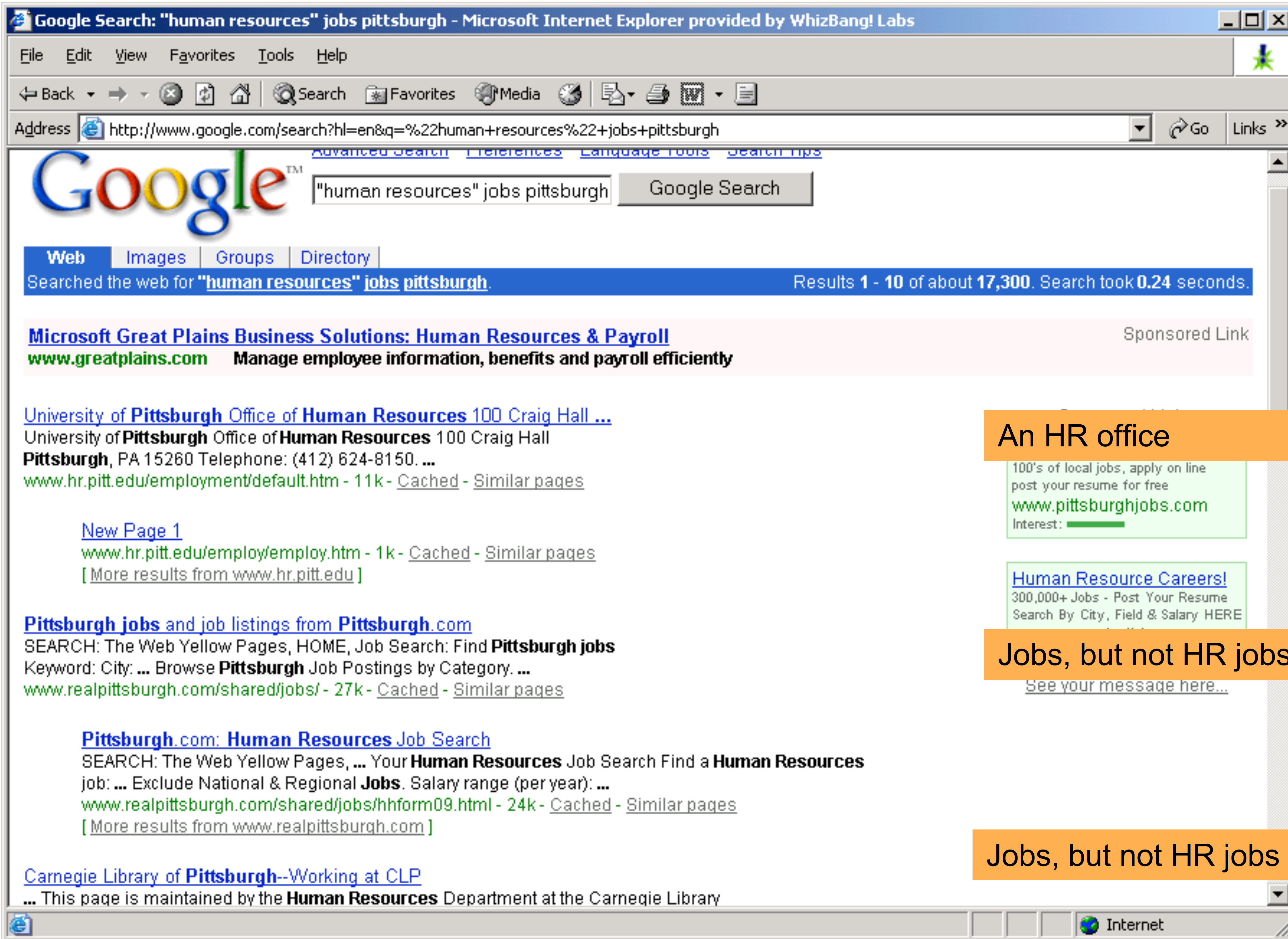
Give the People What They Want: Information Extraction Relation Extraction Question Answering

Introduction to Natural Language Processing
Computer Science 585—Fall 2009
University of Massachusetts Amherst

David Smith
With slides from Andrew McCallum, Chris Manning, Sanda Harabagiu, and Ed Hovy

Goal:

Mine actionable knowledge from unstructured text.



An HR office

Jobs, but not HR jobs

Jobs, but not HR jobs

Extracting Job Openings from the Web

IE from Research Papers

[McCallum et al '99]

Reinforcement Learning: A Survey

Leslie Pack Kaelbling LPK@CS.BROWN.I
Michael L. Littman MLITTMAN@CS.BROWN.I
Computer Science Department, Box 1910, Brown University
Providence, RI 02912-1910 USA

Andrew W. Moore AWM@CS.CMU.I
Smith Hall 221, Carnegie Mellon University, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

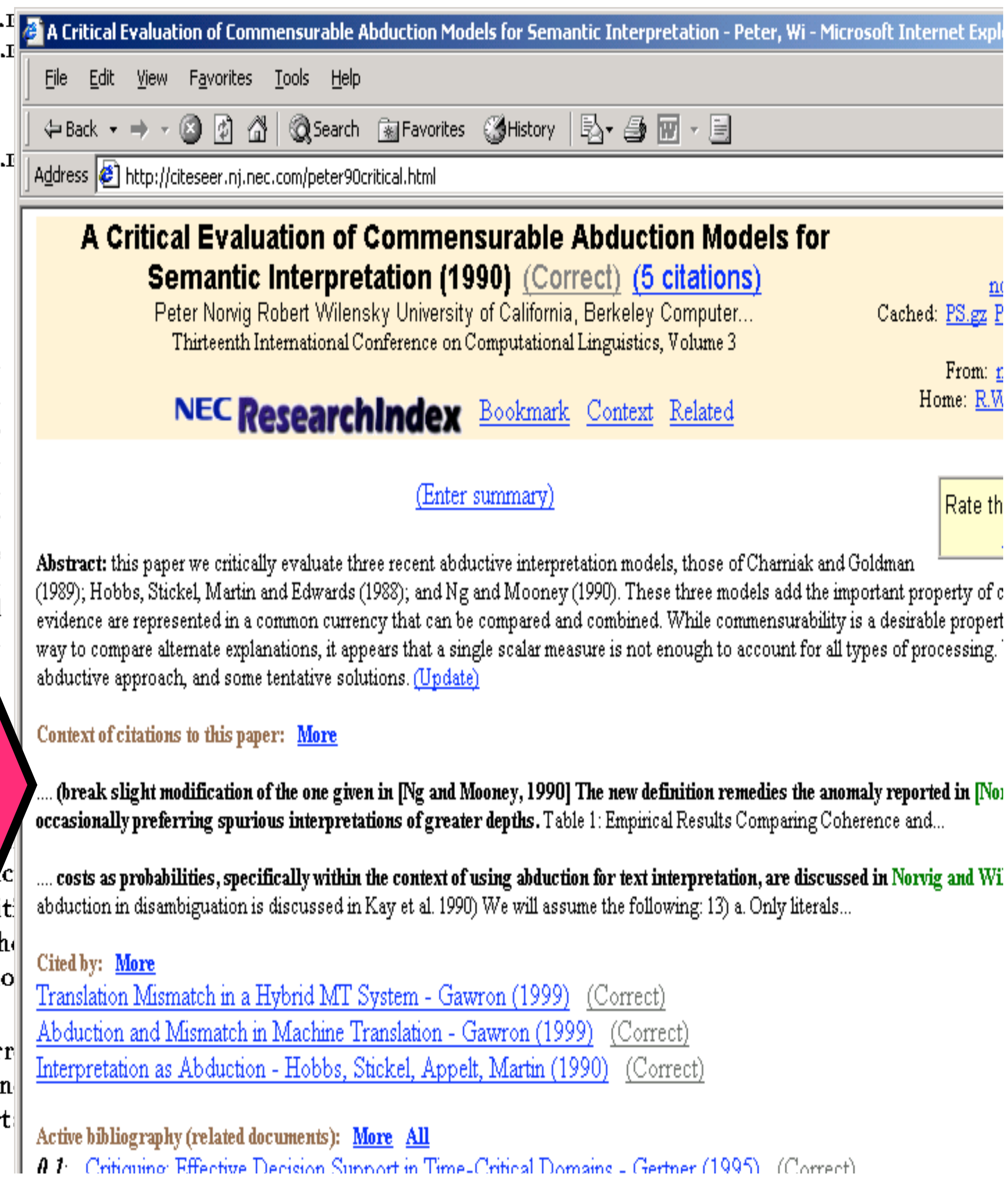
Abstract

This paper surveys the field of reinforcement learning from a computer-science perspective. It is written to be accessible to researchers familiar with machine learning. Both the historical basis of the field and a broad selection of current work are summarized. Reinforcement learning is the problem faced by an agent that learns behavior through trial-and-error interactions with a dynamic environment. The work described here has a resemblance to work in psychology, but differs considerably in the details and in the use of the word "reinforcement." The paper discusses central issues of reinforcement learning, including trading off exploration and exploitation, establishing the foundations of the field via Markov decision theory, learning from delayed reinforcement, constructing empirical models to accelerate learning, making use of generalization and hierarchy, and coping with hidden state. It concludes with a survey of some implemented systems and an assessment of the practical utility of current methods for reinforcement learning.

1. Introduction

Reinforcement learning dates back to the early days of cybernetics and work in stat psychology, neuroscience, and computer science. In the last five to ten years, it has attrac rapidly increasing interest in the machine learning and artificial intelligence communit Its promise is beguiling—a way of programming agents by reward and punishment with needing to specify *how* the task is to be achieved. But there are formidable computatio obstacles to fulfilling the promise.

This paper surveys the historical basis of reinforcement learning and some of the curr work from a computer science perspective. We give a high-level overview of the field an taste of some specific approaches. It is, of course, impossible to mention all of the import work in the field; this should not be taken to be an exhaustive account.



A Critical Evaluation of Commensurable Abduction Models for Semantic Interpretation - Peter, Wi - Microsoft Internet Expl

File Edit View Favorites Tools Help

Back Search Favorites History

Address http://citeseer.nj.nec.com/peter90critical.html

A Critical Evaluation of Commensurable Abduction Models for Semantic Interpretation (1990) (Correct) (5 citations)

Peter Norvig Robert Wilensky University of California, Berkeley Computer...
Thirteenth International Conference on Computational Linguistics, Volume 3

Cached: PS.gz P
From: r
Home: R.W

NEC ResearchIndex Bookmark Context Related

(Enter summary)

Rate th

Abstract: this paper we critically evaluate three recent abductive interpretation models, those of Charniak and Goldman (1989); Hobbs, Stickel, Martin and Edwards (1988); and Ng and Mooney (1990). These three models add the important property of c evidence are represented in a common currency that can be compared and combined. While commensurability is a desirable propert way to compare alternate explanations, it appears that a single scalar measure is not enough to account for all types of processing. abductive approach, and some tentative solutions. (Update)

Context of citations to this paper: More

.... (break slight modification of the one given in [Ng and Mooney, 1990] The new definition remedies the anomaly reported in [No occasionally preferring spurious interpretations of greater depths. Table 1: Empirical Results Comparing Coherence and...

.... costs as probabilities, specifically within the context of using abduction for text interpretation, are discussed in Norvig and Wi abduction in disambiguation is discussed in Kay et al. 1990) We will assume the following: 13) a. Only literals...

Cited by: More
Translation Mismatch in a Hybrid MT System - Gawron (1999) (Correct)
Abduction and Mismatch in Machine Translation - Gawron (1999) (Correct)
Interpretation as Abduction - Hobbs, Stickel, Appelt, Martin (1990) (Correct)

Active bibliography (related documents): More All
0.1: Critiquing: Effective Decision Support in Time-Critical Domains - Gertner (1995) (Correct)

Mining Research Papers

Most cited authors in Computer Science - June 2004 (CiteSeer.IST)

Generated from documents in the CiteSeer.IST database. This list does not inclu
where one or more authors of the citing and cited articles match, or citations whe
relevant author is an editor. An entry may correspond to multiple authors (e.g. J.
list is automatically generated and may contain errors. Citation counts may differ
results because this list is generated in batch mode whereas the database is contin
updated. A total of 703686 authors were found.

1. D. Johnson: 13216
2. J. Ullman: 11724
3. A. Gupta: 8968
4. R. Milner: 8464
5. R. Rivest: 7552
6. M. Garey: 7295
7. R. Tarjan: 7106
8. J. Dongarra: 7007
9. V. Jacobson: 6937
10. L. Lamport: 6780
11. J. Smith: 6563
12. S. Shenker: 6411
13. D. Knuth: 6352
14. E. Clarke: 6272
15. S. Floyd: 6133
16. A. Aho: 5795
17. J. Hennessy: 5759
18. R. Agrawal: 5702
19. C. Papadimitriou: 5690
20. R. Johnson: 5613
21. A. Pnueli: 5598
22. L. Zhang: 5438
23. D. Goldberg: 5414

[Rosen-Zvi, Griffiths, Steyvers, Smyth, 2004]

TOPIC 19	
WORD	**PROB.**
LIKELIHOOD	0.0539
MIXTURE	0.0509
EM	0.0470
DENSITY	0.0398
GAUSSIAN	0.0349
ESTIMATION	0.0314
LOG	0.0263
MAXIMUM	0.0254
PARAMETERS	0.0209
ESTIMATE	0.0204
AUTHOR	**PROB.**
Tresp_V	0.0333
Singer_Y	0.0281
Jebara_T	0.0207
Ghahramani_Z	0.0196
Ueda_N	0.0170
Jordan_M	0.0150
Roweis_S	0.0123

TOPIC 24	
WORD	**PROB.**
RECOGNITION	0.0400
CHARACTER	0.0336
CHARACTERS	0.0250
TANGENT	0.0241
HANDWRITTEN	0.0169
DIGITS	0.0159
IMAGE	0.0157
DISTANCE	0.0153
DIGIT	0.0149
HAND	0.0126
AUTHOR	**PROB.**
Simard_P	0.0694
Martin_G	0.0394
LeCun_Y	0.0359
Denker_J	0.0278
Henderson_D	0.0256
Revow_M	0.0229
Platt_J	0.0226

What is "Information Extraction"

As a task: **Filling slots in a database from sub-segments of text.**

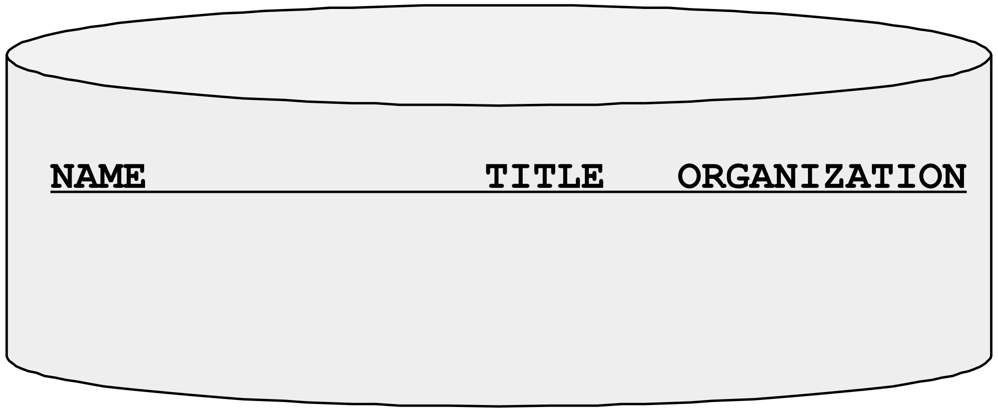
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…



What is “Information Extraction”

As a task: **Filling slots in a database from sub-segments of text.**

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

IE

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

What is "Information Extraction"

As a family of techniques:

Information Extraction =
segmentation + classification + clustering + association

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

What is "Information Extraction"

As a family of techniques:

Information Extraction = segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

What is "Information Extraction"

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation CEO Bill Gates
Microsoft Gates
Microsoft
Bill Veghte Microsoft VP
Richard Stallman founder Free Software Foundation

What is "Information Extraction"

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

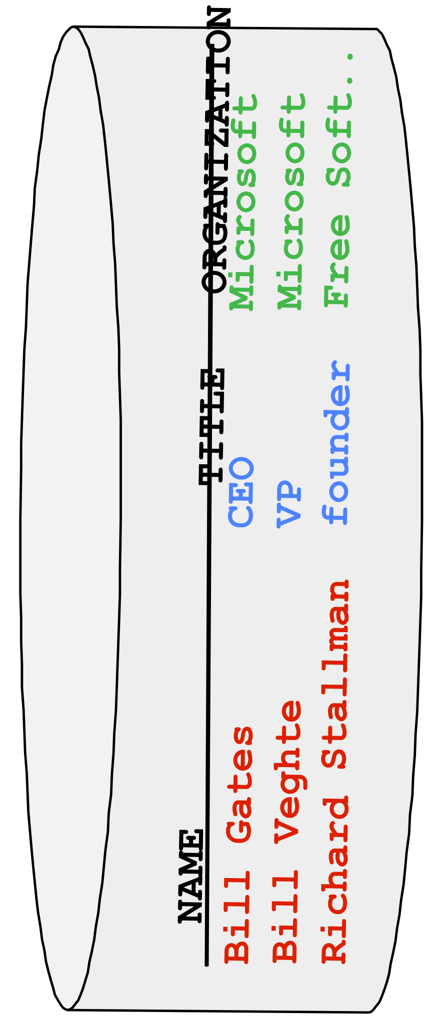
* Microsoft Corporation
CEO
Bill Gates

* Microsoft
Gates

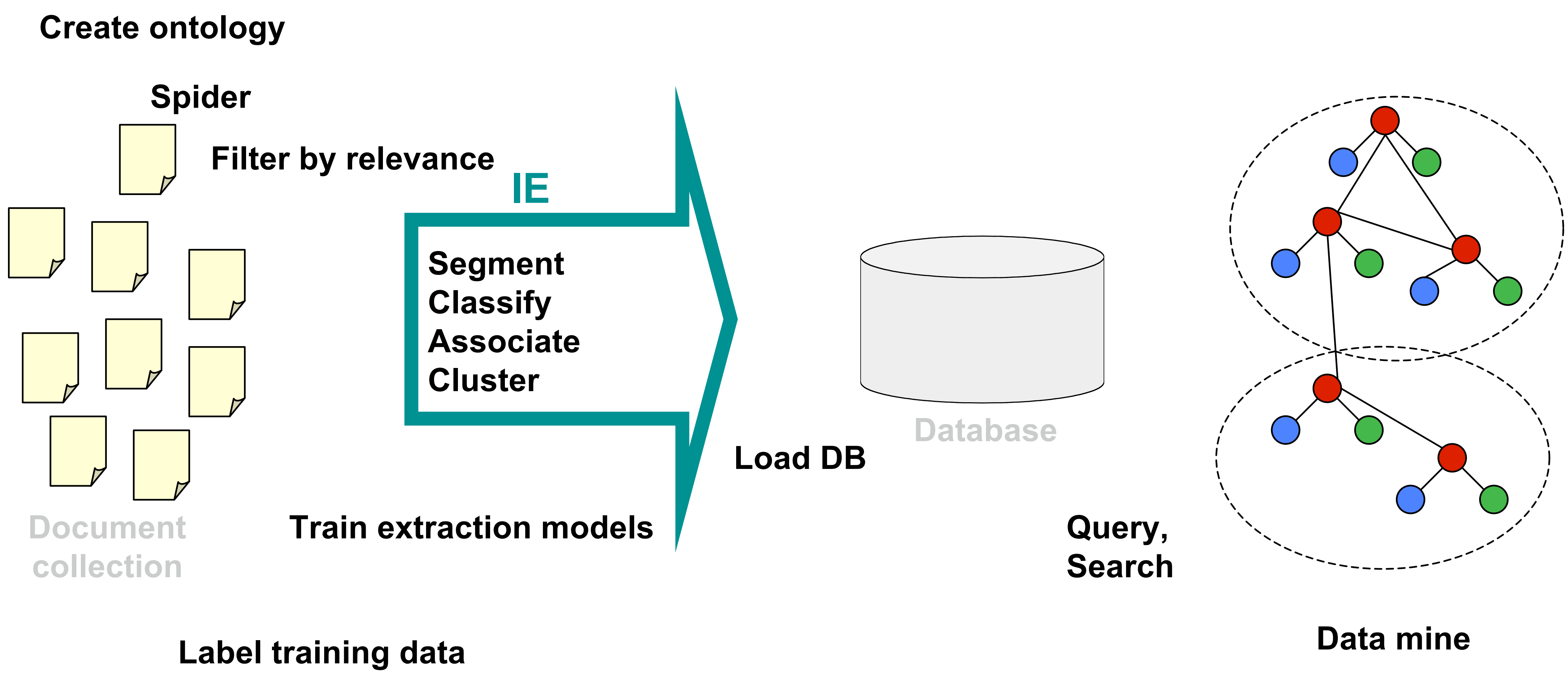
* Microsoft

Bill Veghte
* Microsoft
VP

Richard Stallman
founder
Free Software Foundation



IE in Context



Why Information Extraction (IE)?

- Science
  - Grand old dream of AI: Build large KB* and reason with it.
    IE enables the automatic creation of this KB.
  - IE is a complex problem that inspires new advances in machine learning.
- Profit
  - Many companies interested in leveraging data currently “locked in unstructured text on the Web”.
  - Not yet a monopolistic winner in this space.
- Fun!
  - Build tools that we researchers like to use ourselves:
    Cora & CiteSeer, MRQE.com, FAQFinder,…
  - See our work get used by the general public.

