 2018 Submissions



Annual Meeting of the Association for Computational Linguistics (ACL)

lecture of Internet-based IE technologies

# ACL2019 Topic Analysis

Top three:

1) 信息抽取与文本挖掘 (占 ACL 2019 有效提交论文总数的 **9.2%**)。

2) 机器学习 ( **8.2%** ACL 2019)

3) 机器翻译 ( **7.7%** ACL 2019)

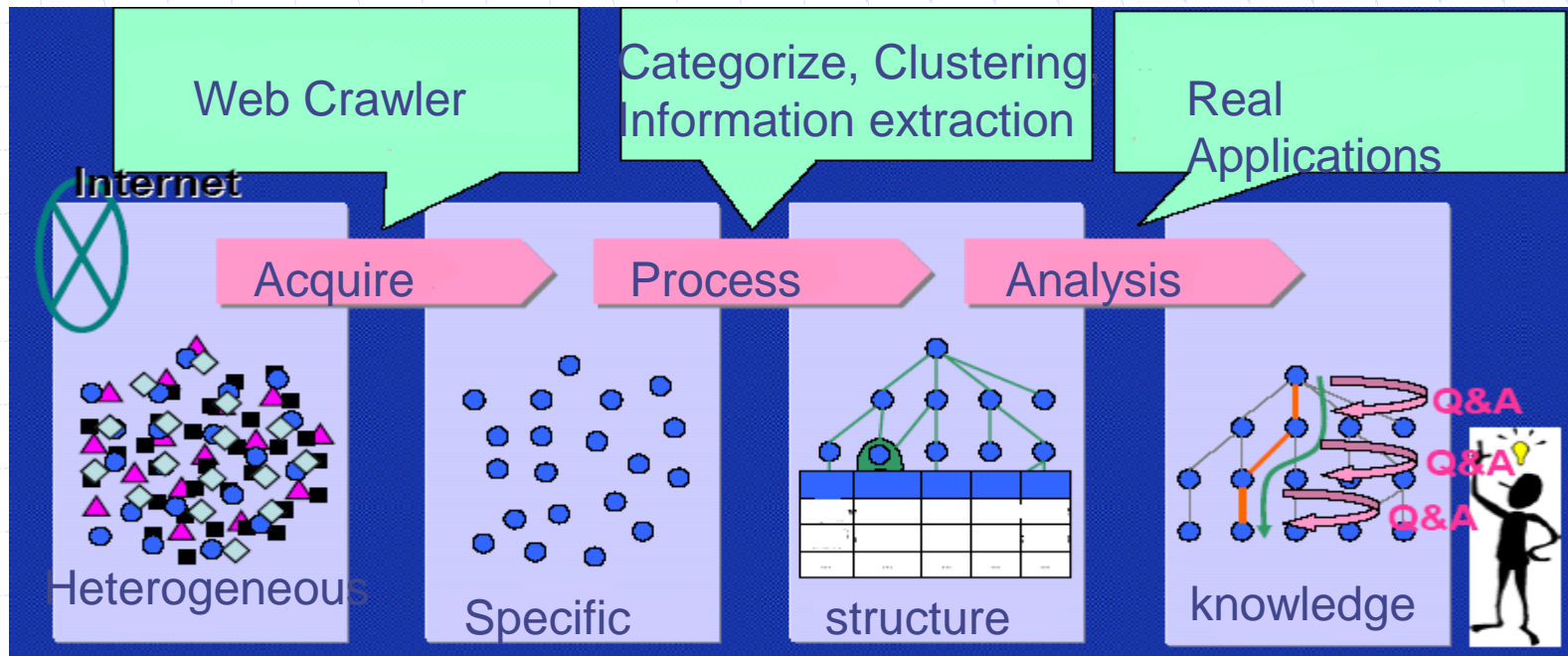
	Area	Long	Short	Total
1.	Information Extraction, Text Mining	156	93	249
2.	Machine Learning	148	73	221
3.	Machine Translation	102	105	207
4.	Dialogue and Interactive Systems	125	57	182
5.	Generation	97	58	155
6.	Question Answering	99	55	154
7.	Sentiment Analysis, Argument Mining	91	60	151
8.	Word-level Semantics	78	59	137
9.	Applications	65	72	137
10.	Resources and Evaluation	70	60	130
11.	Multidisciplinary, AC COI	70	44	114
12.	Sentence-level Semantics	70	42	112
13.	Tagging, Chunking, Syntax, Parsing	50	49	99
14.	Social Media	51	42	93
15.	Summarization	48	35	83

# The Aim of the Lecture

- ◆ Let computer to **understand** the meaning of these web pages which are written by **natural languages**.
- ◆ Let computer to **extract** the knowledge of text written by **natural language**

# Internet-based Information Extraction

## ◆ Web as an information resource



# Data – Knowledge – Information

- ◆ **Data:** recoded facts or figures
- ◆ **Knowledge:** the understanding required to convert data into information and to apply it to real-world situations
- ◆ **Information:** the value derived from data through the application of knowledge

**Information = data + knowledge**



# Data vs. Knowledge

**Knowledge are data with meaning**, e.g., a property (or feature) of an object (size of a human, name of a company). Note that the same data element might have several possible interpretations.

# For example (data & knowledge)

2002	ford
thundebird	5,500
Red,ABS,6 CD changer	\$33,000
(916)972- 9117	

## Car Advertisement



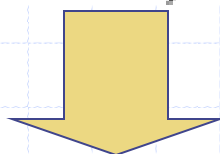
Year 2002  
Make Ford  
Model Thunderbird  
Mileage 5,500 miles  
Features Red  
ABS  
6 CD changer  
keyless entry  
Price \$33,000  
Phone (916) 972-9117

# Knowledge vs. Information

- ◆ **Knowledge:** a model of the world (structural and functional properties of the real world)
- ◆ **Information:**
  - is that **part of knowledge** which is used to **solve a certain problem** (Information System view)
  - **Information systems extract that knowledge just in time, a user needs in context of a given situation.**

# Information Extraction (IE): the **subdiscipline** of Artificial Intelligence.

19 March — A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb — allegedly detonated by urban guerrilla commandos — blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).