* KB = “Knowledge Base”

Outline

- Examples of IE and Data Mining
- $\boxed{\text{Landscape of problems and solutions}}$
- Techniques for Segmentation and Classification
  - Sliding Window and Boundary Detection
  - IE with Hidden Markov Models
  - Introduction to Conditional Random Fields (CRFs)
  - Examples of IE with CRFs
- IE + Data Mining

IE History

Pre-Web

- Mostly news articles
  - De Jong's *FRUMP* [1982]
    - Hand-built system to fill Schank-style "scripts" from news wire
  - *Message Understanding Conference (MUC)* DARPA ['87-'95], *TIPSTER* ['92-'96]
- Most early work dominated by hand-built models
  - E.g. SRI's *FASTUS*, hand-built FSMs.
  - But by 1990's, some machine learning: Lehnert, Cardie, Grishman and then HMMs: Elkan [Leek '97], BBN [Bikel et al '98]

Web

- AAAI '94 Spring Symposium on "Software Agents"
  - Much discussion of ML applied to Web. Maes, Mitchell, Etzioni.
- Tom Mitchell's WebKB, '96
  - Build KB's from the Web.
- Wrapper Induction
  - Initially hand-build, then ML: [Soderland '96], [Kushmeric '97],…

What makes IE from the Web Different?

Less grammar, but more formatting & linking

Newswire

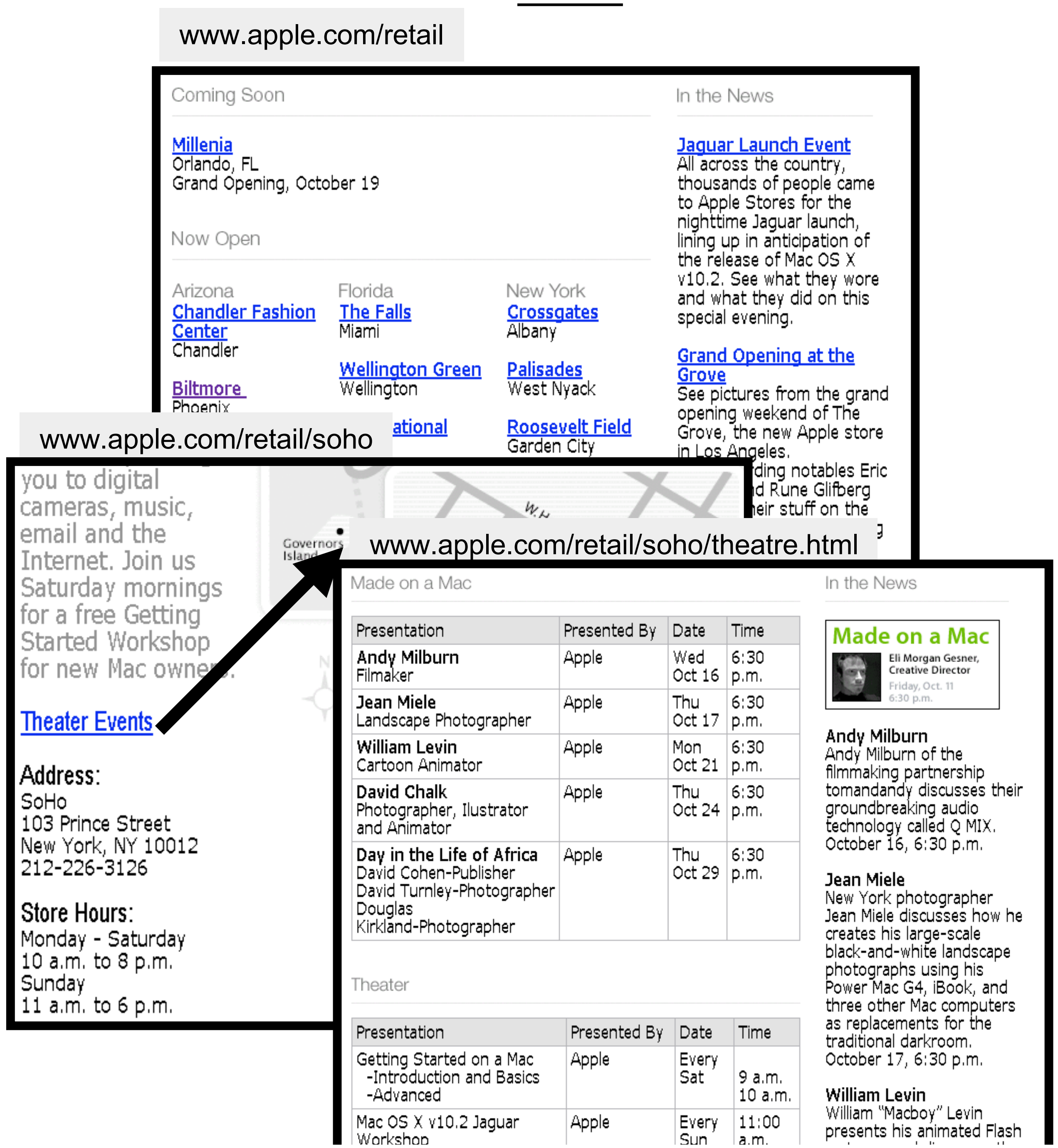
Apple to Open Its First Retail Store in New York City

MACWORLD EXPO, NEW YORK--July 17, 2002--Apple's first retail store in New York City will open in Manhattan's SoHo district on Thursday, July 18 at 8:00 a.m. EDT. The SoHo store will be Apple's largest retail store to date and is a stunning example of Apple's commitment to offering customers the world's best computer shopping experience.

"Fourteen months after opening our first retail store, our 31 stores are attracting over 100,000 visitors each week," said Steve Jobs, Apple's CEO. "We hope our SoHo store will surprise and delight both Mac and PC users who want to see everything the Mac can do to enhance their digital lifestyles."

The directory structure, link structure, formatting & layout of the Web is its own new grammar.

Web



Evaluation of Single Entity Extraction

TRUTH:

Michael Kearns and Sebastian Seung will start Monday’s tutorial, followed by Richard M. Karpe and Martin Cooke.

PRED:

Michael Kearns and Sebastian Seung will start Monday’s tutorial, followed by Richard M. Karpe and Martin Cooke.

$$\text{Precision} = \frac{\text{\# correctly predicted segments}}{\text{\# predicted segments}} = \frac{2}{6}$$

$$\text{Recall} = \frac{\text{\# correctly predicted segments}}{\text{\# true segments}} = \frac{2}{4}$$

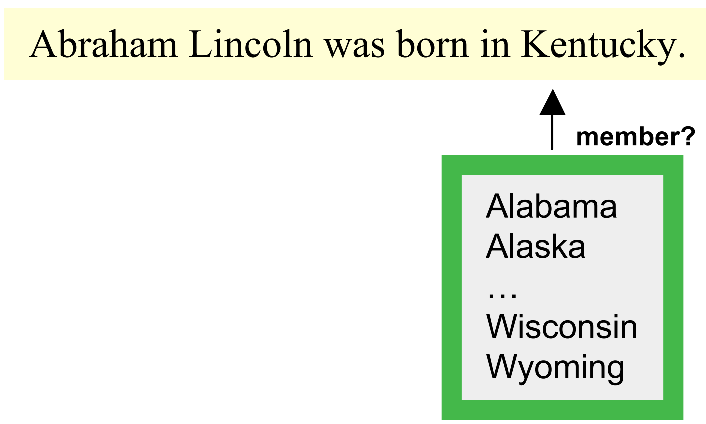
$$\text{F1} = \text{Harmonic mean of Precision \& Recall} = \frac{1}{((1/P) + (1/R)) / 2}$$

State of the Art Performance

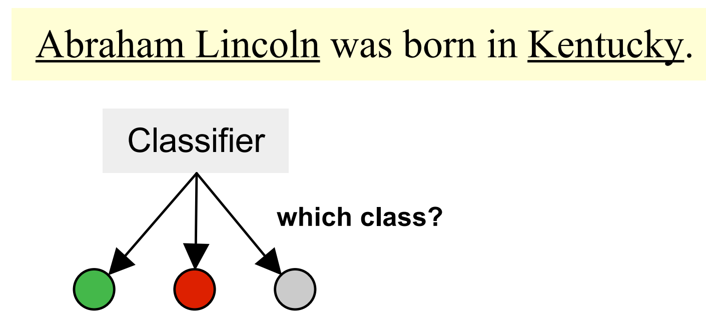
- Named entity recognition
  - Person, Location, Organization, …
  - F1 in high 80's or low- to mid-90's
- Binary relation extraction
  - Contained-in (Location1, Location2)
    Member-of (Person1, Organization1)
  - F1 in 60's or 70's or 80's
- Wrapper induction
  - Extremely accurate performance obtainable
  - Human effort (~30min) required on each site

Landscape of IE Techniques (1/1): Models

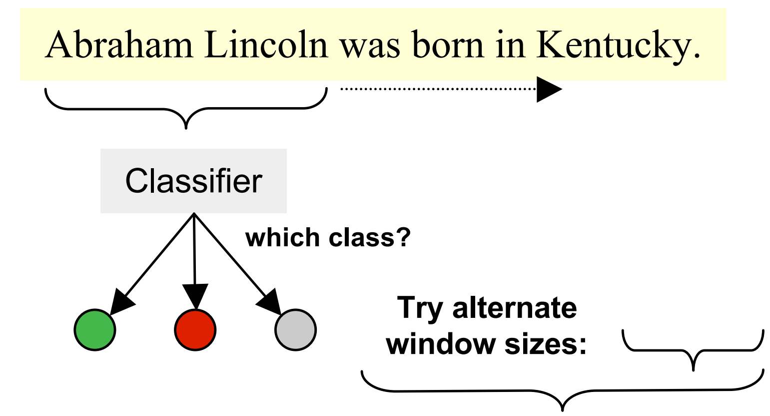
Lexicons



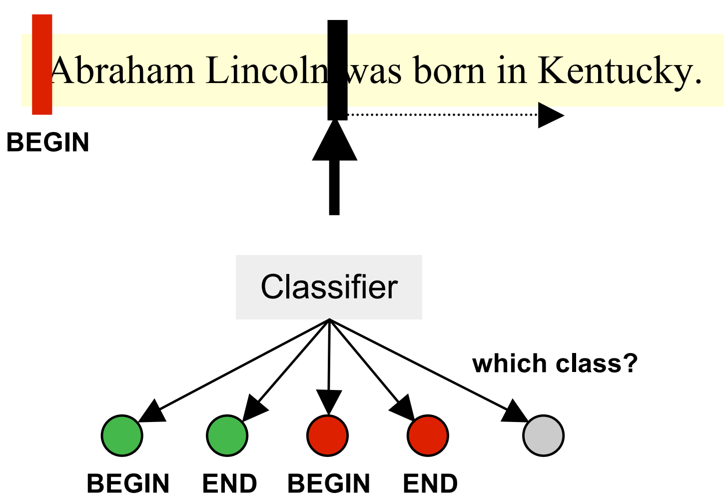
Classify Pre-segmented Candidates



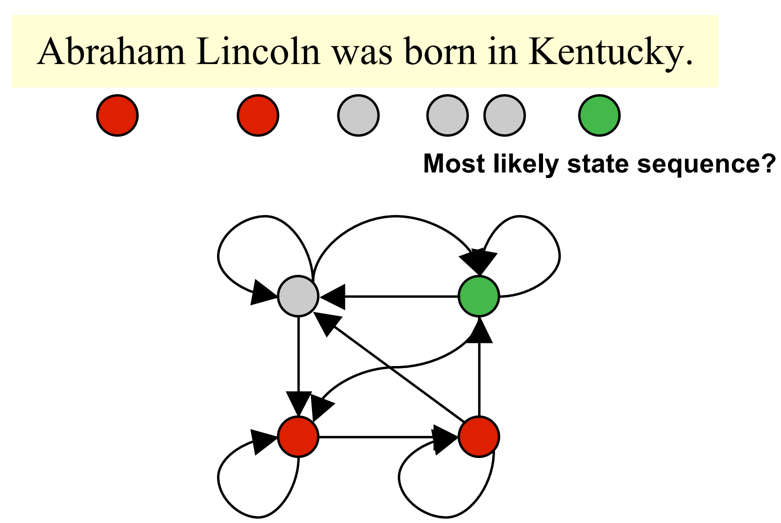
Sliding Window



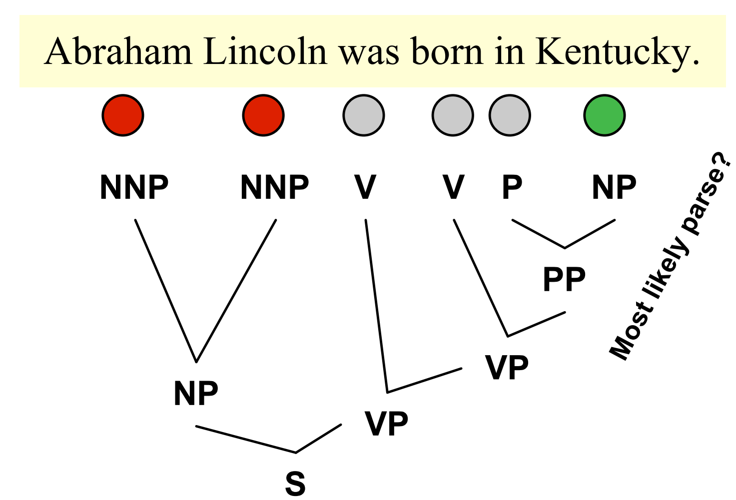
Boundary Models



Finite State Machines



Context Free Grammars



…and beyond

Any of these models can be used to capture words, formatting or both.

Table Extraction from Government Reports

Cash receipts from marketings of milk during 1995 at $19.9 billion dollars, was slightly below 1994. Producer returns averaged $12.93 per hundredweight, $0.19 per hundredweight below 1994. Marketings totaled 154 billion pounds, 1 percent above 1994. Marketings include whole milk sold to plants and dealers as well as milk sold directly to consumers.

An estimated 1.56 billion pounds of milk were used on farms where produced, 8 percent less than 1994. Calves were fed 78 percent of this milk with the remainder consumed in producer households.

Milk Cows and Production of Milk and Milkfat:
United States, 1993-95

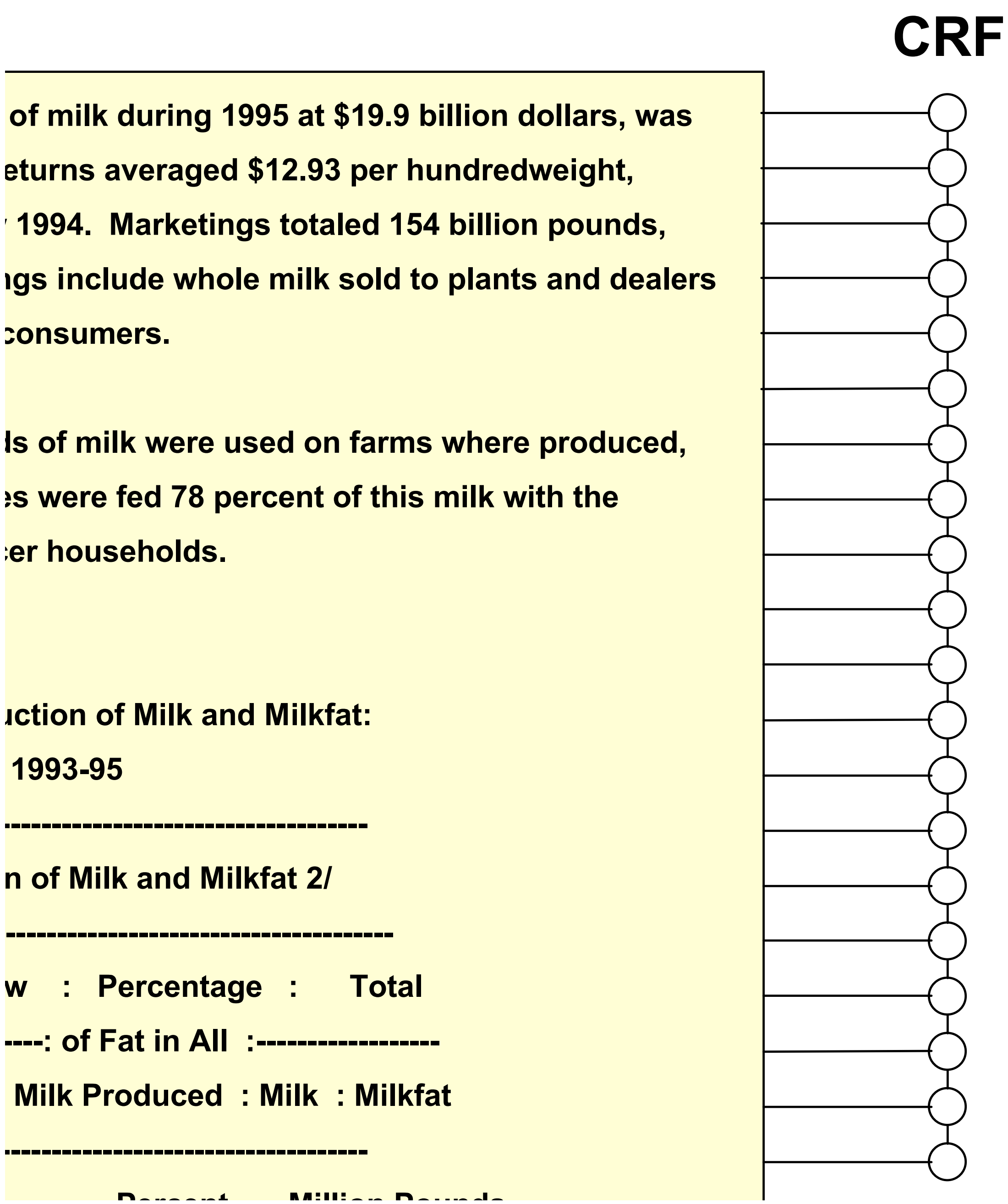
Year	Number of Milk Cows 1/	Production of Milk and Milkfat 2/				
		Per Milk Cow		Percentage of Fat in All Milk Produced	Total	
		Milk	Milkfat		Milk	Milkfat
	1,000 Head	--- Pounds ---		Percent	Million Pounds	
1993	9,589	15,704	575	3.66	150,582	5,514.4
1994	9,500	16,175	592	3.66	153,664	5,623.7
1995	9,461	16,451	602	3.66	155,644	5,694.3

1/ Average number during year, excluding heifers not yet fresh.
2/ Excludes milk sucked by calves.

Table Extraction from Government Reports

[Pinto, McCallum, Wei, Croft, 2003 SIGIR]

100+ documents from www.fedstats.gov



Labels:

- Non-Table
- Table Title
- Table Header
- Table Data Row
- Table Section Data Row
- Table Footnote
- ... *(12 in all)*

Features:

- Percentage of digit chars
- Percentage of alpha chars
- Indented
- Contains 5+ consecutive spaces
- Whitespace in this line aligns with prev.
- ...
- Conjunctions of all previous features, time offset: {0,0}, {-1,0}, {0,1}, {1,2}.

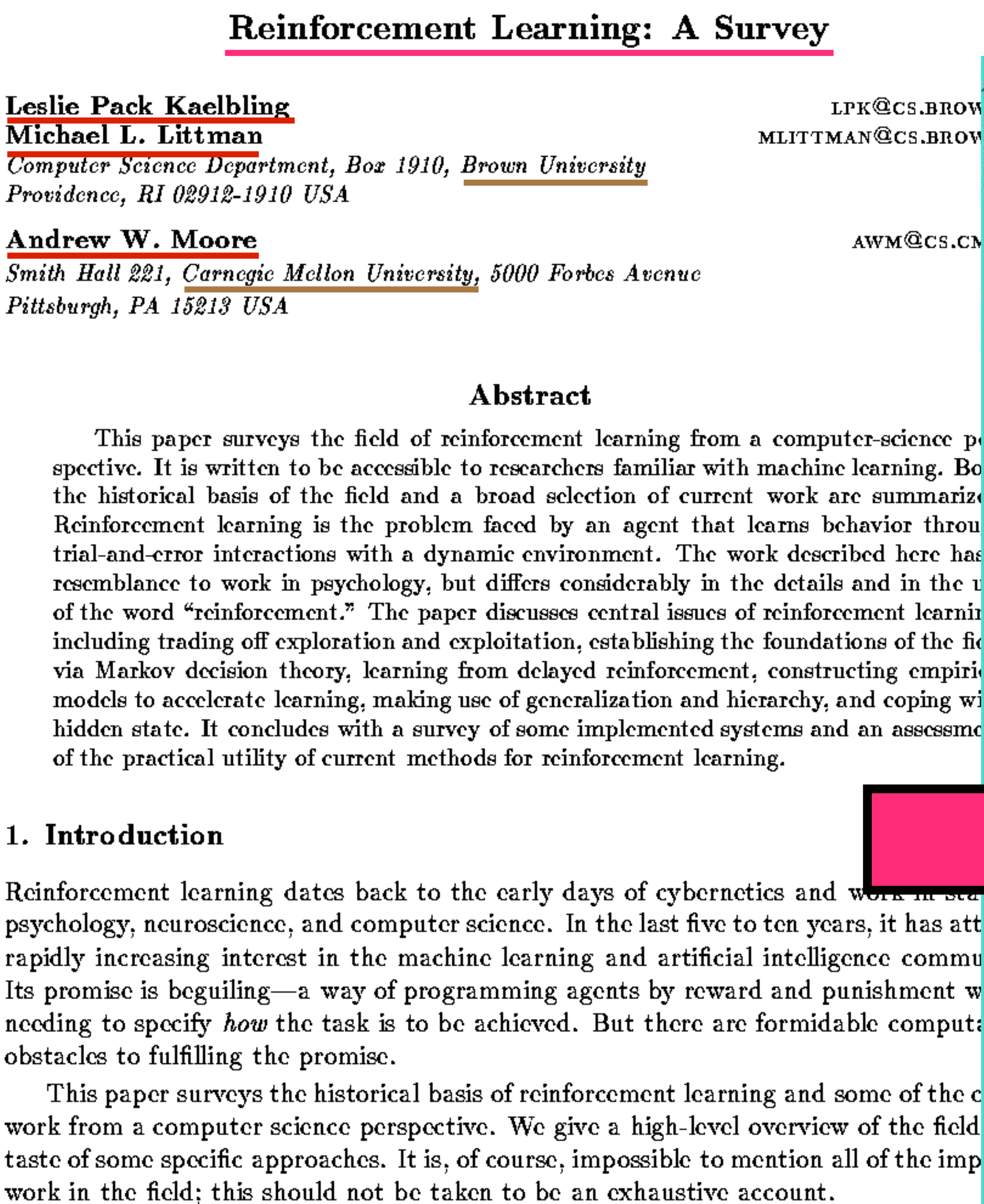
Table Extraction Experimental Results

[Pinto, McCallum, Wei, Croft, 2003 SIGIR]

	Line labels, percent correct	Table segments, F1
HMM	65 %	64 %
Stateless MaxEnt	85 %	-
CRF	95 %	92 %

IE from Research Papers

[McCallum et al '99]



Reinforcement Learning: A Survey

Leslie Pack Kaelbling LPK@CS.BROW
Michael L. Littman MLITTMAN@CS.BROW
Computer Science Department, Box 1910, Brown University
Providence, RI 02912-1910 USA

Andrew W. Moore AWM@CS.CM
Smith Hall 221, Carnegie Mellon University, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

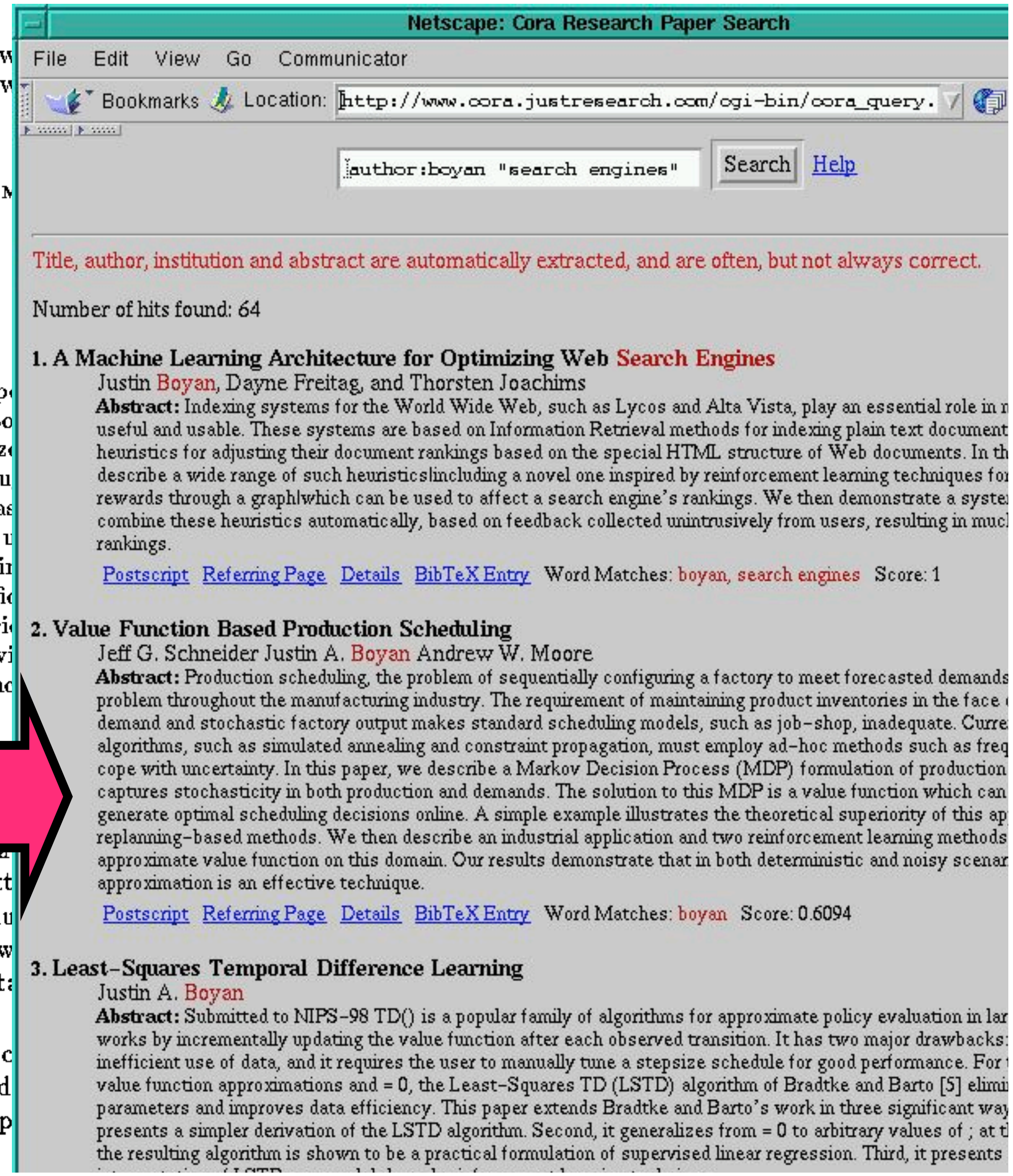
Abstract

This paper surveys the field of reinforcement learning from a computer-science p spective. It is written to be accessible to researchers familiar with machine learning. Bo the historical basis of the field and a broad selection of current work are summariz Reinforcement learning is the problem faced by an agent that learns behavior throu trial-and-error interactions with a dynamic environment. The work described here has resemblance to work in psychology, but differs considerably in the details and in the u of the word "reinforcement." The paper discusses central issues of reinforcement learnin including trading off exploration and exploitation, establishing the foundations of the fi via Markov decision theory, learning from delayed reinforcement, constructing empiri models to accelerate learning, making use of generalization and hierarchy, and coping wi hidden state. It concludes with a survey of some implemented systems and an assessme of the practical utility of current methods for reinforcement learning.

1. Introduction

Reinforcement learning dates back to the early days of cybernetics and work in sta psychology, neuroscience, and computer science. In the last five to ten years, it has att rapidly increasing interest in the machine learning and artificial intelligence commu Its promise is beguiling—a way of programming agents by reward and punishment w needing to specify *how* the task is to be achieved. But there are formidable comput obstacles to fulfilling the promise.

This paper surveys the historical basis of reinforcement learning and some of the c work from a computer science perspective. We give a high-level overview of the field taste of some specific approaches. It is, of course, impossible to mention all of the imp work in the field; this should not be taken to be an exhaustive account.



IE from Research Papers

	Field-level F1
Hidden Markov Models (HMMs) *[Seymore, McCallum, Rosenfeld, 1999]*	**75.6**
Support Vector Machines (SVMs) *[Han, Giles, et al, 2003]*	**89.7**
Conditional Random Fields (CRFs) *[Peng, McCallum, 2004]*	**93.9**

Δ error 40%

Named Entity Recognition

CRICKET -
MILLNS SIGNS FOR BOLAND

CAPE TOWN 1996-08-22

South African provincial side Boland said on Thursday they had signed Leicestershire fast bowler David Millns on a one year contract.
Millns, who toured Australia with England A in 1992, replaces former England all-rounder Phillip DeFreitas as Boland's overseas professional.

Labels:	Examples:
PER	Yayuk Basuki Innocent Butare
ORG	3M KDP Cleveland
LOC	Cleveland Nirmal Hriday The Oval
MISC	Java Basque 1,000 Lakes Rally

Automatically Induced Features

[McCallum & Li, 2003, CoNLL]

Index	*Feature*
0	inside-noun-phrase (o_{t-1})
5	stopword (o_t)
20	capitalized (o_{t+1})
75	word=the (o_t)
100	in-person-lexicon (o_{t-1})
200	word=in (o_{t+2})
500	word=Republic (o_{t+1})
711	word=RBI (o_t) & header=BASEBALL
1027	header=CRICKET (o_t) & in-English-county-lexicon (o_t)
1298	company-suffix-word ($\text{firstmention}_{t+2}$)
4040	location (o_t) & POS=NNP (o_t) & capitalized (o_t) & stopword (o_{t-1})
4945	moderately-rare-first-name (o_{t-1}) & very-common-last-name (o_t)
4474	word=the (o_{t-2}) & word=of (o_t)

Named Entity Extraction Results

[McCallum & Li, 2003, CoNLL]

Method	F1
HMMs BBN's Identifinder	73%
CRFs w/out Feature Induction	83%
CRFs with Feature Induction based on LikelihoodGain	90%

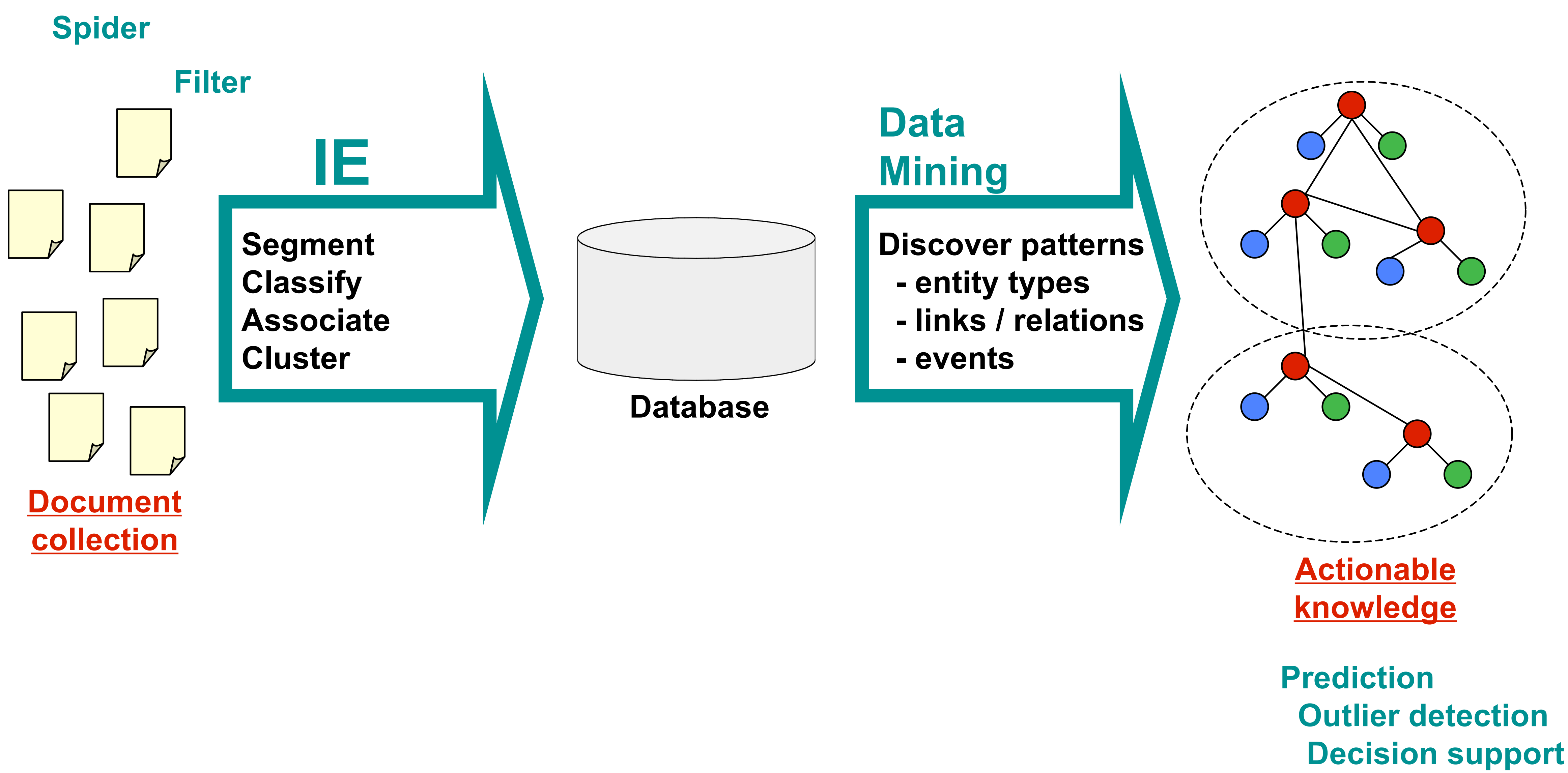
Related Work

- CRFs are widely used for information extraction ...including more complex structures, like trees:
  - [Zhu, Nie, Zhang, Wen, ICML 2007] Dynamic Hierarchical Markov Random Fields and their Application to Web Data Extraction
  - [Viola & Narasimhan]: Learning to Extract Information from Semi-structured Text using a Discriminative Context Free Grammar
  - [Jousse et al 2006]: Conditional Random Fields for XML Trees

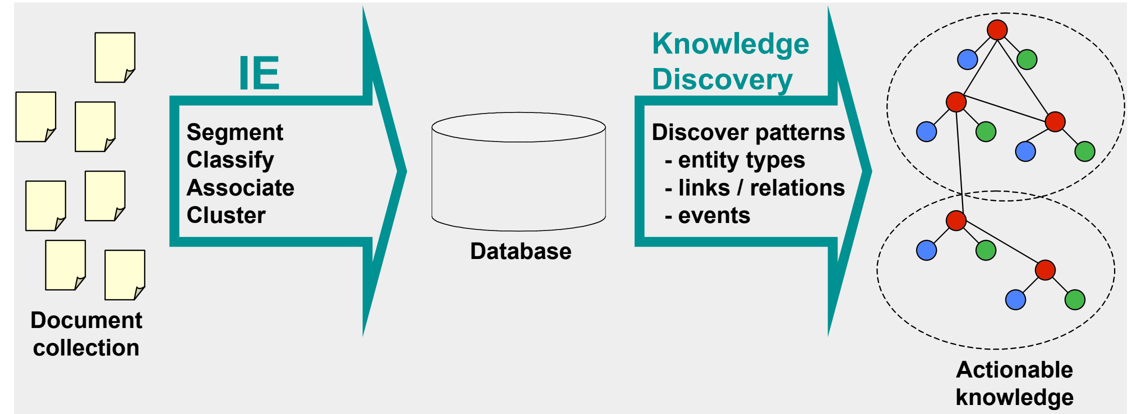
Outline

- Examples of IE and Data Mining
- Landscape of problems and solutions
- Techniques for Segmentation and Classification
  - Sliding Window and Boundary Detection
  - IE with Hidden Markov Models
  - Introduction to Conditional Random Fields (CRFs)
  - Examples of IE with CRFs
- IE + Data Mining

From Text to Actionable Knowledge



Problem:

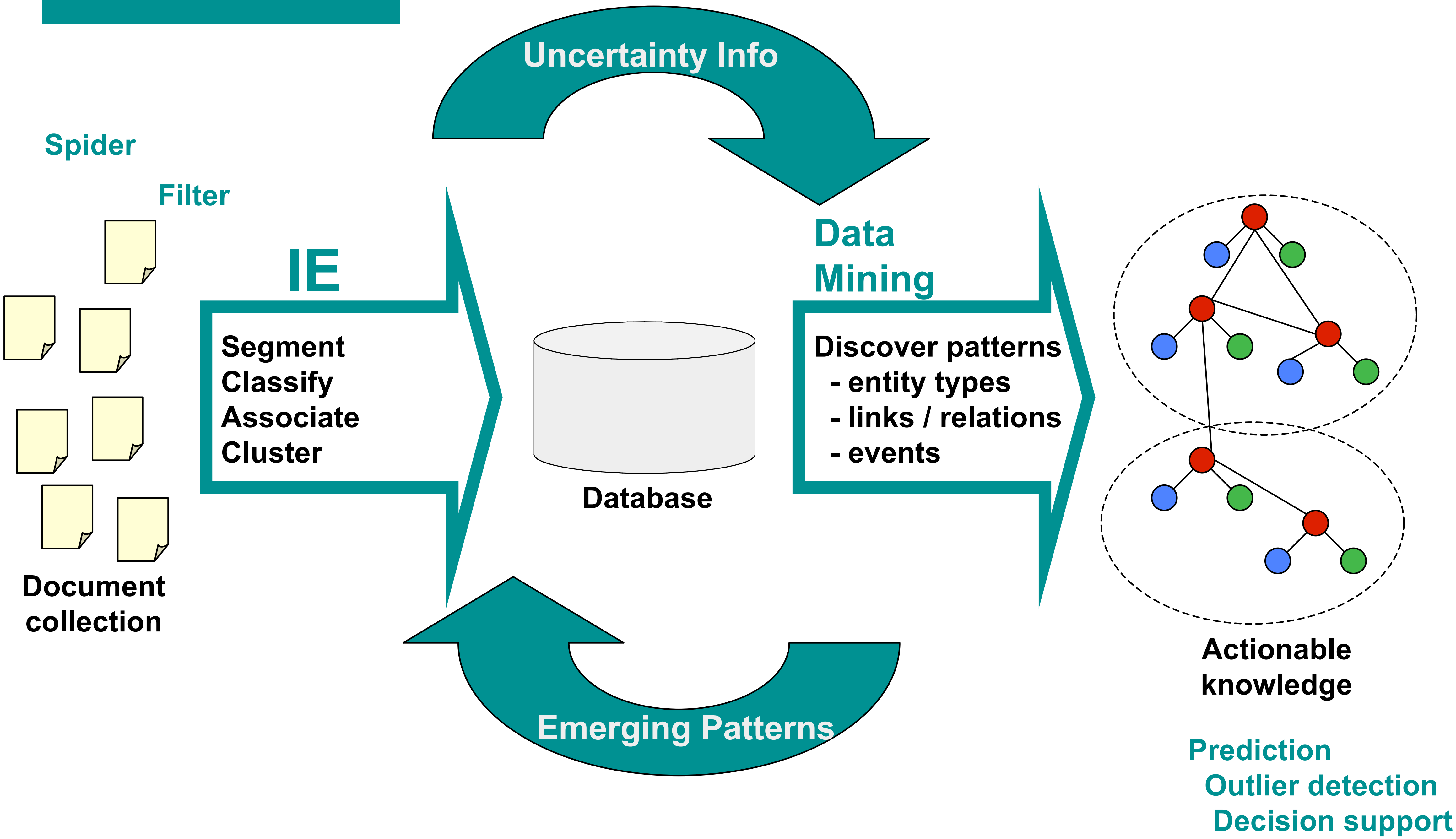


Combined in serial juxtaposition, IE and DM are unaware of each others' weaknesses and opportunities.

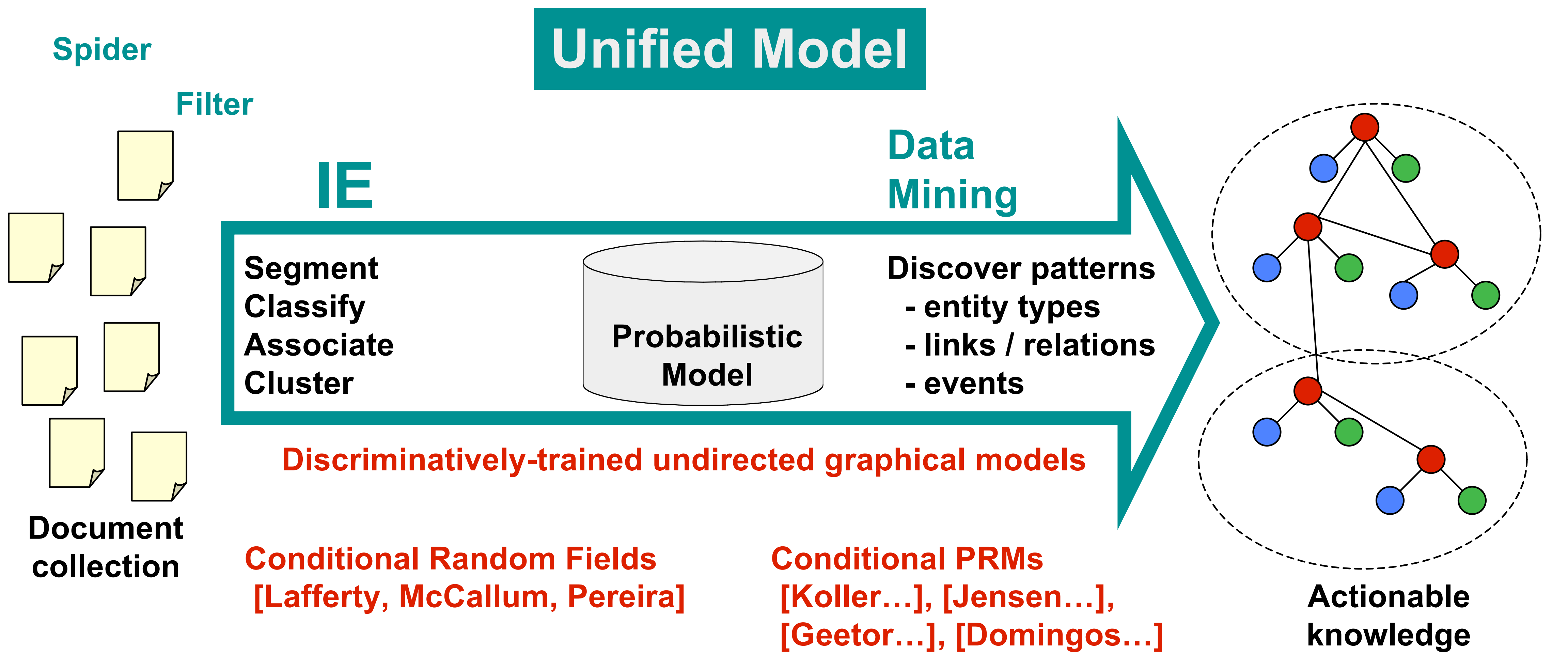
1) DM begins from a populated DB, unaware of where the data came from, or its inherent errors and uncertainties.
2) IE is unaware of emerging patterns and regularities in the DB.