Text →  
structural data

INCIDENT TYPE	bombing
DATE	March 19
LOCATION	El Salvador: San Salvador (city)
PERPETRATOR	urban guerrilla commandos
PHYSICAL TARGET	power tower
HUMAN TARGET	-
EFFECT ON PHYSICAL TARGET	destroyed
EFFECT ON HUMAN TARGET	no injury or death
INSTRUMENT	bomb

# Internet-based Information Extraction

from **William W. Cohe**

## Web site specific

Formatting

Amazon.com Book Pages

## Genre specific

Layout

Resumes

## Wide, non-specific

Language

University Names

amazon.com. VIEW CART

WELCOME YOUR STORE BOOKS ELECTRONICS DVD TOYS & GAMES M

SEARCH BROWSE SUBJECTS BESTSELLERS MAGAZINES CORPORATE ACCOUNT

Get \$5 off

LOOK INSIDE! MACHIN LEARNING

Machine Learning by Tom M. Mitchell

NEW Super Saver Shipping FREE

Learning in Graphical Models by Michael Irwin Jordan (Editor)

List Price: \$60.00 Price: \$60.00

This item ships for FREE with Super

Availability: Usually ships within 2 to 3 days

Used & new from \$20.00

Edition: Paperback | All Editions

See more product details

Great Buy

Buy this book with Probabilistic Reasoning in Intelligent Systems

Buy Together Today: \$128.95

Buy both now!

Jason D. M. Rennie

Massachusetts Institute of Technology  
MIT AI Lab NE43-733  
200 Technology Sq.  
Cambridge, MA 02139

http://www.ai.mit.edu/people/jrennie  
jrennie@ai.mit.edu  
(617) 253-5339

Research Interests

My main interests lie in the automated analysis of data for the purposes of classification, estimation and the acquiring of new knowledge. I have both interests in applying such techniques to real-world problems, and in the analysis of existing algorithms and the creation of new ones.

L. Douglas Baker

Home Address available upon request  
Wean Hall, 8102

Office Address School of Computer Science  
Carnegie Mellon University  
5000 Forbes Avenue  
Pittsburgh, PA 15213

Office Phone (412) 683-6036

Home Page http://www.cs.cmu.edu/~ldbapp

Objective A position in a dynamic, highly-skilled applied research and development team using statistical machine learning to solve large-scale, real-world tasks such as Information Retrieval and Text Classification.

Education Carnegie Mellon University Pittsburgh, PA  
Ph.D., Computer Science, in progress  
M.S., Computer Science, 1999  
Technical University of Berlin Berlin, Germany  
Exchange Fellow, 1992-1993  
University of Michigan Ann Arbor, MI  
M.S.E., Computer Science and Engineering, 1994 B.S.E.,  
Computer Engineering, Summa Cum Laude, 1992

Research Experience Carnegie Mellon University 1994-present

I am currently pursuing my dissertation research: a hierarchical probabilistic model for novelty detection in text. This work is being done as part of the Topic Detection and Tracking project at CMU, under the direction of Yimin Yang. The

8:30 - 9:30 AM	Invited Talk: Plausibility Measures: A General Approach <i>Joseph Y. Halpern, Cornell University</i>			
9:30 - 10:00 AM	Coffee Break			
10:00 - 11:30 AM	Technical Paper Sessions:			
<b>Cognitive Robotics</b>	<b>Logic Programming</b>	<b>Natural Language Generation</b>	<b>Complexity Analysis</b>	<b>Novelty Detection</b>
739: A Logical Account of Causal and Topological Maps <i>Emilio Remolina and Benjamin Kuipers</i>	116: A-System: Problem Solving through Abduction <i>Marc Denecker, Antonis Kakas, and Bert Van</i>	758: Title Generation for Machine-Translated Documents <i>Rong Jin and Alexander G. Hauptmann</i>	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories <i>Marco Cadoli,</i>	17: Ex Co from Ne Ka M W

**Dr. Steven Minton - Founder/CTO**  
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC's Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

**Frank Huybrechts - COO**  
Mr. Huybrechts has over 20 years of

- Press
- General information
- Directions maps

technologies

# Different tasks of IE

Jack Welch will retire as CEO of General Electric tomorrow.  
The top role at the Connecticut company will be filled by Jeffrey Immelt.