The accuracy of both suffers, and significant mining of complex text sources is beyond reach.

Solution:


Uncertainty Info
Spider
Filter
IE
Segment
Classify
Associate
Cluster
Document
collection
Database
Data
Mining
Discover patterns
- entity types
- links / relations
- events
Actionable
knowledge
Prediction
Outlier detection
Decision support
Emerging Patterns

Solution:



Complex Inference and Learning

Just what we researchers like to sink our teeth into!

Prediction
Outlier detection
Decision support

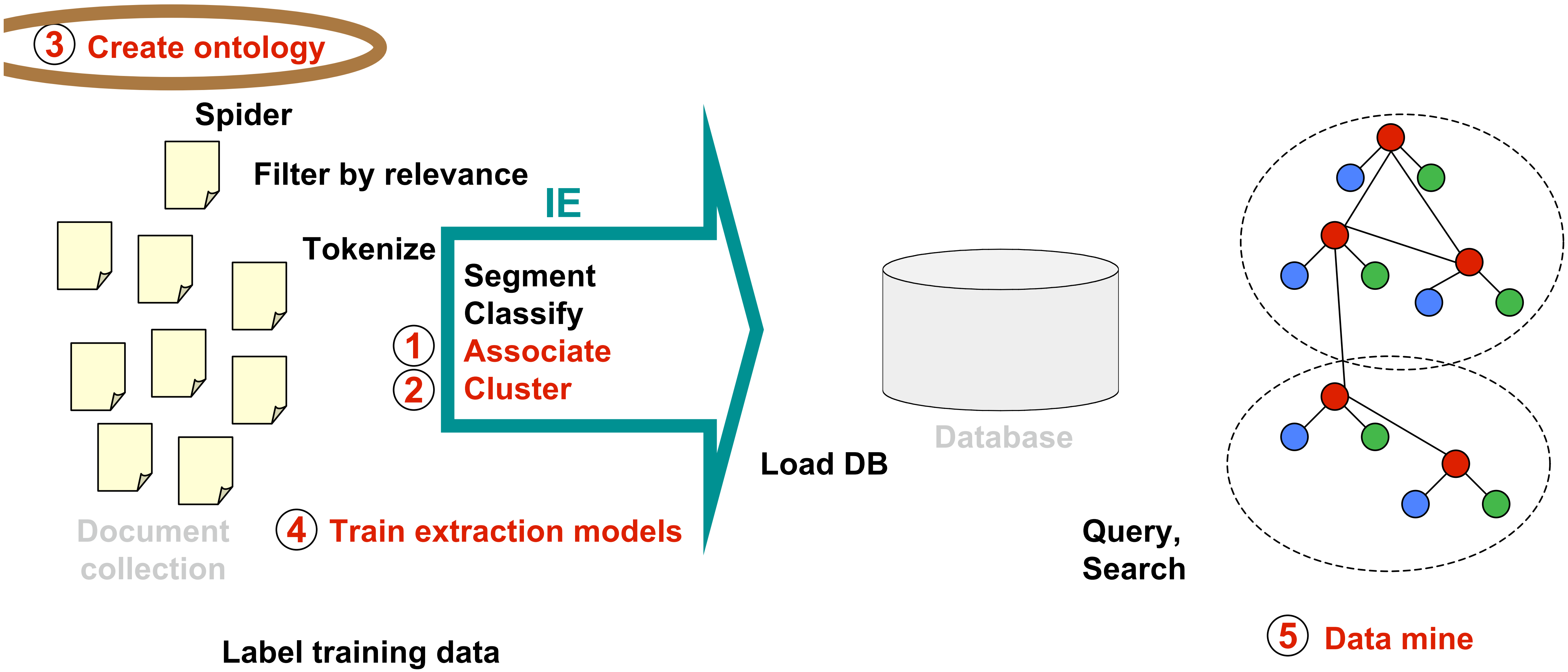
Scientific Questions

- What model structures will capture salient dependencies?
- Will joint inference actually improve accuracy?

- How to do ***inference*** in these large graphical models?
- How to do ***parameter estimation*** efficiently in these models, which are built from multiple large components?
- How to do ***structure discovery*** in these models?

Broader View

Now touch on some other issues



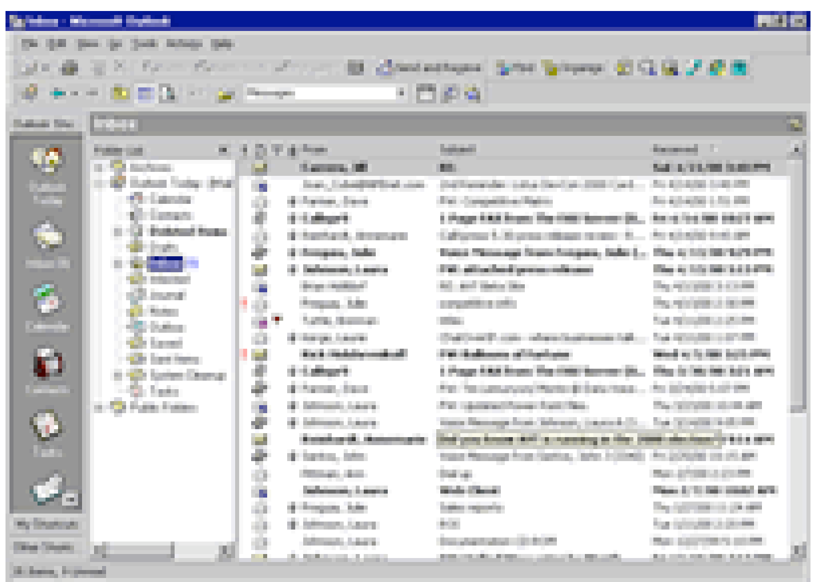
1

Managing and Understanding Connections of People in our Email World

Workplace effectiveness ~ Ability to leverage network of acquaintances

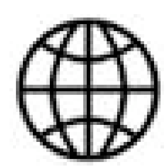
But filling Contacts DB by hand is tedious, and incomplete.

Email Inbox

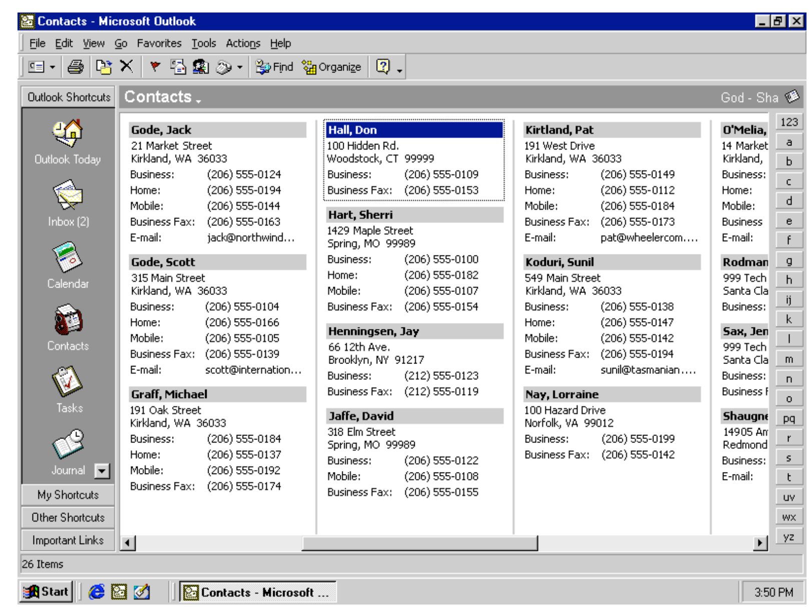


Automatically

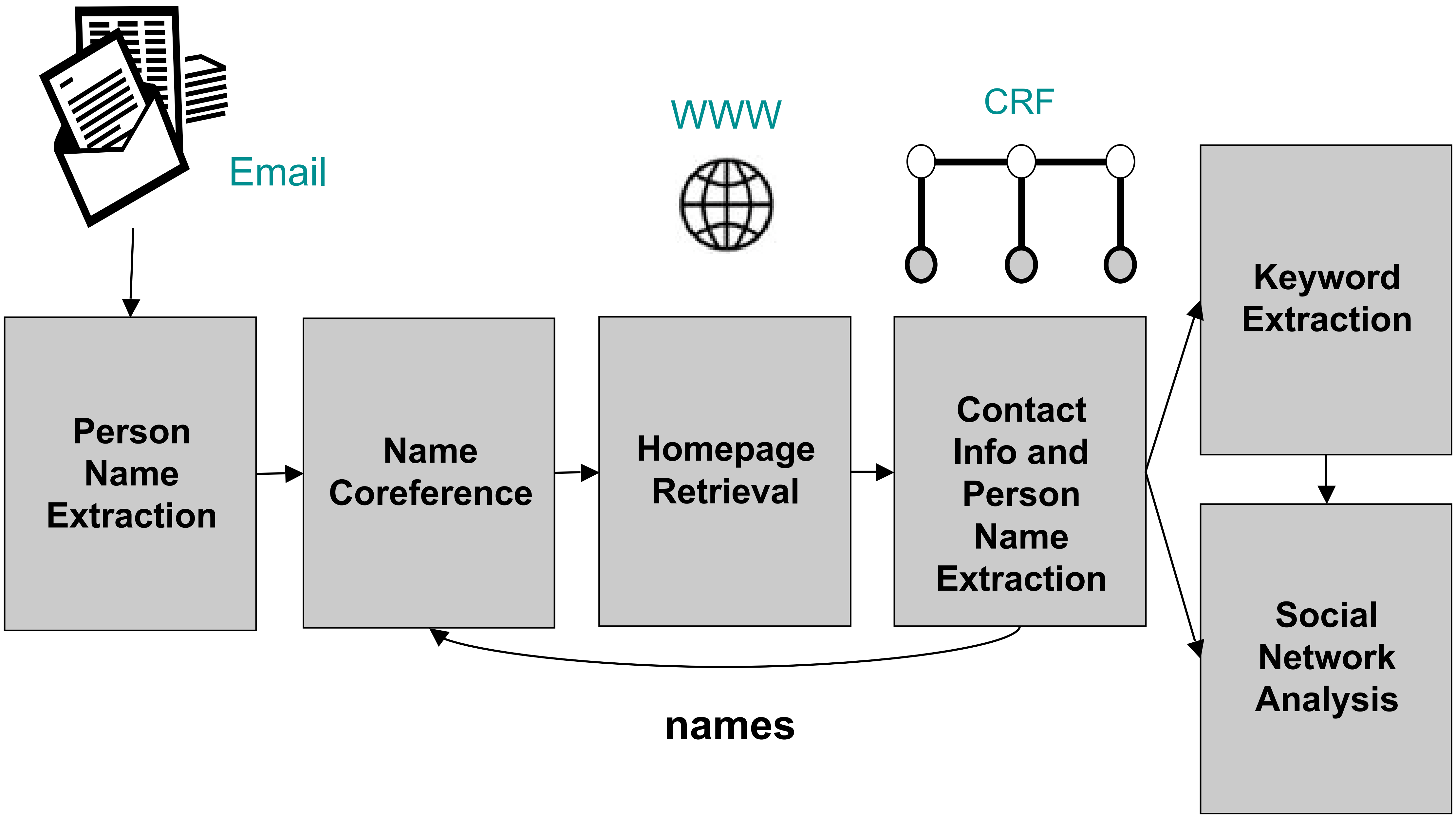
WWW



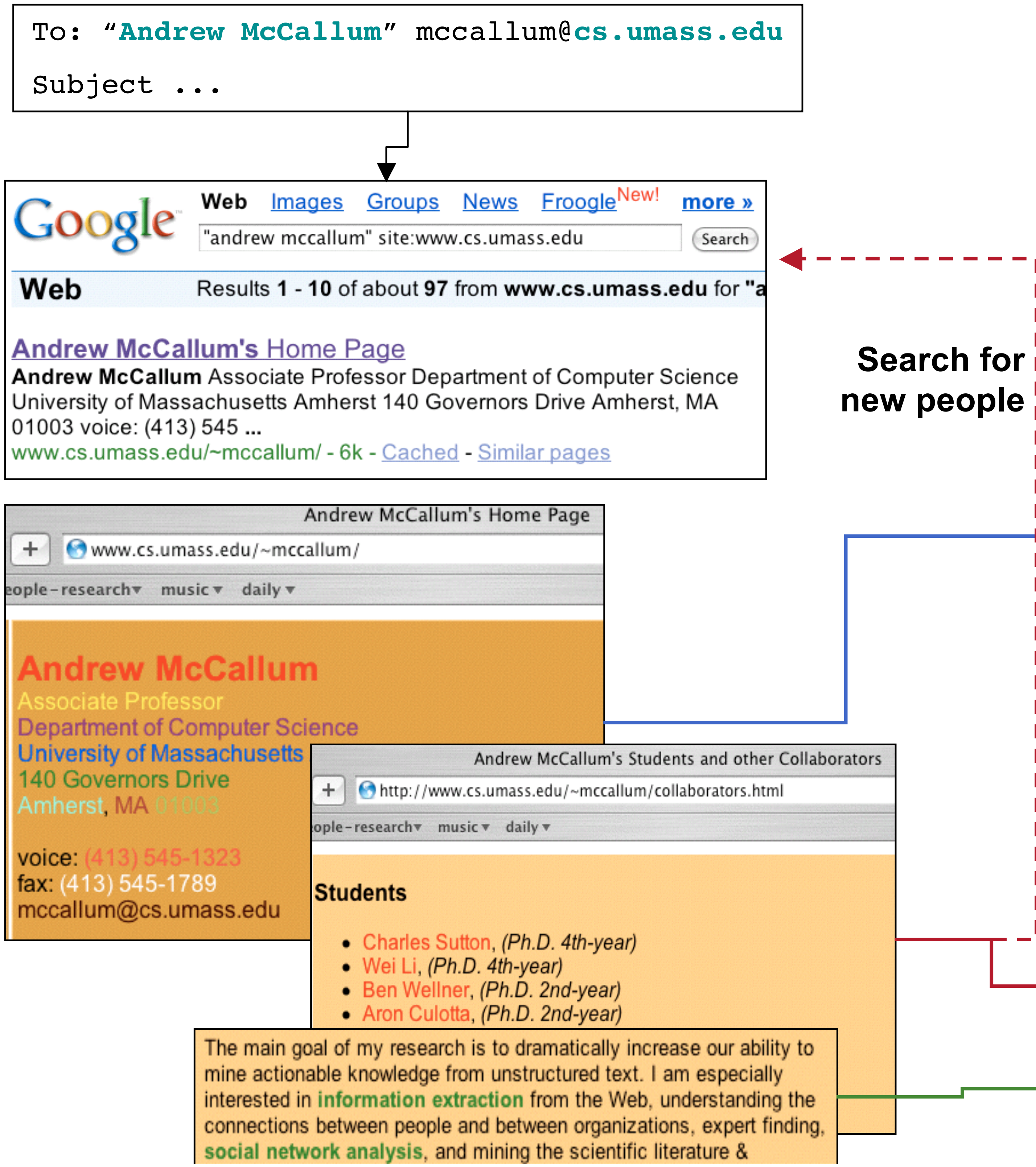
Contacts DB



System Overview



An Example



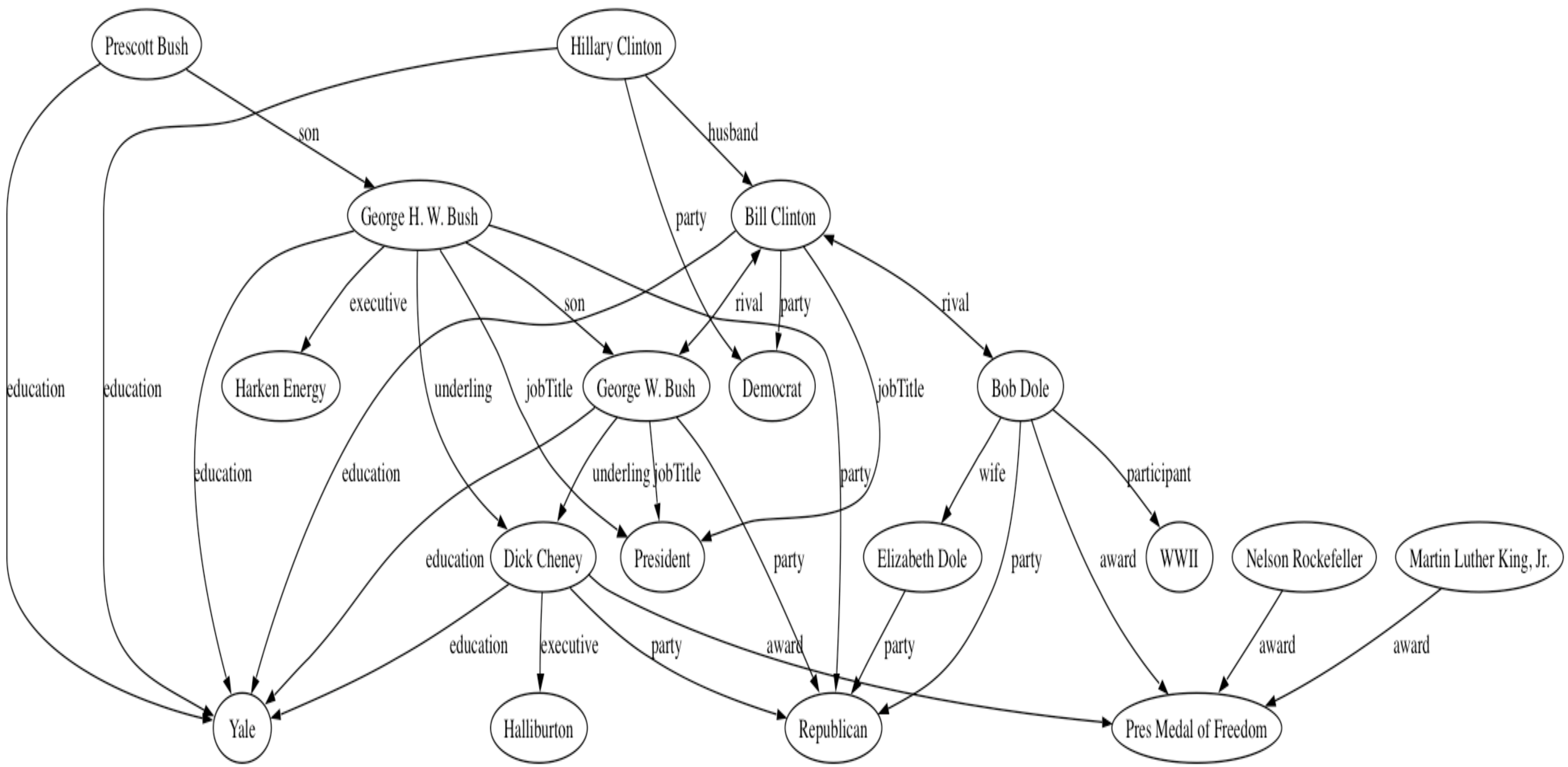
First Name:	Andrew
Middle Name:	Kachites
Last Name:	McCallum
JobTitle:	Associate Professor
Company:	University of Massachusetts
Street Address:	140 Governor's Dr.
City:	Amherst
State:	MA
Zip:	01003
Company Phone:	(413) 545-1323
Links:	Fernando Pereira, Sam Roweis,…
Key Words:	Information extraction, social network,…

Relation Extraction - Data

- 270 Wikipedia articles
- 1000 paragraphs
- 4700 relations

- 52 relation types
  - JobTitle, BirthDay, Friend, Sister, Husband, Employer, Cousin, Competition, Education, …

- Targeted for density of relations
  - Bush/Kennedy/Manning/Coppola families and friends


Prescott Bush
Hillary Clinton
son
husband
George H. W. Bush
party
Bill Clinton
executive
son
rival
party
rival
education
education
Harken Energy
underling
jobTitle
George W. Bush
Democrat
jobTitle
Bob Dole
education
education
underling
jobTitle
party
wife
participant
education
Dick Cheney
President
party
Elizabeth Dole
party
award
WWII
Nelson Rockefeller
Martin Luther King, Jr.
education
executive
party
award
party
award
award
Yale
Halliburton
Republican
Pres Medal of Freedom

George W. Bush

…his father George H. W. Bush…

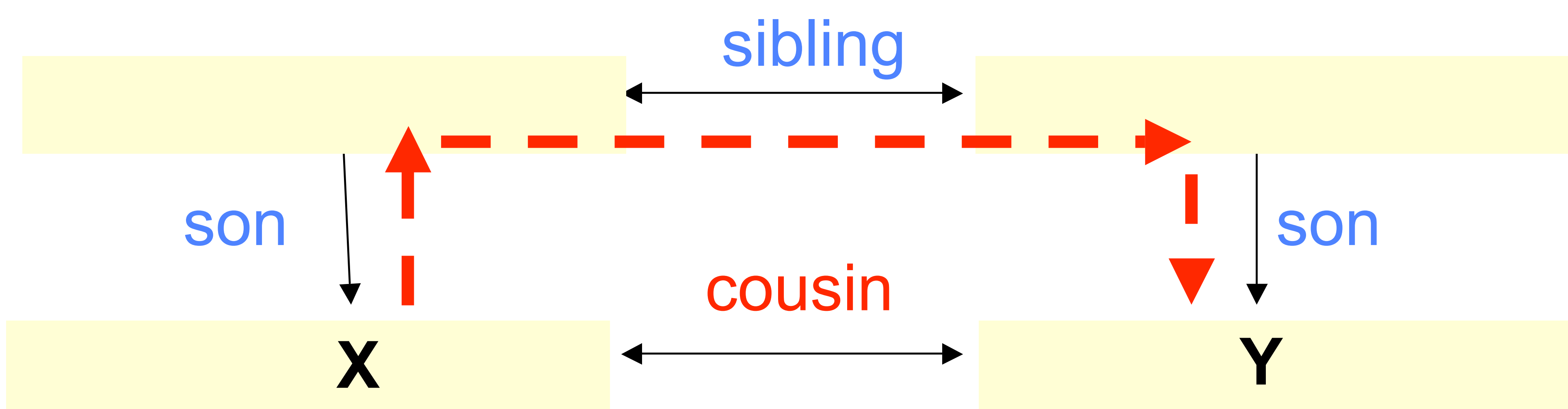
George H. W. Bush

…his sister Nancy Ellis Bush…

Nancy Ellis Bush

…her son John Prescott Ellis…

Cousin = Father's Sister's Son

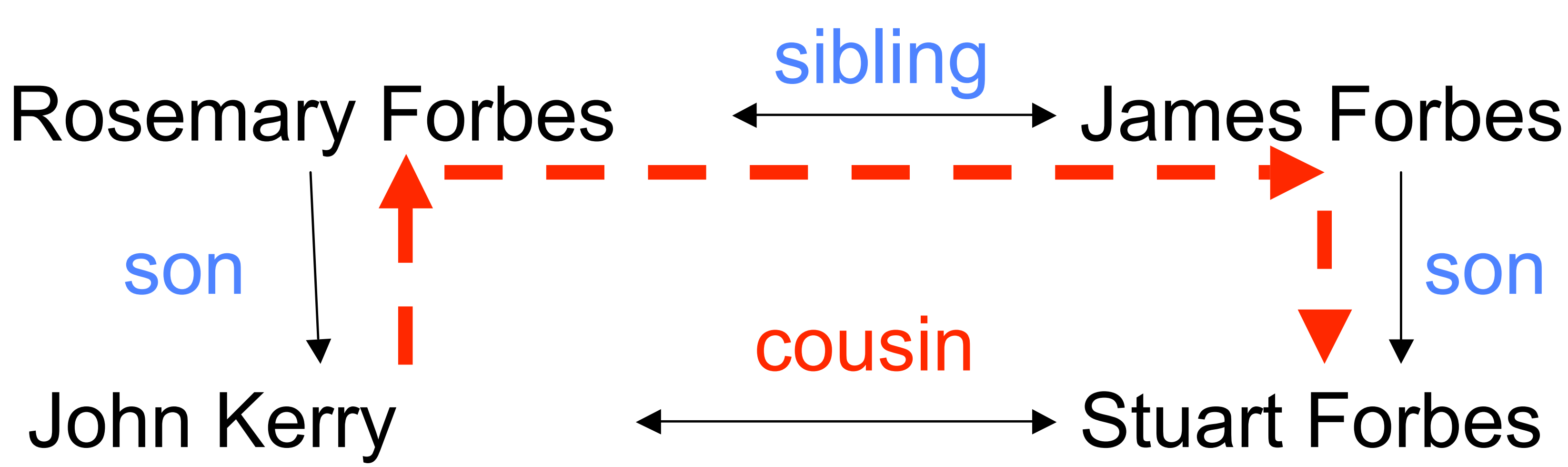


John Kerry likely a cousin

…celebrated with Stuart Forbes…

Name	Son
Rosemary Forbes	John Kerry
James Forbes	Stuart Forbes

Name	Sibling
Rosemary Forbes	James Forbes



Examples of Discovered Relational Features

- Mother: Father→Wife
- Cousin: Mother→Husband→Nephew
- Friend: Education→Student
- Education: Father→Education
- Boss: Boss→Son
- MemberOf: Grandfather→MemberOf
- Competition: PoliticalParty→Member→Competition

What is "Information Extraction"

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

* Microsoft Corporation
CEO
Bill Gates

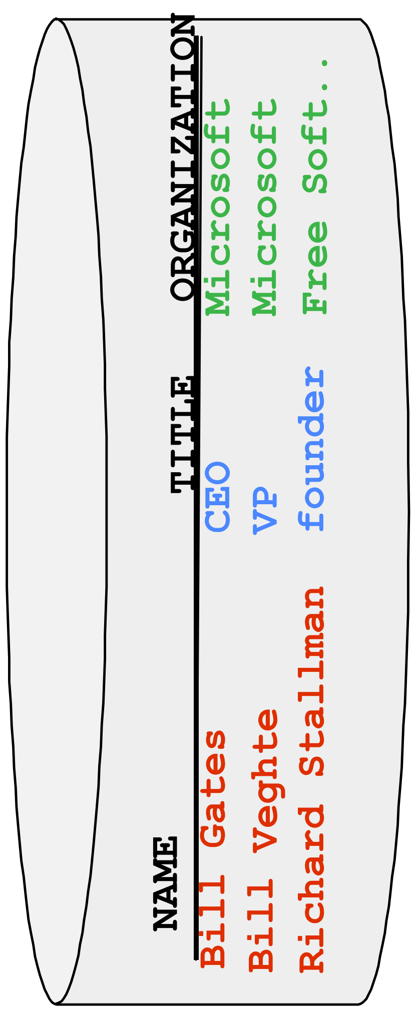
* Microsoft

* Gates

* Bill Veghte
Microsoft

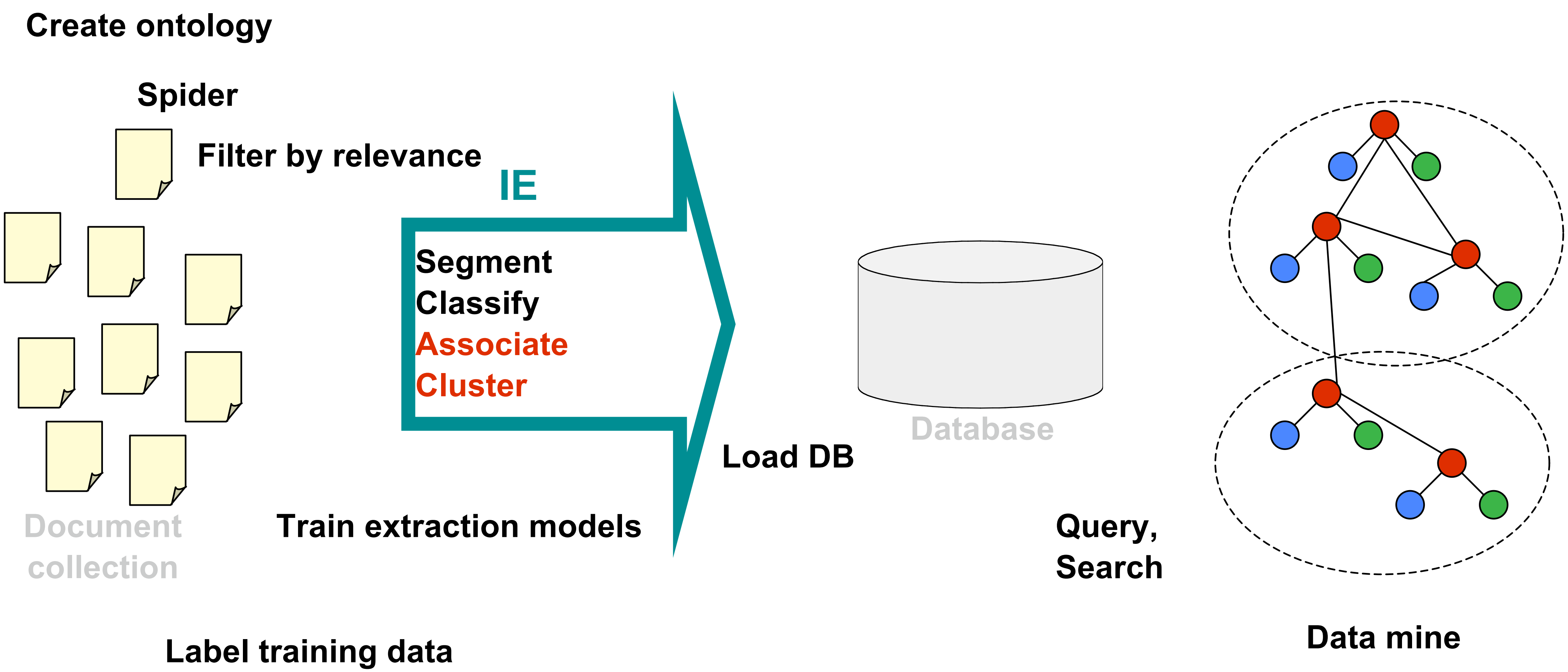
VP

Richard Stallman
founder
Free Software Foundation



3

IE in Context



Coreference Resolution

Coreference Resolution

AKA "record linkage", "database record deduplication", "citation matching", "object correspondence", "identity uncertainty"

Input

News article, with named-entity "mentions" tagged

Today Secretary of State Colin Powell met with . . . .
. . . . he . . . . . .
. . . . Condoleezza Rice . . . . .
. . . . Mr Powell . . . . .she . . . .
. . . . Powell . . . .
. . . President Bush . . . .
. . . . Rice . . . .
. . . . . . Bush . . . .
. . . .
. . . .

Output

Number of entities, *N = 3*

#1
- Secretary of State Colin Powell
- he
- Mr. Powell
- Powell

#2
- Condoleezza Rice
- she
- Rice

#3
- President Bush
- Bush

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

Queen Elizabeth set about transforming her husband,
King George VI, into a viable monarch. Logue,
a renowned speech therapist, was summoned to help
the King overcome his speech impediment...

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

Queen Elizabeth set about transforming her husband,
King George VI, into a viable monarch. Logue,
a renowned speech therapist, was summoned to help
the King overcome his speech impediment...

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

> Queen Elizabeth set about transforming her husband,
> King George VI, into a viable monarch. Logue,
> a renowned speech therapist, was summoned to help
> the King overcome his speech impediment...

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. Logue, a renowned speech therapist, was summoned to help the King overcome his speech impediment...

IE Example: Coreference

SAN SALVADOR, 15 JAN 90 (ACAN-EFE) -- [TEXT] ARMANDO CALDERON SOL, PRESIDENT OF THE NATIONALIST REPUBLICAN ALLIANCE (ARENA), THE RULING SALVADORAN PARTY, TODAY CALLED FOR AN INVESTIGATION INTO ANY POSSIBLE CONNECTION BETWEEN THE MILITARY PERSONNEL IMPLICATED IN THE ASSASSINATION OF JESUIT PRIESTS.

"IT IS SOMETHING SO HORRENDOUS, SO MONSTROUS, THAT WE MUST INVESTIGATE THE POSSIBILITY THAT THE FMLN (FARABUNDO MARTI NATIONAL LIBERATION FRONT) STAGED THESE MURDERS TO DISCREDIT THE GOVERNMENT," CALDERON SOL SAID.

SALVADORAN PRESIDENT ALFREDO CRISTIANI IMPLICATED FOUR OFFICERS, INCLUDING ONE COLONEL, AND FIVE MEMBERS OF THE ARMED FORCES IN THE ASSASSINATION OF SIX JESUIT PRIESTS AND TWO WOMEN ON 16 NOVEMBER AT THE CENTRAL AMERICAN UNIVERSITY.

Why It's Hard

Many sources of information play a role

- head noun matches
  - IBM *executives* = the *executives*
- syntactic constraints
  - John helped himself to...
  - John helped him to…
- number and gender agreement
- discourse focus, recency, syntactic parallelism, semantic class, world knowledge, …

Why It's Hard

- No single source is a completely reliable indicator
  - number agreement
    - the assassination = these murders
- Identifying each of these features automatically, accurately, and in context, is hard
- Coreference resolution subsumes the problem of pronoun resolution…

A Machine Learning Approach

- Classification
  - given a description of two noun phrases, NP_i and NP_j, classify the pair as *coreferent* or *not coreferent*

coref ? *coref ?*

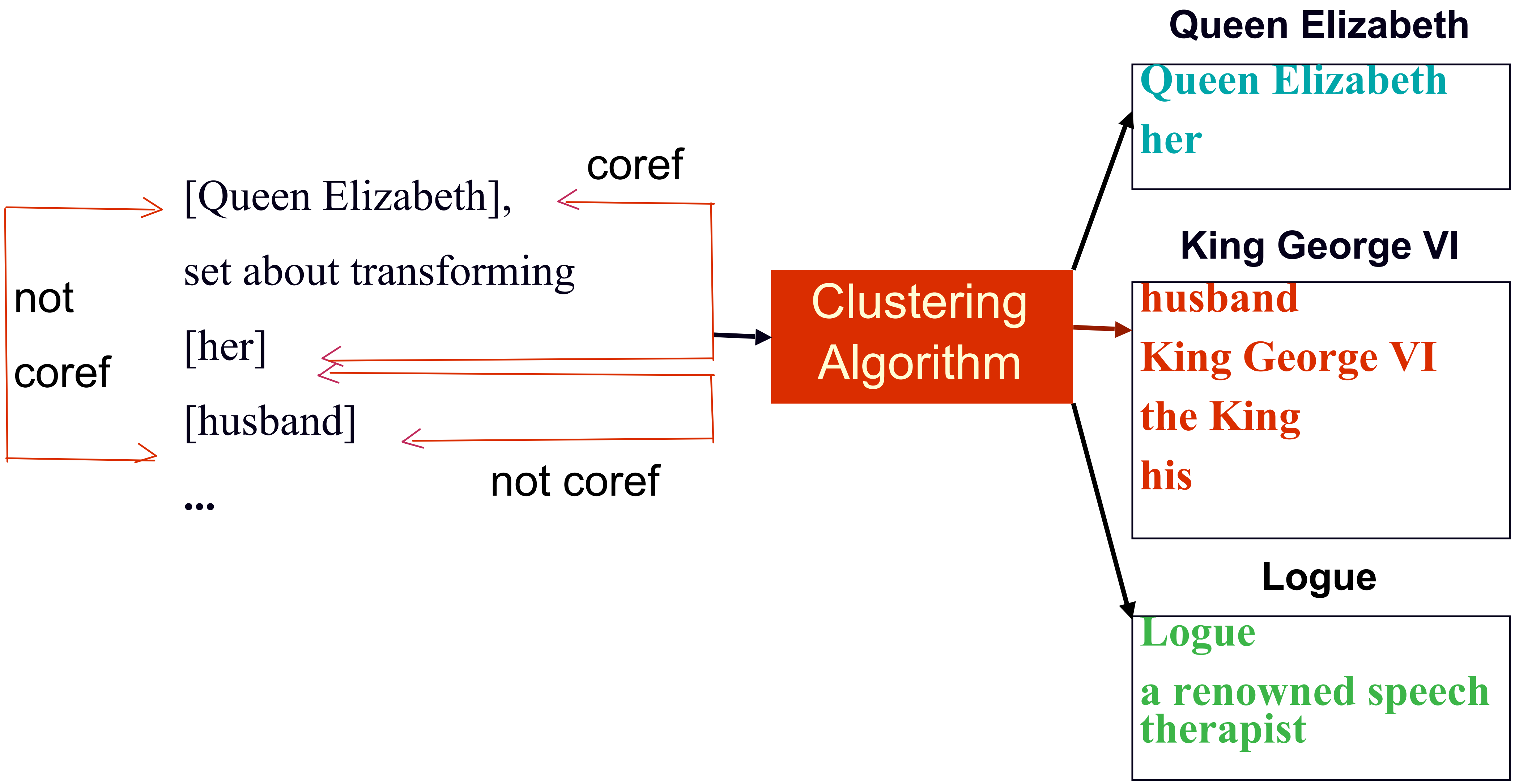
[Queen Elizabeth] set about transforming [her] [husband], ...

not coref ?

Aone & Bennett [1995]; Connolly et al. [1994]; McCarthy & Lehnert [1995]; Soon et al. [2001]; Ng & Cardie [2002]; …

A Machine Learning Approach

- Clustering
  - coordinates pairwise coreference decisions



Machine Learning Issues

- Training data creation
- Instance representation
- Learning algorithm
- Clustering algorithm

Training Data Creation

- Creating training instances
  - texts annotated with coreference information
  - one instance $inst(NP_i, NP_j)$ for each pair of NPs
    - assumption: NP_i precedes NP_j
    - feature vector: describes the two NPs and context
    - class value:

coref	pairs on the same coreference chain
not coref	otherwise

Instance Representation

- 25 features per instance
  - lexical (3)
    - string matching for pronouns, proper names, common nouns
  - grammatical (18)
    - pronoun, demonstrative (the, this), indefinite (it is raining), …
    - number, gender, animacy
    - appositive (george, the king), predicate nominative (a horse is a mammal)
    - binding constraints, simple contra-indexing constraints, …
    - span, maximalnp, …
  - semantic (2)
    - same WordNet class
    - alias
  - positional (1)
    - distance between the NPs in terms of # of sentences
  - knowledge-based (1)
    - naïve pronoun resolution algorithm

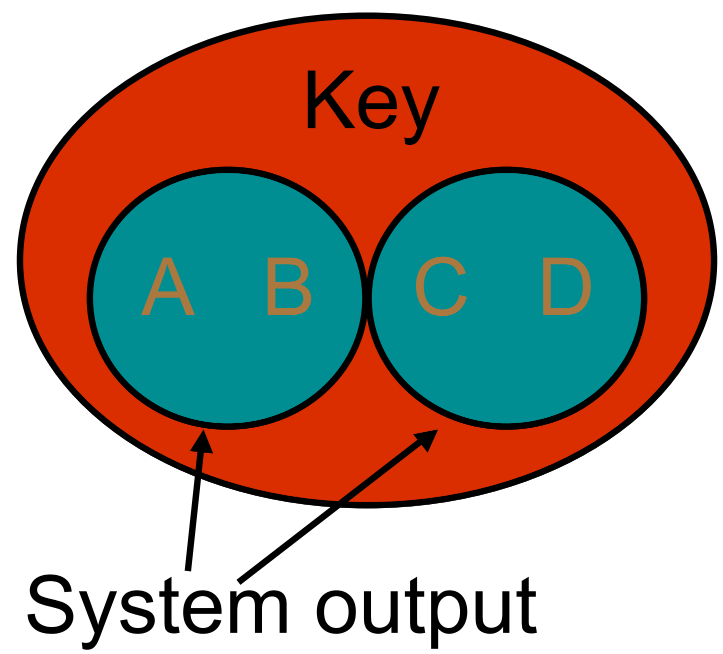
Clustering Algorithm

- Best-first single-link clustering
  - Mark each NP_j as belonging to its own class: $NP_j \in c_j$
  - Proceed through the NPs in left-to-right order.
    - For each NP, NP_j, create test instances, *inst*(NP_i, NP_j), for all of its preceding NPs, NP_i.
    - Select as the antecedent for NP_j the highest-confidence coreferent NP, NP_i, according to the coreference classifier (or none if all have below .5 confidence);

      Merge c_j and c_j .

Evaluation

- MUC-6 and MUC-7 coreference data sets
- documents annotated w.r.t. coreference
- 30 + 30 training texts (dry run)
- 30 + 20 test texts (formal evaluation)
- scoring program
  - recall
  - precision
  - F-measure: 2PR/(P+R)
- Types
  - MUC
  - ACE
  - Bcubed
  - Pairwise

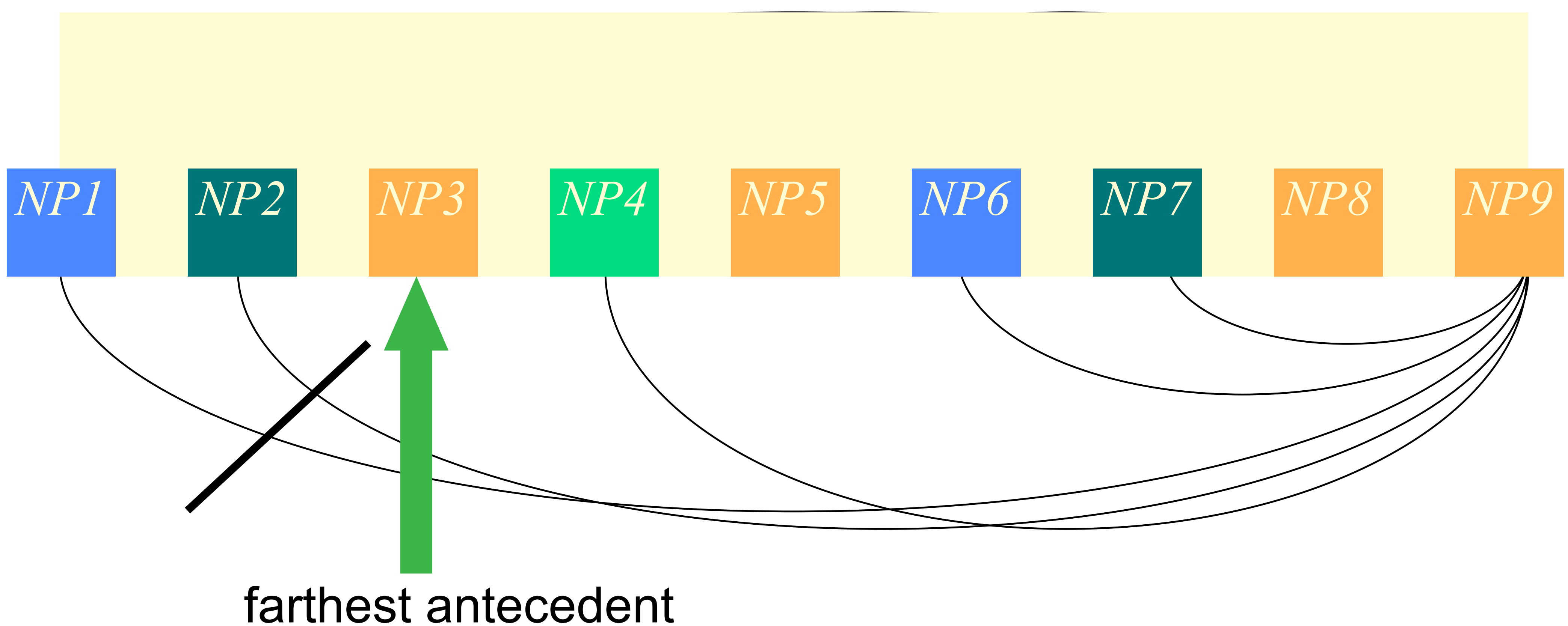


Baseline Results

	MUC-6			MUC-7		
	R	P	F	R	P	F
Baseline	40.7	73.5	**52.4**	27.2	86.3	**41.3**
Worst MUC System	36	44	40	52.5	21.4	30.4
Best MUC System	59	72	65	56.1	68.8	61.8

Problem 1

- Coreference is a rare relation
  - skewed class distributions (2% positive instances)
  - *remove some negative instances*



Problem 2

- Coreference is a discourse-level problem
  - different solutions for different types of NPs
    - proper names: string matching and aliasing
  - inclusion of “hard” positive training instances
  - *positive example selection*: selects easy positive training instances (cf. Harabagiu *et al.* (2001))

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. Logue, the renowned speech therapist, was summoned to help the King overcome his speech impediment...