## Single entity

*Person:* Jack Welch

*Person:* Jeffrey Immelt

*Location:* Connecticut

Named entity  
extraction

## Binary relationship

*Relation:* Person-Title

*Person:* Jack Welch

*Title:* CEO

*Relation:* Company-Location

*Company:* General Electric

*Location:* Connecticut

Relation  
extraction

## N-ary record

*Relation:* Succession

*Company:* General Electric

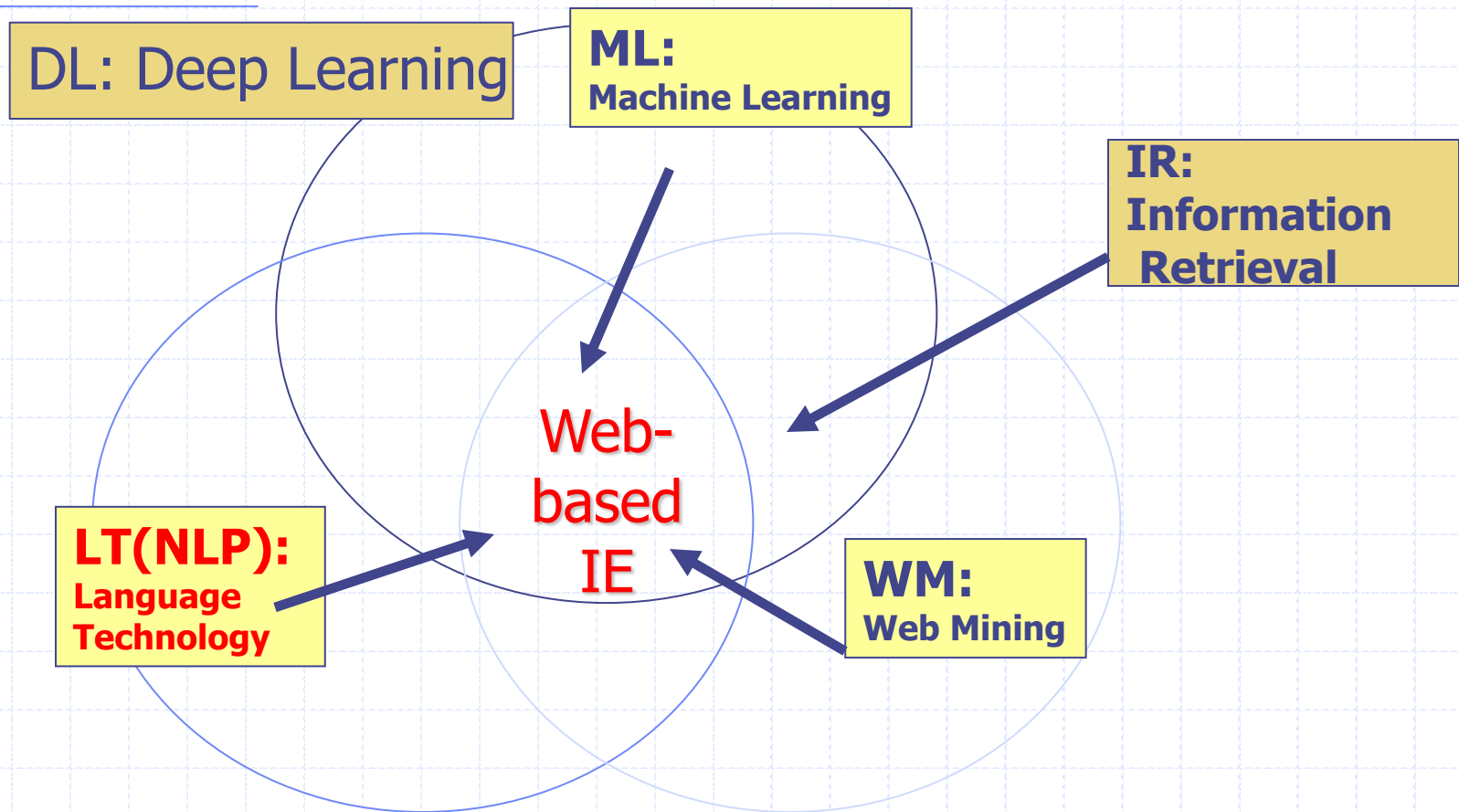
*Title:* CEO

*Out:* Jack Welch

*In:* Jeffrey Immelt

Event  
extraction

# Related Technologies



# What's the difference between machine learning, AI, and NLP?

---



Jason Eisner, computer science professor at Johns Hopkins

Answered Apr 5, 2015 · Upvoted by Venkata Neehar, [Master's Computer Science & Mathematics, University of California, San Diego](#) and Sudeep Narkar, [M.S Computer Science, Binghamton University \(2017\)](#)

AI = building systems that can do intelligent things

NLP = building systems that can understand language  $\subsetneq$  AI

ML = building systems that can learn from experience  $\subsetneq$  AI

$NLP \cap ML$  = building systems that can learn how to understand language

NLP pursues a set of problems within AI.

ML also pursues a set of problems within AI, whose solutions may be useful to help solve other AI problems. Most AI work now involves ML because intelligent behavior requires considerable knowledge, and learning is the easiest way to get that knowledge.



# Language Technology: as **normalization**

IE vs. LT

Natural Language Processing includes:

- ◆ Tokenization
- ◆ Morphological Analysis
- ◆ Special phrases: Date and time, Proper names, Number expressions
- ◆ General Phrases: nominal phrases, prepositional phrases, ...
- ◆ Structural Analysis: complex flat structure
- ◆ Semantic Analysis: agent, role, target

# Natural Language Processing

- **Issues in tokenization:**

Hewlett-Packard → Hewlett Packard?

Lowercase → lower-case, lower case?

San Francisco → one token or two?

- **Issues in morphological analysis**

Windows, window      U.S.A, USA

→ need to **normalize terms**

Fed vs. fed, US vs. us → case is helpful

- **Period "." is quite ambiguous**

1. Sentence boundary
2. Abbreviations like Inc. or Dr.
3. Numbers like .02% or 4.3

# ML & DL: as instruments

## IE vs. ML(DL)

- ◆ Use machine learning algorithms(or DL) to automatically draw extraction patterns or rules for information extraction.
- ◆ Use machine learning methods (or DL) to create IE systems for new domains.