Problem 3

- Coreference is an equivalence relation
  - loss of transitivity
  - need to tighten the connection between classification and clustering
  - *prune learned rules w.r.t. the clustering-level coreference scoring function*

coref ? *coref ?*

[Queen Elizabeth] set about transforming [her] [husband], ...

not coref ?

Results

	MUC-6			MUC-7		
	R	P	F	R	P	F
Baseline	40.7	73.5	52.4	27.2	86.3	41.3
NEG-SELECT	46.5	67.8	55.2	37.4	59.7	46.0
POS-SELECT	53.1	80.8	64.1	41.1	78.0	53.8
NEG-SELECT + POS-SELECT	63.4	76.3	69.3	59.5	55.1	57.2
NEG-SELECT + POS-SELECT + RULE-SELECT	63.3	76.9	**69.5**	54.2	76.3	**63.4**

- Ultimately: large increase in F-measure, due to gains in recall

Comparison with Best MUC Systems

	MUC-6			MUC-7		
	R	P	F	R	P	F
NEG-SELECT + POS-SELECT + RULE-SELECT	63.3	76.9	**69.5**	54.2	76.3	**63.4**
Best MUC System	59	72	**65**	56.1	68.8	**61.8**

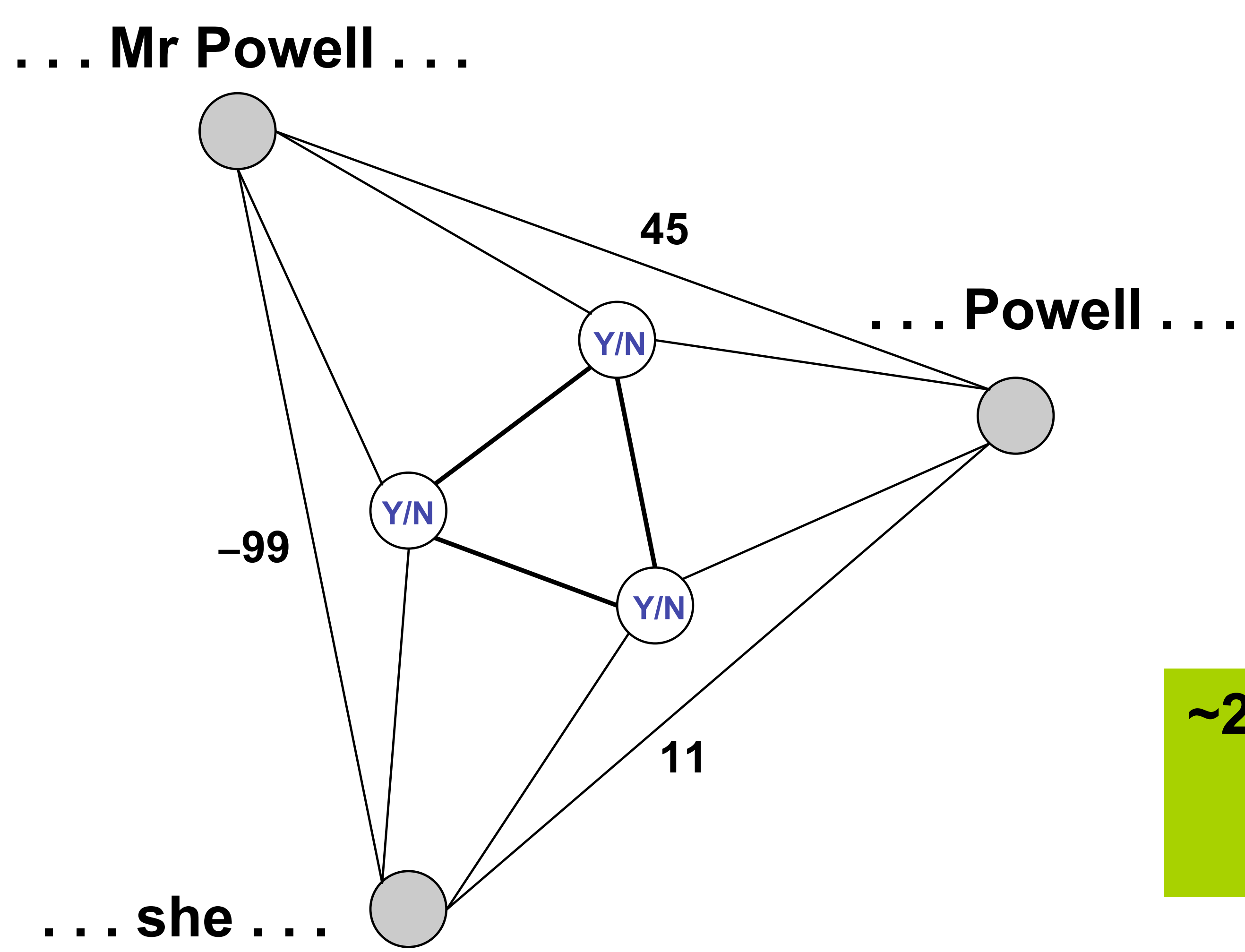
Main Points

Co-reference

- How to cast as classification [Cardie]
- **Joint resolution [McCallum et al]**

Joint co-reference among all pairs
<u>Affinity Matrix CRF</u>

"Entity resolution"
"Object correspondence"



~25% reduction in error on co-reference of proper nouns in newswire.

Inference:
Correlational clustering
graph partitioning

[Bansal, Blum, Chawla, 2002]

[McCallum, Wellner, IJCAI WS 2003, NIPS 2004]

Coreference Resolution

AKA "record linkage", "database record deduplication", "citation matching", "object correspondence", "identity uncertainty"

Input

News article,
with named-entity "mentions" tagged

Today Secretary of State Colin Powell
met with . . . .
. . . . . he . . . . . .
. . . . Condoleezza Rice . . . . .
. . . . Mr Powell . . . . .she . . . .
. . . . Powell . . . .
. . . President Bush . . . .
. Rice . . . .
. . . . Bush . . . .
. . . . .
. . . . .

Output

Number of entities, *N = 3*

#1
Secretary of State Colin Powell
he
Mr. Powell
Powell

#2
Condoleezza Rice
she
Rice

#3
President Bush
Bush

Inside the Traditional Solution

Pair-wise Affinity Metric

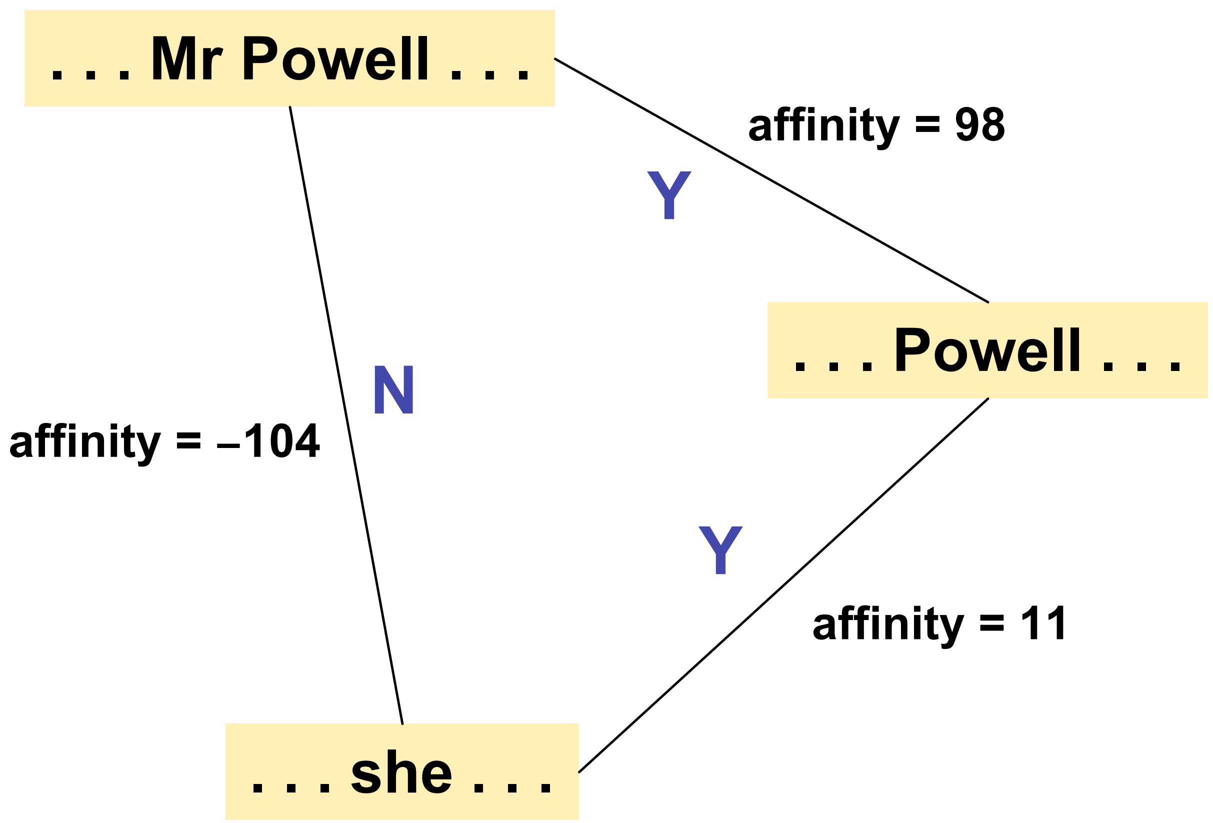
Mention (3) | **Mention (4)**

. . . Mr Powell . . . —— **Y/N?** —— **. . . Powell . . .**

N	Two words in common	29
Y	**One word in common**	**13**
Y	**"Normalized" mentions are string identical**	**39**
Y	**Capitalized word in common**	**17**
Y	**> 50% character tri-gram overlap**	**19**
N	< 25% character tri-gram overlap	-34
Y	**In same sentence**	**9**
Y	**Within two sentences**	**8**
N	Further than 3 sentences apart	-1
Y	**"Hobbs Distance" < 3**	**11**
N	Number of entities in between two mentions = 0	12
N	Number of entities in between two mentions > 4	-3
Y	**Font matches**	**1**
Y	**Default**	**-19**

OVERALL SCORE = 98 > threshold=0

The Problem



Pair-wise merging decisions are being made independently from each other

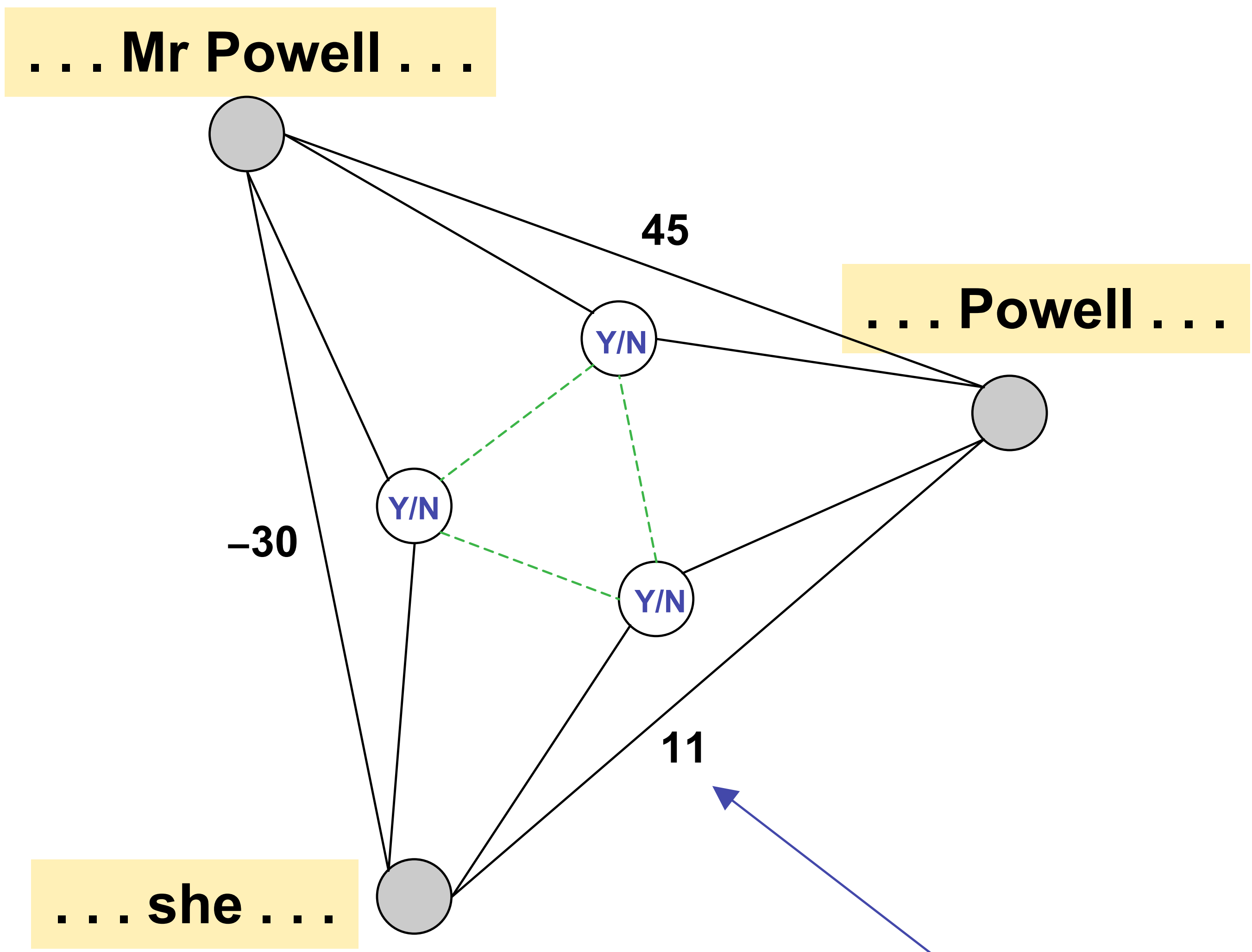
They should be made in relational dependence with each other.

Affinity measures are noisy and imperfect.

A Markov Random Field for Co-reference

(MRF)

[McCallum & Wellner, 2003, ICML]



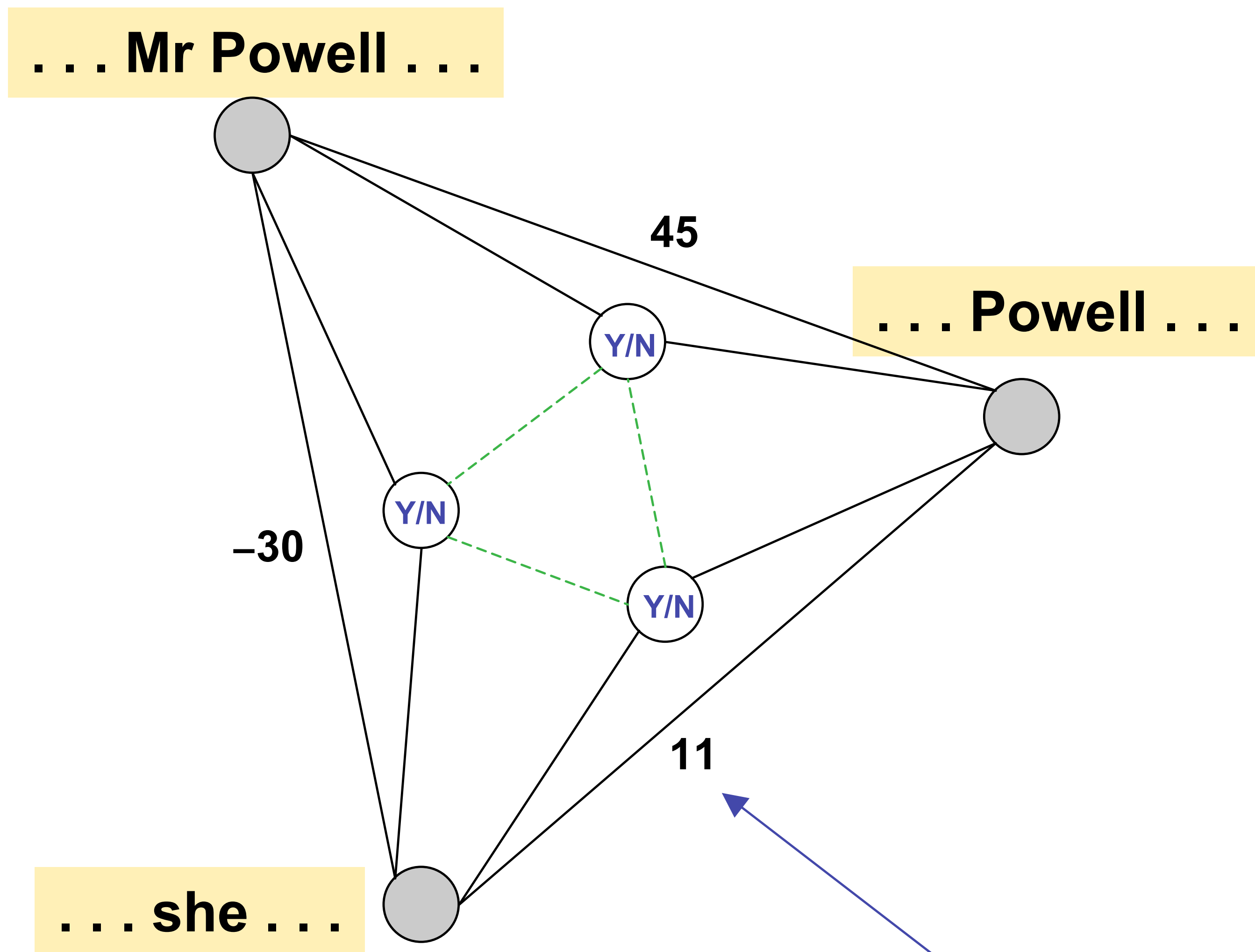
Make pair-wise merging decisions in dependent relation to each other by
- calculating a joint prob.
- including all edge weights
- adding dependence on consistent triangles.

$$P(\vec{y} \mid \vec{x}) = \frac{1}{Z_{\vec{x}}} \exp\left(\sum_{i,j} \sum_{l} \lambda_l f_l(x_i, x_j, y_{ij}) + \sum_{i,j,k} \lambda' f'(y_{ij}, y_{jk}, y_{ik}) \right)$$

A Markov Random Field for Co-reference

(MRF)

[McCallum & Wellner, 2003]



Make pair-wise merging decisions in dependent relation to each other by
- **calculating a joint prob.**
- **including all edge weights**
- **adding dependence on consistent triangles.**

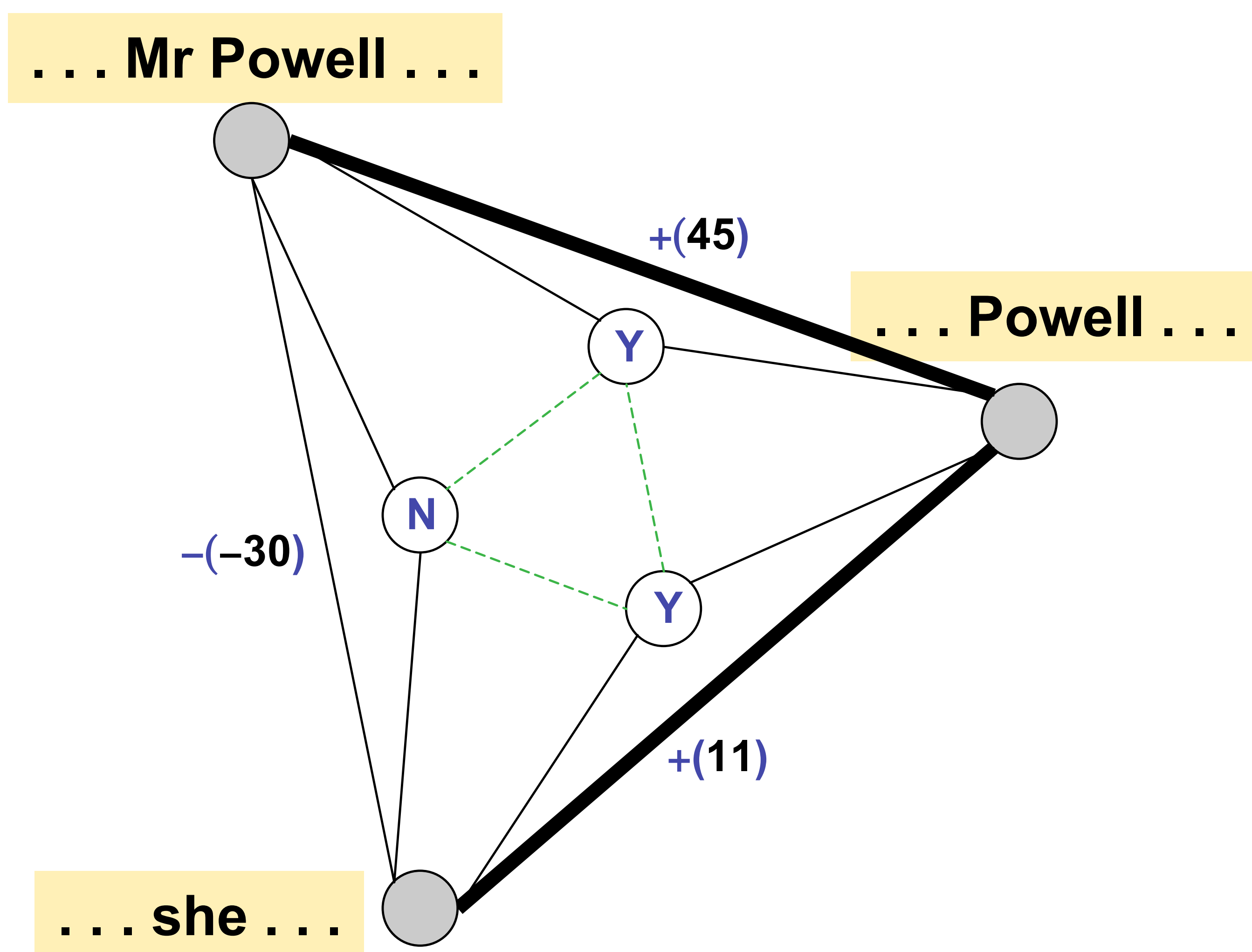
$-\infty$

$$P(\vec{y}\,|\,\vec{x}) = \frac{1}{Z_{\vec{x}}} \exp\left(\sum_{i,j} \sum_{l} \lambda_l f_l(x_i, x_j, y_{ij}) + \sum_{i,j,k} \lambda' f'(y_{ij}, y_{jk}, y_{ik}) \right)$$

A Markov Random Field for Co-reference

(MRF)

[McCallum & Wellner, 2003]



$$P(\vec{y} \mid \vec{x}) = \frac{1}{Z_{\vec{x}}} \exp\left(\sum_{i,j} \sum_{l} \lambda_l f_l(x_i, x_j, y_{ij}) + \underbrace{\sum_{i,j,k} \lambda' f'(y_{ij}, y_{jk}, y_{ik})}_{}\right)$$

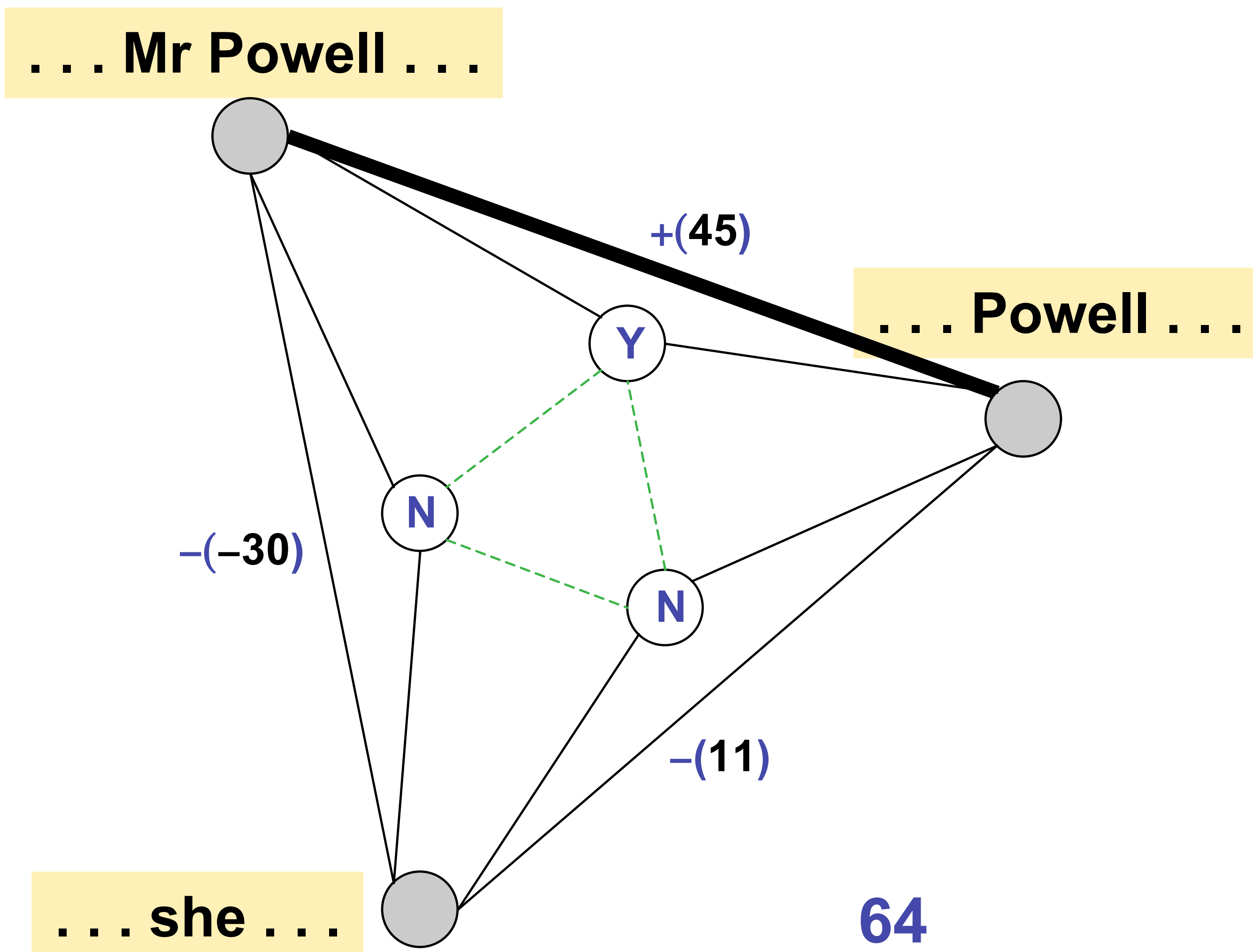
−infinity

40

A Markov Random Field for Co-reference

(MRF)

[McCallum & Wellner, 2003]

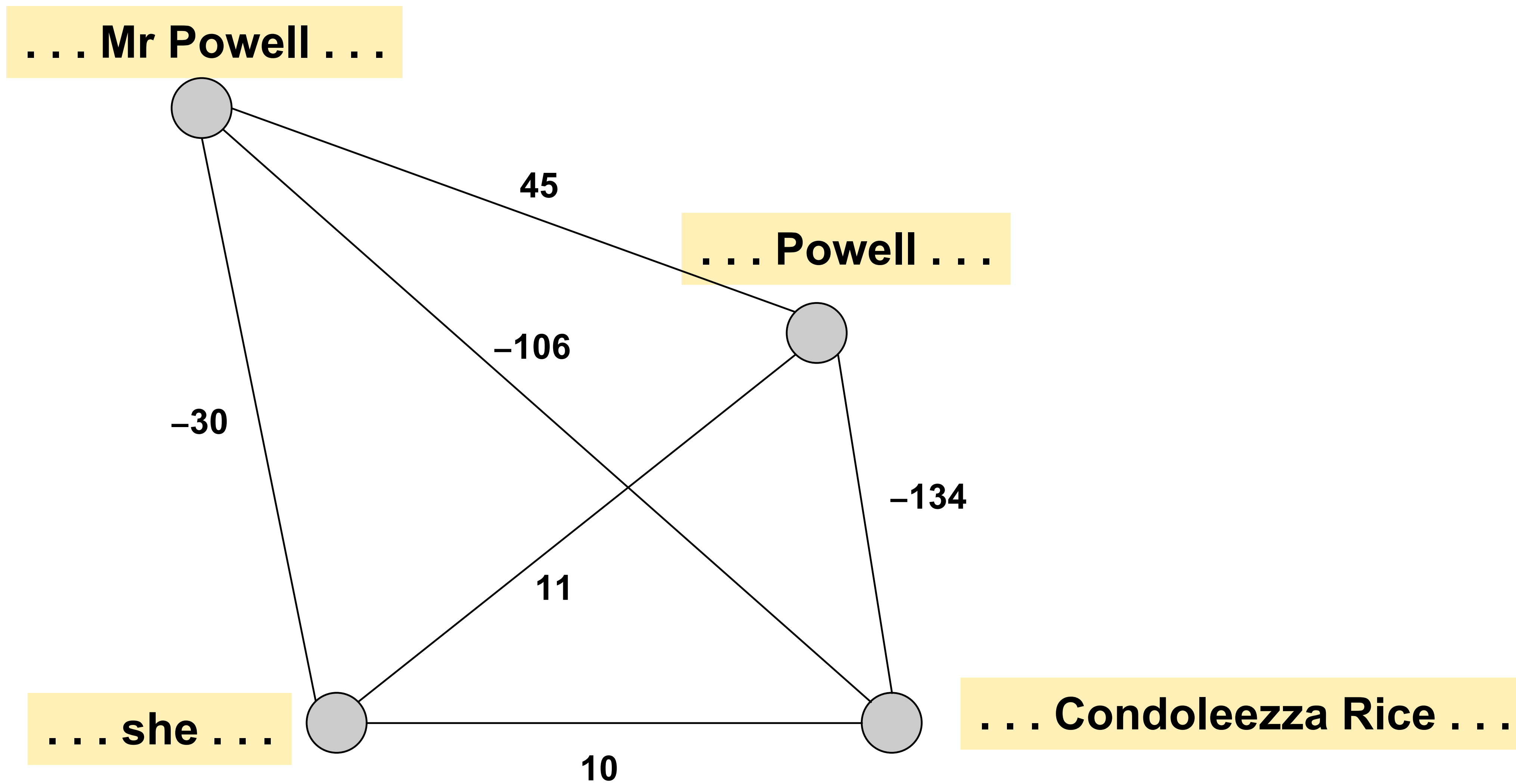


$$P(\vec{y}\,|\,\vec{x}) = \frac{1}{Z_{\vec{x}}}\exp\left(\overbrace{\sum_{i,j}\sum_{l}\lambda_l f_l(x_i,x_j,y_{ij})}^{64} + \sum_{i,j,k}\lambda' f'(y_{ij},y_{jk},y_{ik})\right)$$

41

Inference in these MRFs = Graph Partitioning

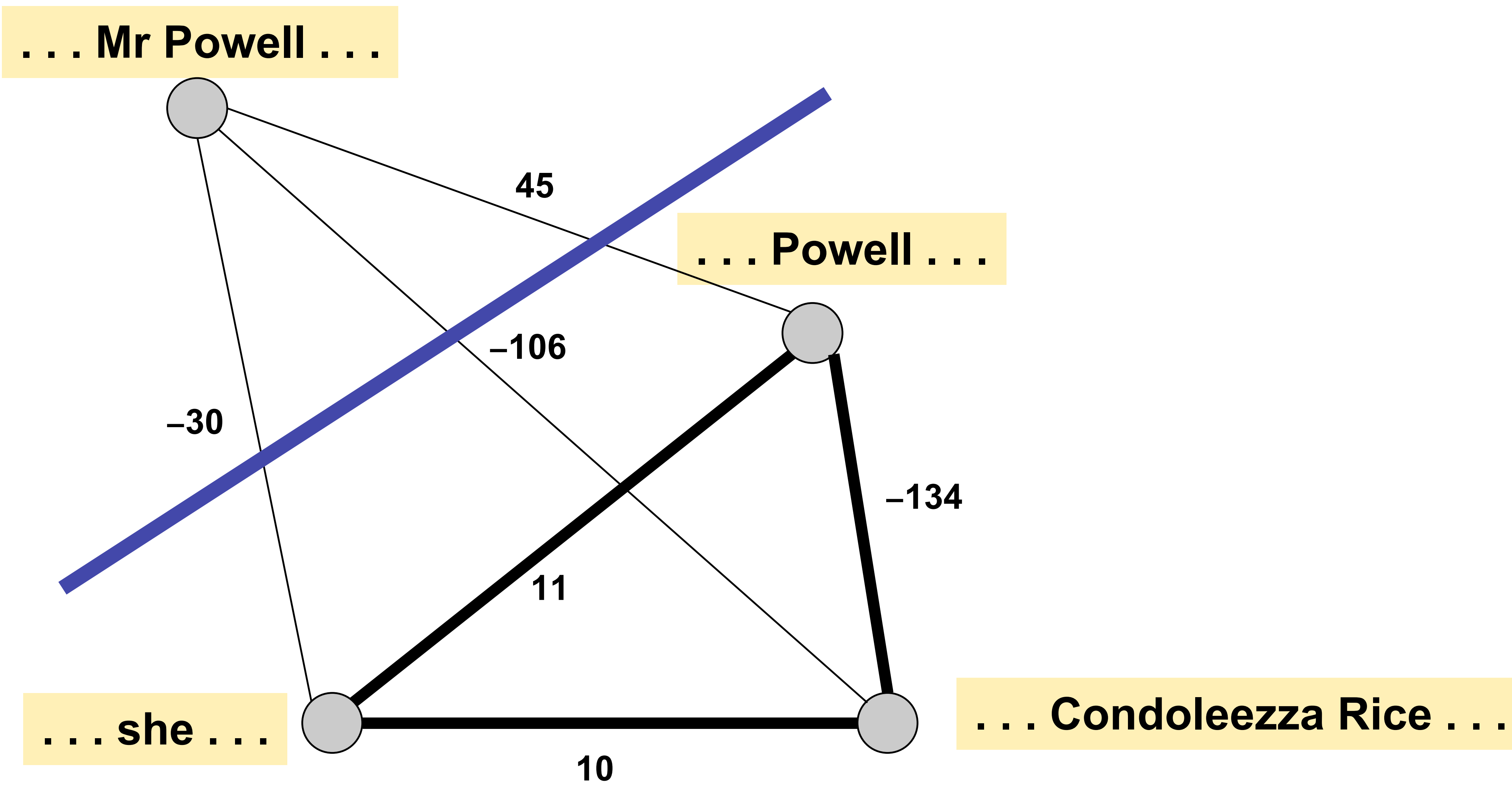
[Boykov, Vekler, Zabih, 1999], [Kolmogorov & Zabih, 2002], [Yu, Cross, Shi, 2002]



$$\log\left(P(\vec{y} \mid \vec{x})\right) \propto \sum_{i,j} \sum_{l} \lambda_l f_l(x_i, x_j, y_{ij}) = \sum_{\substack{i,j \text{ w/in} \\ \text{paritions}}} \mathrm{w}_{\mathrm{ij}} - \sum_{\substack{i,j \text{ across} \\ \text{paritions}}} \mathrm{w}_{\mathrm{ij}}$$

Inference in these MRFs = Graph Partitioning

[Boykov, Vekler, Zabih, 1999], [Kolmogorov & Zabih, 2002], [Yu, Cross, Shi, 2002]

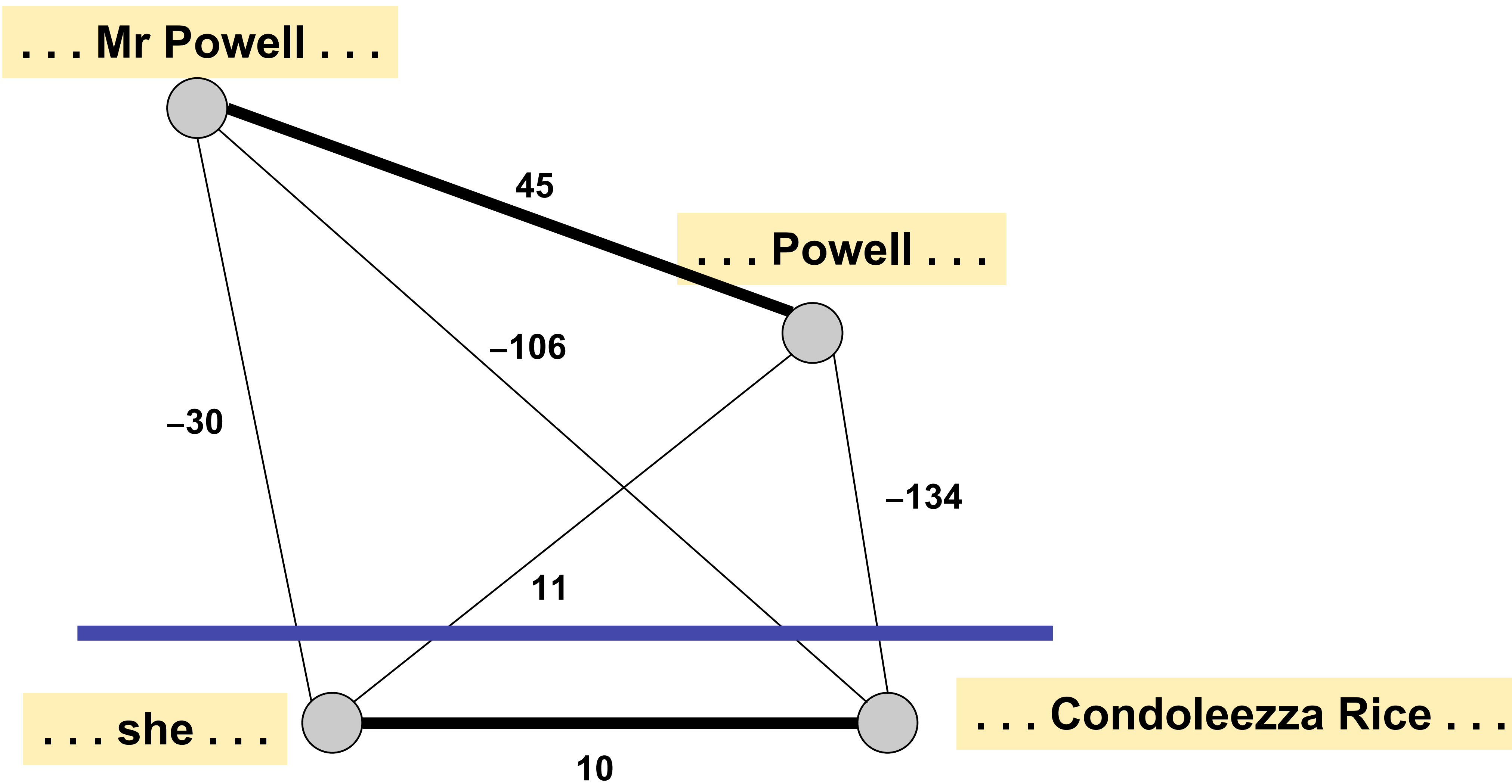


$$\log\left(P(\vec{y} \mid \vec{x})\right) \propto \sum_{i,j} \sum_{l} \lambda_l f_l(x_i, x_j, y_{ij}) = \sum_{\substack{i,j \text{ w/in} \\ \text{paritions}}} \mathrm{w}_{\mathrm{ij}} - \sum_{\substack{i,j \text{ across} \\ \text{paritions}}} \mathrm{w}_{\mathrm{ij}}$$ **= –22**

43

Inference in these MRFs = Graph Partitioning

[Boykov, Vekler, Zabih, 1999], [Kolmogorov & Zabih, 2002], [Yu, Cross, Shi, 2002]



$$\log\left(P(\vec{y} \mid \vec{x})\right) \propto \sum_{i,j} \sum_{l} \lambda_l f_l(x_i, x_j, y_{ij}) = \sum_{\substack{i,j \text{ w/in} \\ \text{paritions}}} \mathrm{w}_{\mathrm{ij}} + \sum_{\substack{i,j \text{ across} \\ \text{paritions}}} \mathrm{w}'_{\mathrm{ij}}$$ **= 314**

Co-reference Experimental Results

[McCallum & Wellner, 2003]

Proper noun co-reference

DARPA ACE broadcast news transcripts, *117 stories*

	Partition F1	Pair F1
Single-link threshold	16 %	18 %
Best prev match [Morton]	83 %	89 %
MRFs	88 %	92 %
	Δerror=30%	Δerror=28%