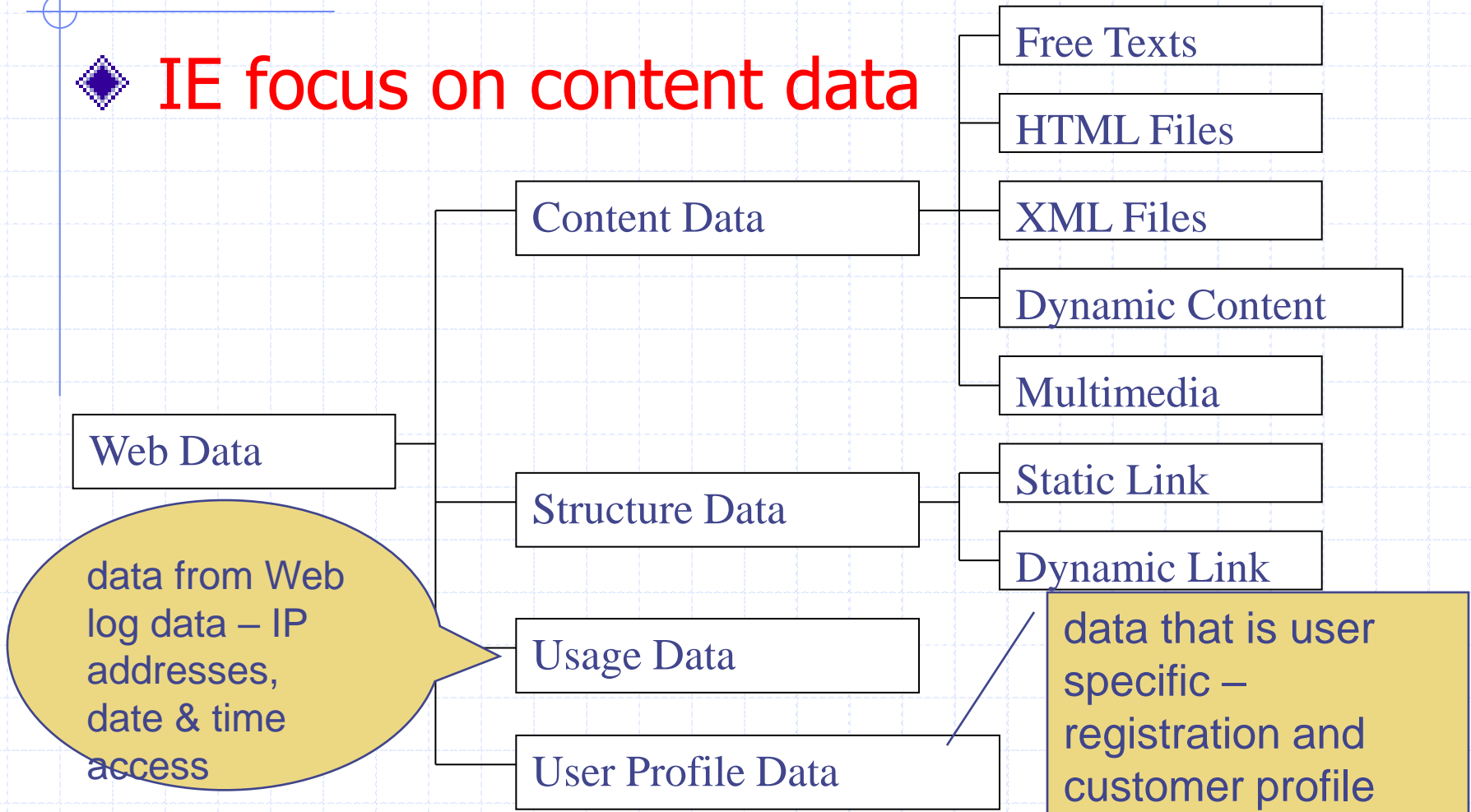
**Aim : to learn from a training corpus or seed examples and predict the new one.**

# Interacts with Web Mining IE vs. WM

- ◆ **Data Mining** (from database): information extraction and discovery of relational data
- ◆ **Text Mining** (from text documents): data mining using domain-independent shallow text processing
- ◆ **Web Mining (from the Internet)** : The use of data mining techniques to uncover hidden patterns or relationships among available Web data.
  - Data on the WWW: all the content information available on-line
  - Web log data: users' on-line activities (Cookies)
  - Web structure data: web linkage information

# Web Mining (cont.) IE vs. WM

## ◆ IE focus on content data



# Information Retrieval: as a precondition

IR vs.IE

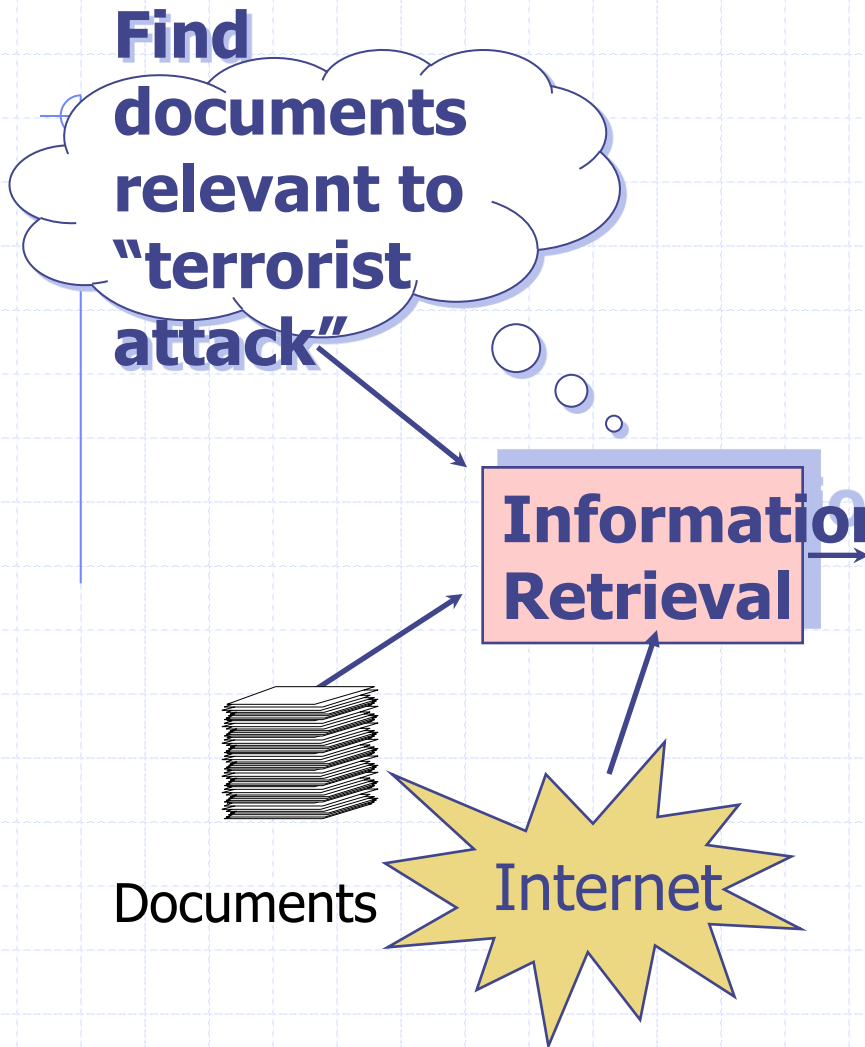
## ◆ Information Retrieval

- user
- dynamic, open domain

## ◆ Information Extraction

- system
- static, predefined -> open domain

# IE vs. IR: What IR System can do



[Terrorist attack on New York City and Washington DC on 2001-SEP-11](#) - [ 翻译此页 ]

[Terrorist attack on New York City and Washington DC on 2001-SEP-11](#).

[www.religioustolerance.org/terr\\_010911.htm](http://www.religioustolerance.org/terr_010911.htm) - 28k - [网页快照](#) - [类似网页](#)

[Emergency Response to Chemical/Biological Terrorist Incidents](#) - [ 翻译此页 ]

... and emergency medical units to cope with a **terrorist attack** that used nuclear, chemical or biological weapons. ... When dealing with any potential **terrorist attack**, past experience has taught that the first necessary task is to ...

[www.emergency.com/cbwlesn1.htm](http://www.emergency.com/cbwlesn1.htm) - 31k - [网页快照](#) - [类似网页](#)

[CNN.com - Cheney: Kerry win risks terror attack - Sep 7, 2004](#) - [ 翻译此页 ]

... United States at risk of another "devastating" **terrorist attack**, Vice President Dick Cheney told supporters Tuesday. ... and another **terrorist attack** occurs, then it's your fault," Edwards said during a stop in Chillicothe, Ohio. ...

[www.cnn.com/2004/ALLPOLITICS/09/07/cheney.terror/](http://www.cnn.com/2004/ALLPOLITICS/09/07/cheney.terror/) - 45k - 2005年4月23日 - [网页快照](#)

[free resources : Terrorist Attack Resources : Youth Specialties](#) - [ 翻译此页 ]

... **Terrorist Attack** Resources. We've compiled some resources and perspectives to help you process the unspeakable tragedy of 9/11—and help your kids do the same:. Articles and Columns; Concert of Prayer; Discussion Meeting; NEW! ...

[www.youthspecialties.com/free/attack/](http://www.youthspecialties.com/free/attack/) - 18k - [网页快照](#) - [类似网页](#)

# IE vs. IR: What IR System can not do

Given documents describing terrorist attacks, identify the involved criminal names

Sort the terrorist attacks based on when they occurred



*IE technologies are required!*



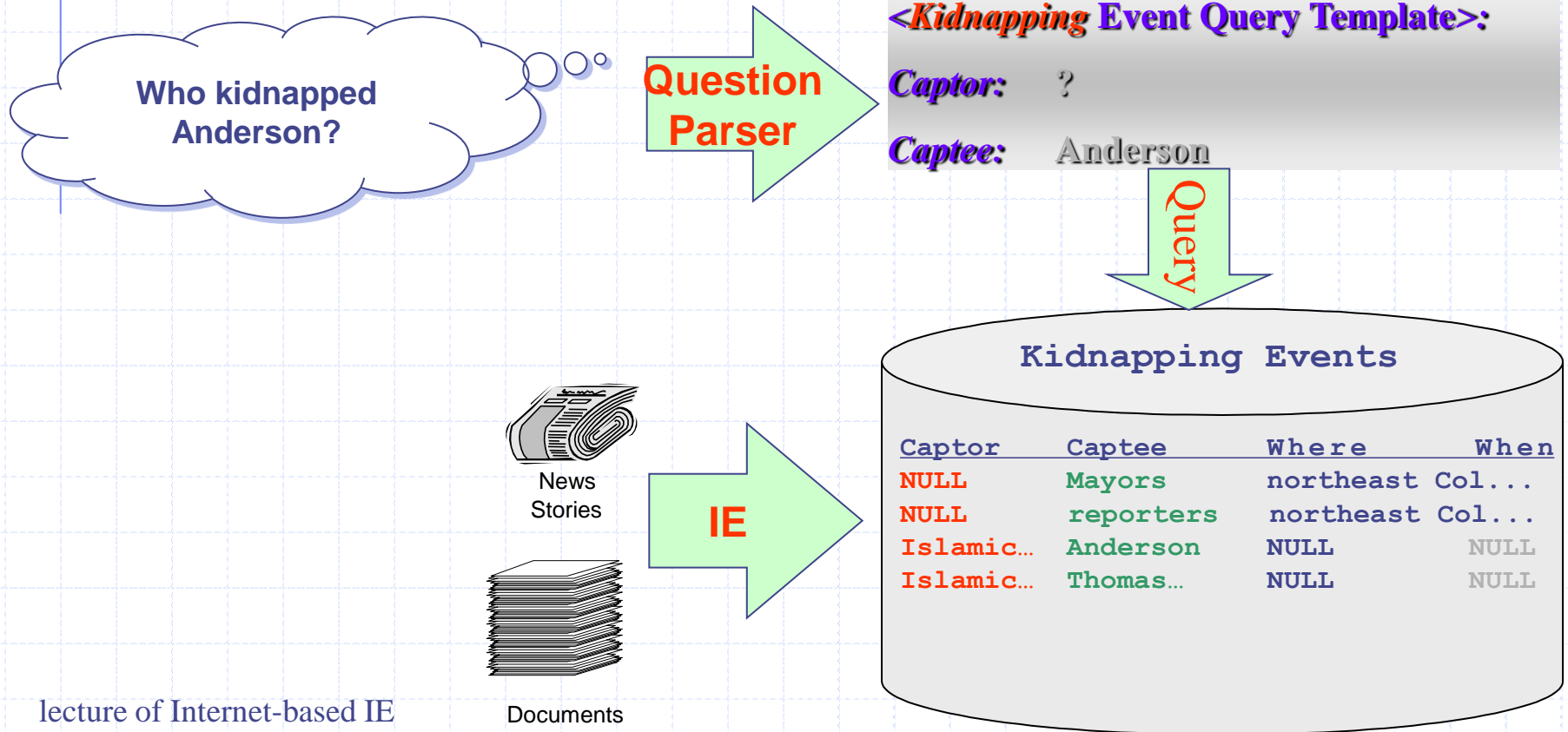
# IE has a high application impact & interacts with

1. Question Answering 问答系统
2. Text Classification and Abstractive Summarization 文本分类和自动摘要
3. Knowledge Graph 知识图谱
4. Dialog System 对话系统

Information extraction is a helpful step in these processes because the data become structured and semantically enriched.

# IE for Question-Answering

- **Question-Answering (QA):** query texts using natural language just like **querying a database using SQL**
- **QA supported by IE:**



## Sample Job Posting:

**Job Title:** Senior DBMS Consultant

**Location:** Dallas, TX

**Responsibilities:**

DBMS Applications consultant works with project teams to define DBMS based solutions that support the enterprise deployment of Electronic Commerce, Sales Force Automation, and Customer Service applications.

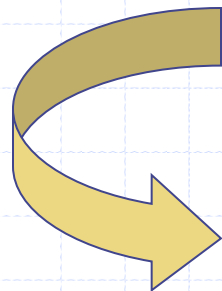
**Desired Requirements:**

3-5 years exp. developing Oracle or SQL Server apps using Visual Basic, C/C++, Powerbuilder, Progress, or similar.

Recent experience related to installing and configuring Oracle or SQL Server in both dev. and deployment environments.

**Desired Skills:**

Understanding of UNIX or NT, scripting language. Know principles of structured software engineering and project management



## Filled Job Template:

title: Senior DBMS Consultant

state: TX

city: Dallas

country: US

language: Powerbuilder, Progress, C, C++, Visual Basic

platform: UNIX, NT

application: SQL Server, Oracle

area: Electronic Commerce, Customer Service

required years of experience: 3

desired years of experience: 5

# IE Application in Real world

## Job Hunting

# IE Real System: Never-Ending Language Learning

- ◆ A research project at Carnegie Mellon University.
- ◆ Learns over time to read the Web since Jan.2010.
- ◆ Perform two tasks each day:
  1. Extract facts from text found in hundreds of millions of web pages
  2. Improve its reading competence, more accurately





















# Never-Ending Language Learning

◆ <http://rtw.ml.cmu.edu/rtw/index.php>

Recently-Learned Facts



Refresh

instance	iteration	date learned	confidence
<a href="#">babenhause</a> is a <a href="#">visualizable scene</a>	1110	24-jun-2018	99.8  
<a href="#">royal_oak_arkansas</a> is a <a href="#">beach</a>	1111	06-jul-2018	92.4  
<a href="#">civic_awards</a> is an <a href="#">award, championship, or tournament trophy</a>	1111	06-jul-2018	100.0  
<a href="#">illness_cycle</a> is a <a href="#">physiological condition</a>	1111	06-jul-2018	95.6  
<a href="#">hotel_lindenufer</a> is a <a href="#">place to ski</a>	1111	06-jul-2018	90.6  
<a href="#">n44</a> is the number of people that <a href="#">died in</a> the event <a href="#">n2008_lake_kivu_earthquake</a>	1115	03-sep-2018	100.0  
<a href="#">blenheim</a> is a city that <a href="#">lies on</a> the river <a href="#">wairau</a>	1111	06-jul-2018	100.0  
<a href="#">harrow</a> is a park <a href="#">in the city central_london</a>	1115	03-sep-2018	100.0  
<a href="#">kommersant_daily</a> is a newspaper <a href="#">in the city moscow</a>	1115	03-sep-2018	96.9  
the sports league <a href="#">first_nascar_uses</a> the venue <a href="#">daytona</a>	1113	15-aug-2018	93.8  

# Course Introduction

- Textbook is not the only reference
- **Research papers** are needed to read
- Different forms of exercises
- **Reading & writing, Discussion & Quiz,**
- Programming & presentation.
- Many models and technologies are involved, not detailed.
- Research oriented course

# Text Book and References

- ◆ Marie-Francine Moens, **Information Extraction: Algorithms and Prospects in a Retrieval context** (2006) Springer
- ◆ Sunita Sarawagi, **Information Extraction** from Foundations and Trends in Database vol.1, No.3(2007) 261-377
- ◆ Jerry R.Hobbs, Ellen Riloff, **Information Extraction** chapter 21 of Handbook of Natural Language Processing, 2nd Edition, Editors: Nitin Indurkha and Fred J. Damerau, Chapman & Hall/CRC Press, Taylor & Francis Group. (2010).
- ◆ Ralph Grishman, **Information Extraction: Capabilities and challenges**

# Course Web Sites

**My personal web sites:**

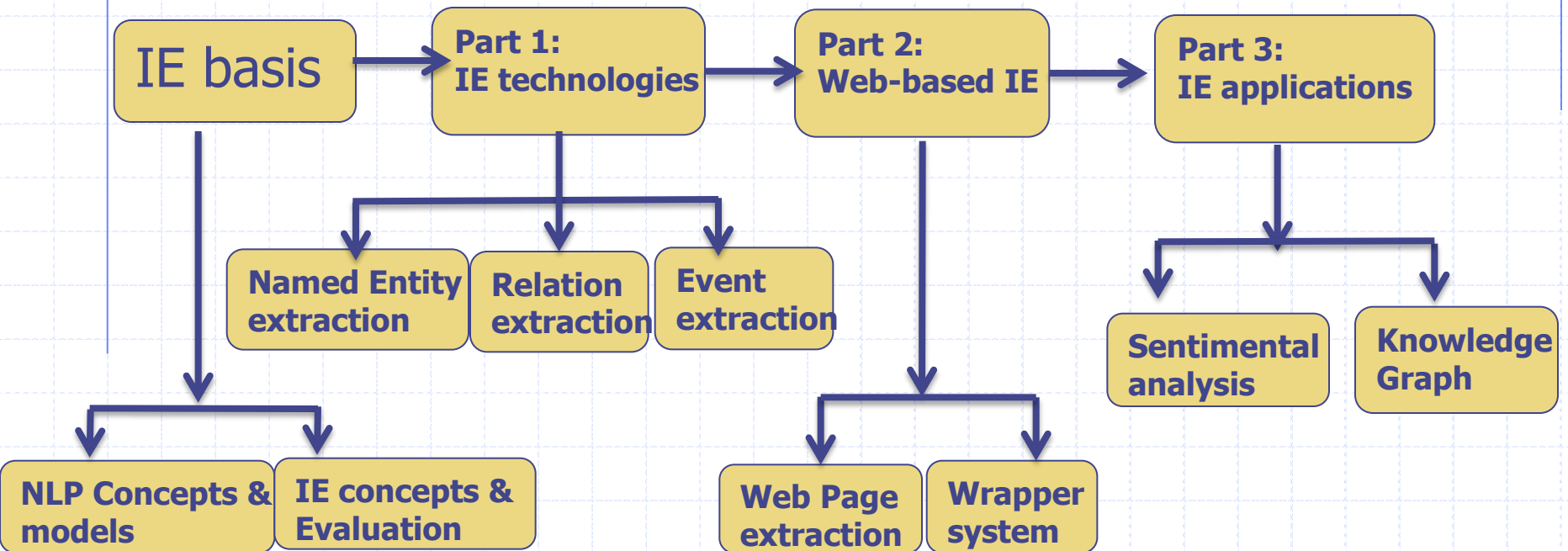
<http://www.cs.sjtu.edu.cn/~li-fang/>

**This course is in:**

<http://www.cs.sjtu.edu.cn/~li-fang/IEdescription.html>



# Course Knowledge Graph



The course organized as following:

**Information Extraction**

**targets:**

Free text

Web pages

**Information extraction**

**Tasks:**

Named Entity extraction

Relation extraction

Event extraction

**Information extraction**

**Research topics:**

Knowledge Graph

Sentimental analysis

# Teaching Methods

- ◆ **Critical Thinking:** 结合批判性思维的思想， 风靡美国50年的思维方法  
《批判性思维工具》
- ◆ **Think-Pair-Share :** 思考-讨论-分享方法

# 批判性思维

- ◆ 批判性思维是一种对思维方式进行思考的艺术，该艺术能够**优化**我们的思维方式。它包括三个紧密关联，互相影响的阶段：**分析**思维方式，**评估**，**提高**思维方式三个阶段。
- ◆ 学习+好的思维→**卓有成效**

# Critical Thinking

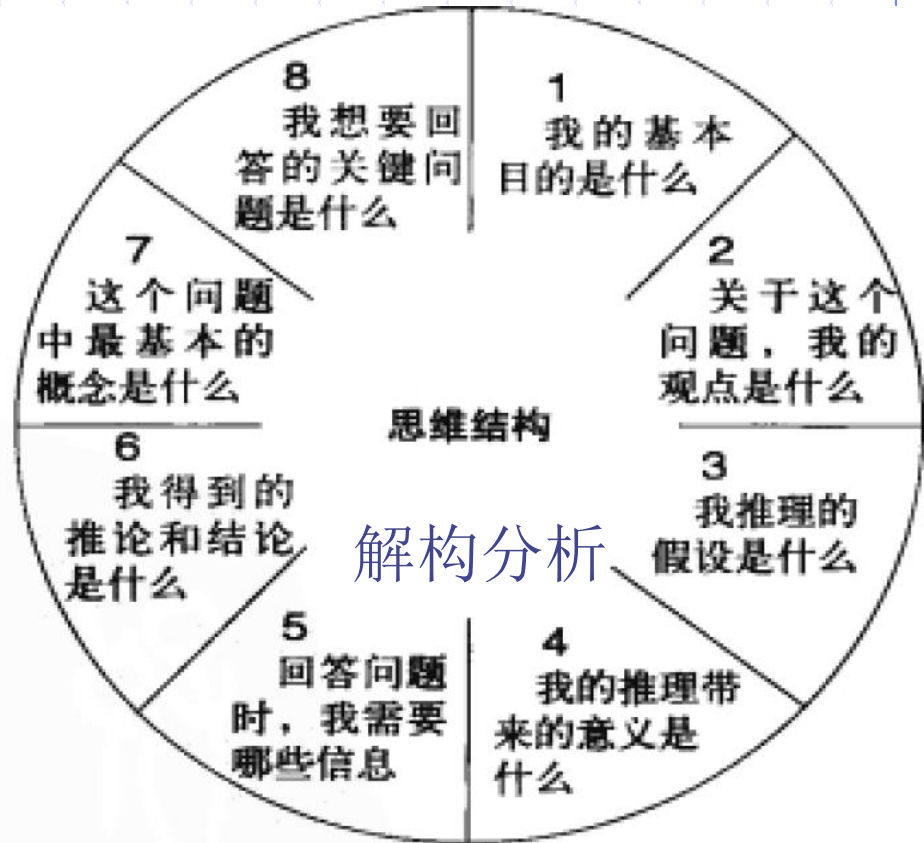
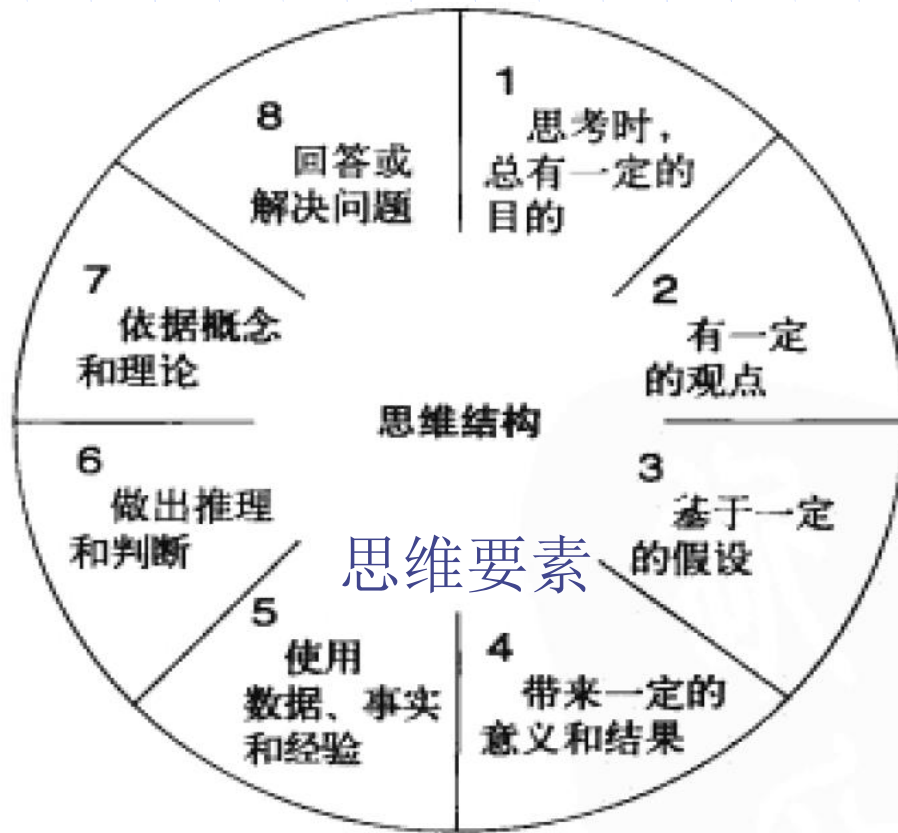
## (批判性思维)

### ◆ 批判性思维

**组成：**目的，问题，观点，信息，推理，概念，意义，假设

**特质：**思维谦逊，自主性，完整性，勇气，坚忍不拔，同理心，公正，对推理的信心。

# 批判性思维的训练与培养



应用在听课, 讨论,  
写作和阅读文章中

# Teaching method 1: Classroom Discussion

- Think-Pair-Share
  - ✓ Think: 3~5 mints
  - ✓ Pair: 5~10 mints
  - ✓ Share: 12~20 mints

# Example of Discussion Topic

**Background:** There are many seminar announcements on the campus.

**Task:** find **free food events** on campus from those announcements. 有否免费的午餐?

Wednesday, August 18

**Free Speech, Free Internet, and Free Pizza:** Join FreeThinkers@USF for free pizza and a discussion on privacy and protected speech in the era of advanced technology and social media. This event takes place in MSC 2709 from  12:30 p.m. to 2:30 p.m.

**WOW Kick-Off:** New Student Connections opens Week of Welcome with a celebration for USF students and families. Enjoy free food, music, prizes and more! This event takes place in MSC Plaza from 3:30 p.m. to 5:30 p.m.

**Chill with the Queens:** Learn about the service organization Eternal Legendary Queens while enjoying free ice cream. This event takes place in MSC 3707 from 6 p.m. to 8 p.m.

**MCWW Welcome Jam:** As part of Week of Welcome, the Office of Multicultural Affairs hosts the Multicultural Welcome Week (MCWW), which features events designed to bridge the divide between incoming and returning students from diverse cultural backgrounds. Join their welcome jam for music, performances and free food. This event takes place in MSC Ampitheatre from 9 p.m. to 12 a.m.

*Designed by Greg Woloschyn (a former CMU student)  
<http://food-bot.com> (2010-2015)*



# Food-Bot: Problem Statement

Given an arbitrary text  $d$ , determine whether  $d$  contains information about a **free food event**, and if so, return an array of correctly-associated information about each event (**date/time, location, and food type**).

Question: How to do it?

Think-pair-share

# Mistakes made by Food-bot

**Date:**

Tuesday, February 15, 2011 - 12:00pm

**Location:**

12:00 pm - 1:00 pm <br /> Stanley Hall

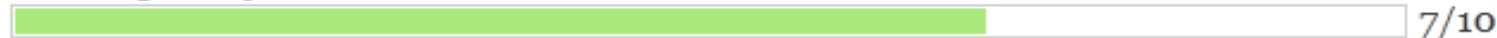
**Food Type:**

lunch

**Organization:**

unspecified

**Food Quality:**



**Food Quantity:**



**Time Commitment:**



**Awkwardness:**



Speaker/Performer: Erica Whitney QB3-Berkeley is offering a series of workshops for graduate students and postdoctoral researchers in QB3-affiliated labs on understanding the research grant process and how to write fundable proposals.

place

time

Each workshop is offered twice per week once on Tuesdays and again on Wednesday. They are held in 221 Stanley Hall from 12-1pm. Please feel free to bring a bag lunch.

Creators Rating:

# Teaching Method 2: Reading Paper & “Writing”

- ◆ **Aim:** 这篇文章目的是什么？
- ◆ **Problem:** 文章要解决什么问题？
- ◆ **Information collect:** 利用哪些信息来解决问题
- ◆ **Method:** 使用哪些概念，技术和方法来解决问题？
- ◆ **Reason & assumption :** 提出方法的理由和假设是什么？
- ◆ **Conclusion:** 最后结论是什么？

# Grading (考核标准)

◆ Classroom (30%)

Attendance+Discussion+quiz

◆ Reading paper (20%)

◆ Group Project (50%)

Coding + workshop Presentation

• No final examination

# Course Projects: two of three

**Task 1:** Relation Extraction (design algorithm)

- ◆ Corpus: News Reports
- ◆ Extract: Employment (person-organization)

**Task 2:** Sentiment Analysis (design algorithm)

- ◆ Corpus: TV Series Reviews
- ◆ Classify: positive or negative.

**Task 3:** Web page extraction (design algorithm)

- Specific site or
- Same type of web page

# Student Workshop

Each group presents their work (using **PPT** ) about the following points :

- ◆ Group member ( $\leq 4$ ) introduction.
- ◆ methods & resources introduction.
- ◆ preliminary evaluations.

What are your expectations  
from this courses?  
feedbacks are welcome!