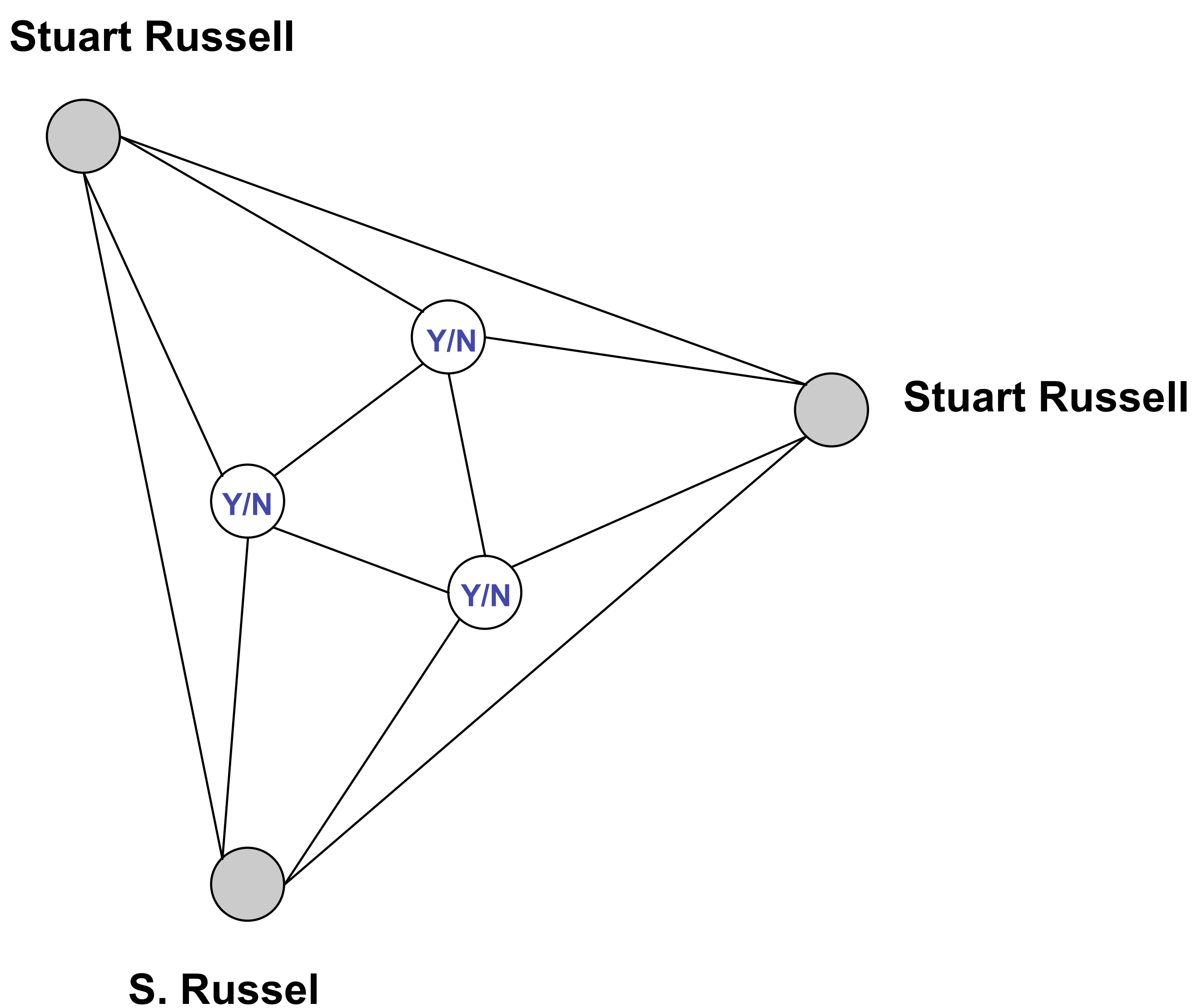
DARPA MUC-6 newswire article corpus, *30 stories*

	Partition F1	Pair F1
Single-link threshold	11%	7 %
Best prev match [Morton]	70 %	76 %
MRFs	74 %	80 %
	Δerror=13%	Δerror=17%

Joint Co-reference for Multiple Entity Types

[Culotta & McCallum 2005]

<u>People</u>

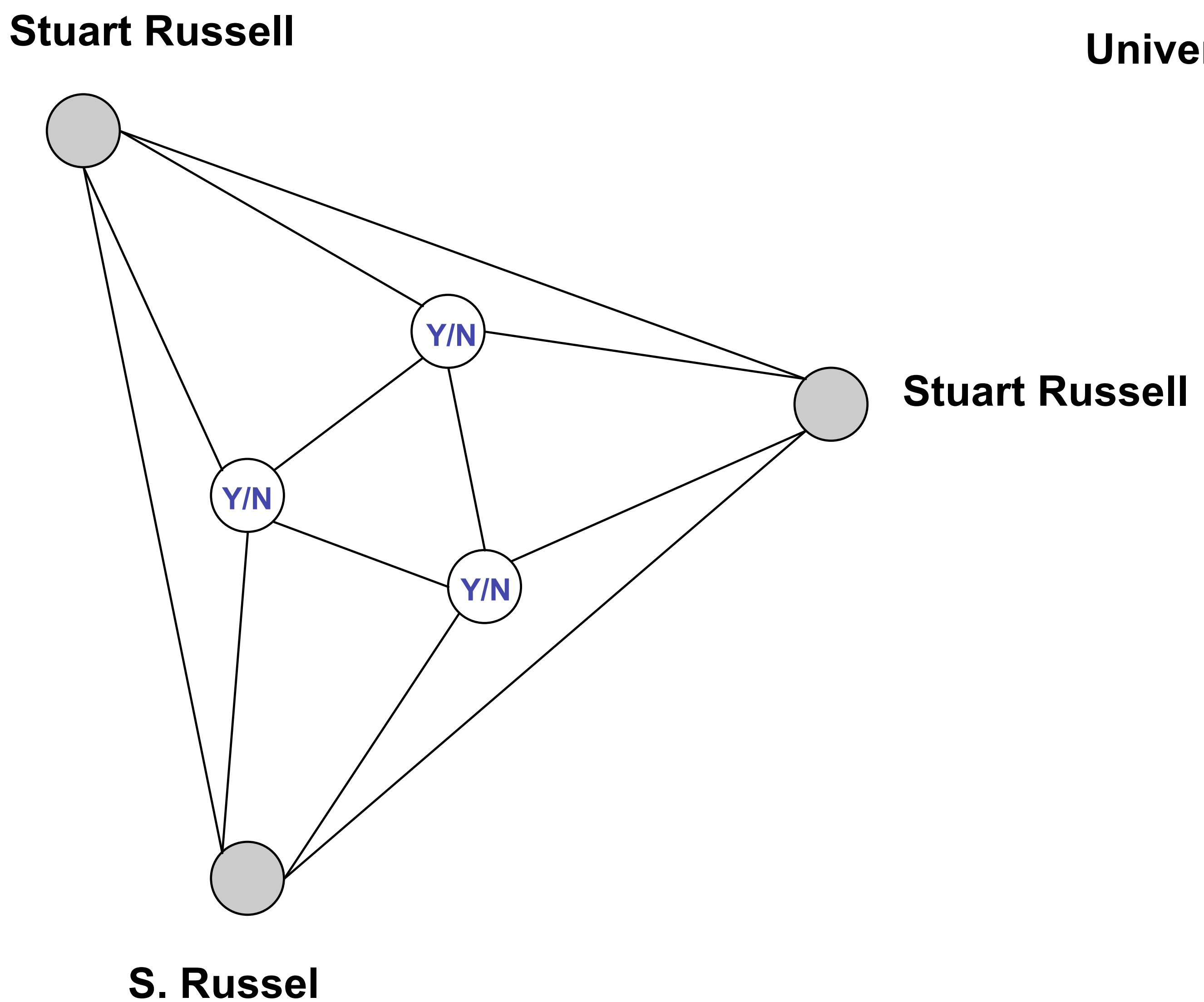


46

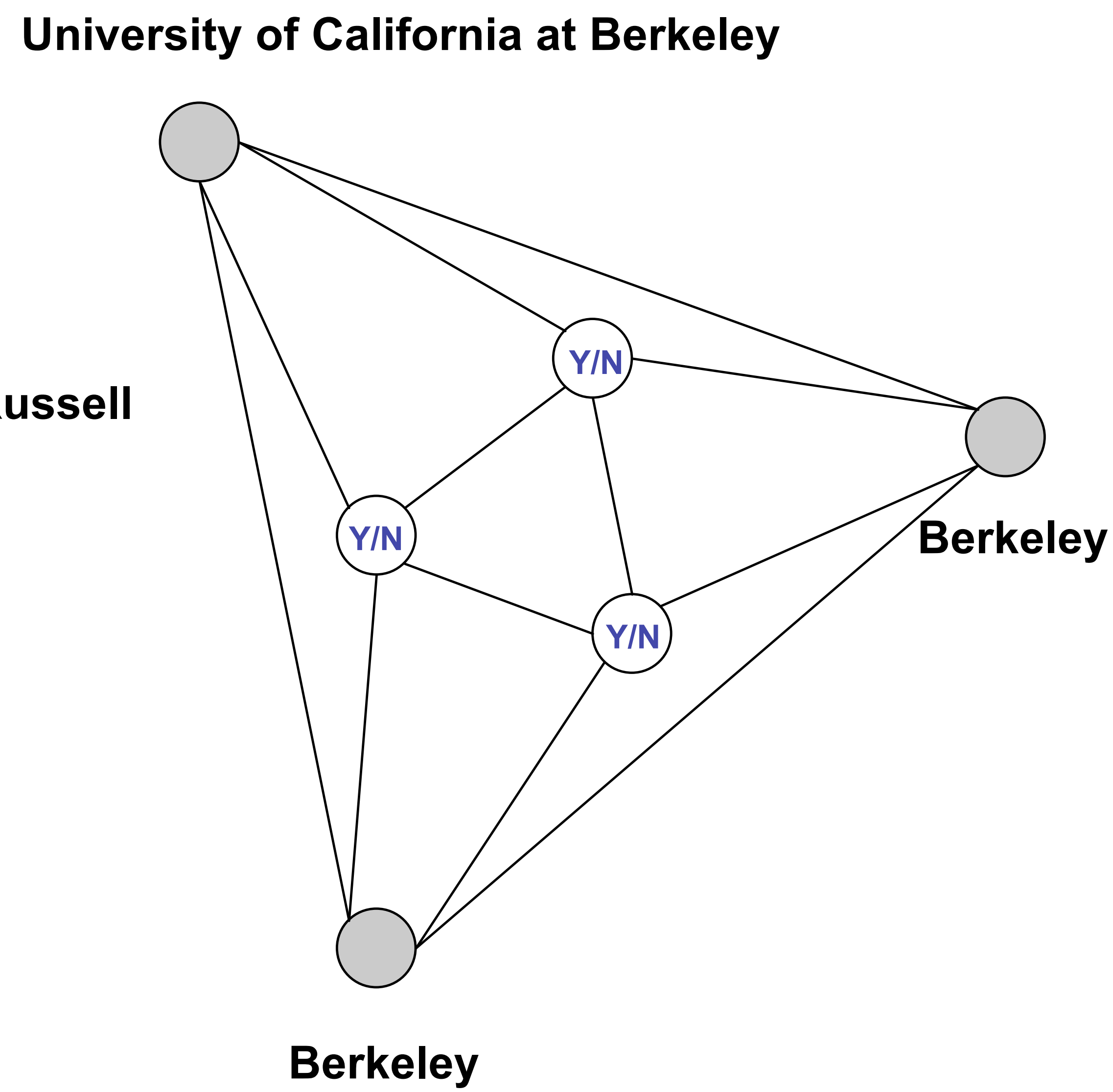
Joint Co-reference for Multiple Entity Types

[Culotta & McCallum 2005]

People

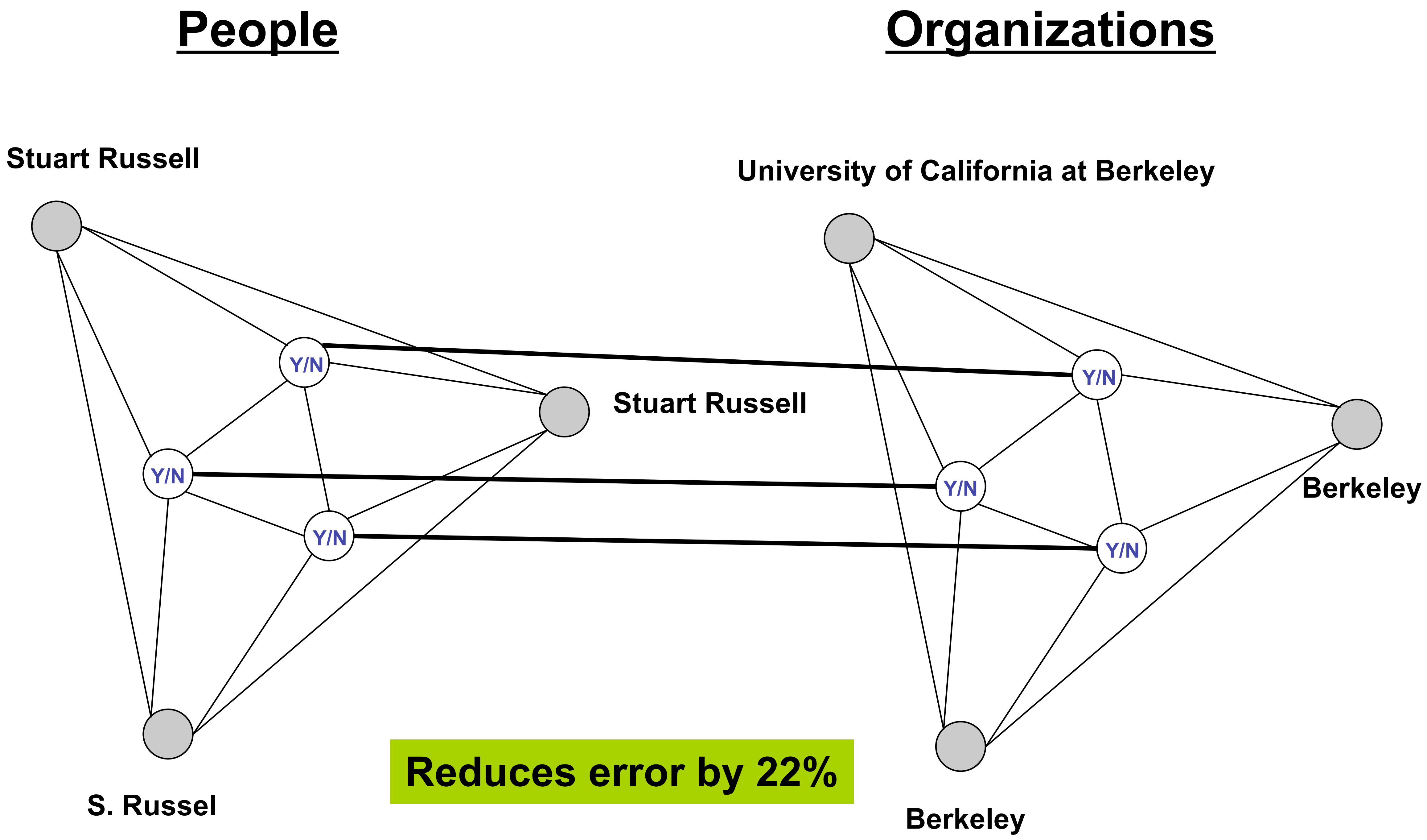


Organizations



Joint Co-reference for Multiple Entity Types

[Culotta & McCallum 2005]



Reduces error by 22%

Question Answering

Question Answering from Text

- The common person's view? [From a novel]
  - "I like the Internet. Really, I do. Any time I need a piece of shareware or I want to find out the weather in Bogota … I'm the first guy to get the modem humming. But as a source of information, it sucks. You got a billion pieces of data, struggling to be heard and seen and downloaded, and anything I want to know seems to get trampled underfoot in the crowd."
    - M. Marshall. *The Straw Men.* HarperCollins Publishers, 2002.

- Question Answering:
  - Give the user a (short) answer to their question, perhaps supported by evidence.
  - An idea originating from the IR community
  - With massive collections of full-text documents, simply finding *relevant documents* is of limited use: we want *answers* from textbases

People *want* to ask questions?

Examples of search queries

who invented surf music?
how to make stink bombs
where are the snowdens of yesteryear?
which english translation of the bible is used in official catholic liturgies?
how to do clayart
how to copy psx
how tall is the sears tower?
how can i find someone in texas
where can i find information on puritan religion?
what are the 7 wonders of the world
how can i eliminate stress
What vacuum cleaner does Consumers Guide recommend

Around 10–15% of query logs

AskJeeves (Classic)

- Probably the most hyped example of "question answering"
- It largely did pattern matching to match your question to their own knowledge base of questions
- If that works, you get the human-curated answers to that known question (which are presumably good)
- If that fails, it falls back to regular web search
- A potentially interesting middle ground, but not full QA

A Brief (Academic) History

- Question answering is not a new research area
- Question answering systems can be found in many areas of NLP research, including:
    - Natural language database systems
        - A lot of early NLP work on these
    - Spoken dialog systems
        - Currently very active and commercially relevant
- The focus on open-domain QA is new
    - MURAX (Kupiec 1993): Encyclopedia answers
    - Hirschman: Reading comprehension tests
    - TREC QA competition: 1999–

Question Answering at TREC

- Question answering competition at TREC consists of answering a set of 500 fact-based questions, e.g., *“When was Mozart born*?”.
- For the first three years systems were allowed to return 5 ranked answer snippets (50/250 bytes) to each question.
  - IR think
  - Mean Reciprocal Rank (MRR) scoring:
    - 1, 0.5, 0.33, 0.25, 0.2, 0 for 1, 2, 3, 4, 5, 6+ doc
  - Mainly Named Entity answers (person, place, date, …)
- From 2002 the systems are only allowed to return a single *exact* answer and the notion of confidence has been introduced.

The TREC Document Collection

- One recent round: news articles from:
  - AP newswire, 1998-2000
  - New York Times newswire, 1998-2000
  - Xinhua News Agency newswire, 1996-2000
- In total 1,033,461 documents in the collection.
- 3GB of text
- While small in some sense, still too much text to process using advanced NLP techniques (on the fly at least)
- Systems usually have initial information retrieval followed by advanced processing.
- Many supplement this text with use of the web, and other knowledge bases

Sample TREC questions

1. Who is the author of the book, "The Iron Lady: A Biography of Margaret Thatcher"?
2. What was the monetary value of the Nobel Peace Prize in 1989?
3. What does the Peugeot company manufacture?
4. How much did Mercury spend on advertising in 1993?
5. What is the name of the managing director of Apricot Computer?
6. Why did David Koresh ask the FBI for a word processor?
7. What debts did Qintex group leave?
8. What is the name of the rare neurological disease with symptoms such as: involuntary movements (tics), swearing, and incoherent vocalizations (grunts, shouts, etc.)?

Top Performing Systems

- Currently the best performing systems at TREC can answer approximately 70% of the questions
- Approaches and successes have varied a fair deal
  - Knowledge-rich approaches, using a vast array of NLP techniques stole the show in 2000, 2001, still do well
    - Notably Harabagiu, Moldovan et al. – SMU/UTD/LCC
  - AskMSR system stressed how much could be achieved by very simple methods with enough text (and now various copycats)
  - Middle ground is to use large collection of surface matching patterns (ISI)

Ravichandran and Hovy 2002
Learning Surface Patterns

- Use of Characteristic Phrases
- "When was <person> born"
  - Typical answers
    - "Mozart was born in 1756."
    - "Gandhi (1869-1948)..."
  - Suggests phrases like
    - "<NAME> was born in <BIRTHDATE>"
    - "<NAME> (<BIRTHDATE>-"
  - as Regular Expressions can help locate correct answer

Use Pattern Learning

- Example: Start with “Mozart 1756”
  - Results:
    - “The great composer Mozart (1756-1791) achieved fame at a young age”
    - “Mozart (1756-1791) was a genius”
    - “The whole world would always be indebted to the great music of Mozart (1756-1791)”
  - Longest matching substring for all 3 sentences is "Mozart (1756-1791)”
  - Suffix tree would extract "Mozart (1756-1791)" as an output, with score of 3
- Reminiscent of IE pattern learning

Pattern Learning (cont.)

- Repeat with different examples of same question type
  - “Gandhi 1869”, “Newton 1642”, etc.
- Some patterns learned for BIRTHDATE
  - a. born in <ANSWER>, <NAME>
  - b. <NAME> was born on <ANSWER> ,
  - c. <NAME> (<ANSWER> -
  - d. <NAME> (<ANSWER> -)

Experiments: (R+H, 2002)

- 6 different Question types
  - from Webclopedia QA Typology (Hovy et al., 2002a)
    - BIRTHDATE
    - LOCATION
    - INVENTOR
    - DISCOVERER
    - DEFINITION
    - WHY-FAMOUS

Experiments: pattern precision

- BIRTHDATE table:
  - 1.0 <NAME> (<ANSWER> -)
  - 0.85 <NAME> was born on <ANSWER>,
  - 0.6 <NAME> was born in <ANSWER>
  - 0.59 <NAME> was born <ANSWER>
  - 0.53 <ANSWER> <NAME> was born
  - 0.50 - <NAME> (<ANSWER>
  - 0.36 <NAME> (<ANSWER> -
- INVENTOR
  - 1.0 <ANSWER> invents <NAME>
  - 1.0 the <NAME> was invented by <ANSWER>
  - 1.0 <ANSWER> invented the <NAME> in

Experiments (cont.)

- WHY-FAMOUS
  - 1.0 <ANSWER> <NAME> called
  - 1.0 laureate <ANSWER> <NAME>
  - 0.71 <NAME> is the <ANSWER> of
- LOCATION
  - 1.0 <ANSWER>'s <NAME>
  - 1.0 regional : <ANSWER> : <NAME>
  - 0.92 near <NAME> in <ANSWER>
- Depending on question type, get high MRR (0.6–0.9), with higher results from use of Web than TREC QA collection

Shortcomings & Extensions

- Need for POS &/or semantic types
    - "Where are the Rocky Mountains?”
    - "Denver's new airport, topped with white fiberglass cones in imitation of the Rocky Mountains in the background , continues to lie empty”
    - <NAME> in <ANSWER>
- NE tagger &/or ontology could enable system to determine "background" is not a location

Shortcomings... (cont.)

- Long distance dependencies
  - "Where is London?"
  - "London, which has one of the busiest airports in the world, lies on the banks of the river Thames"
  - would require pattern like:
    <QUESTION>, (<any_word>)*, lies on <ANSWER>
- But: abundance & variety of Web data helps system to find an instance of patterns w/o losing answers to long distance dependencies

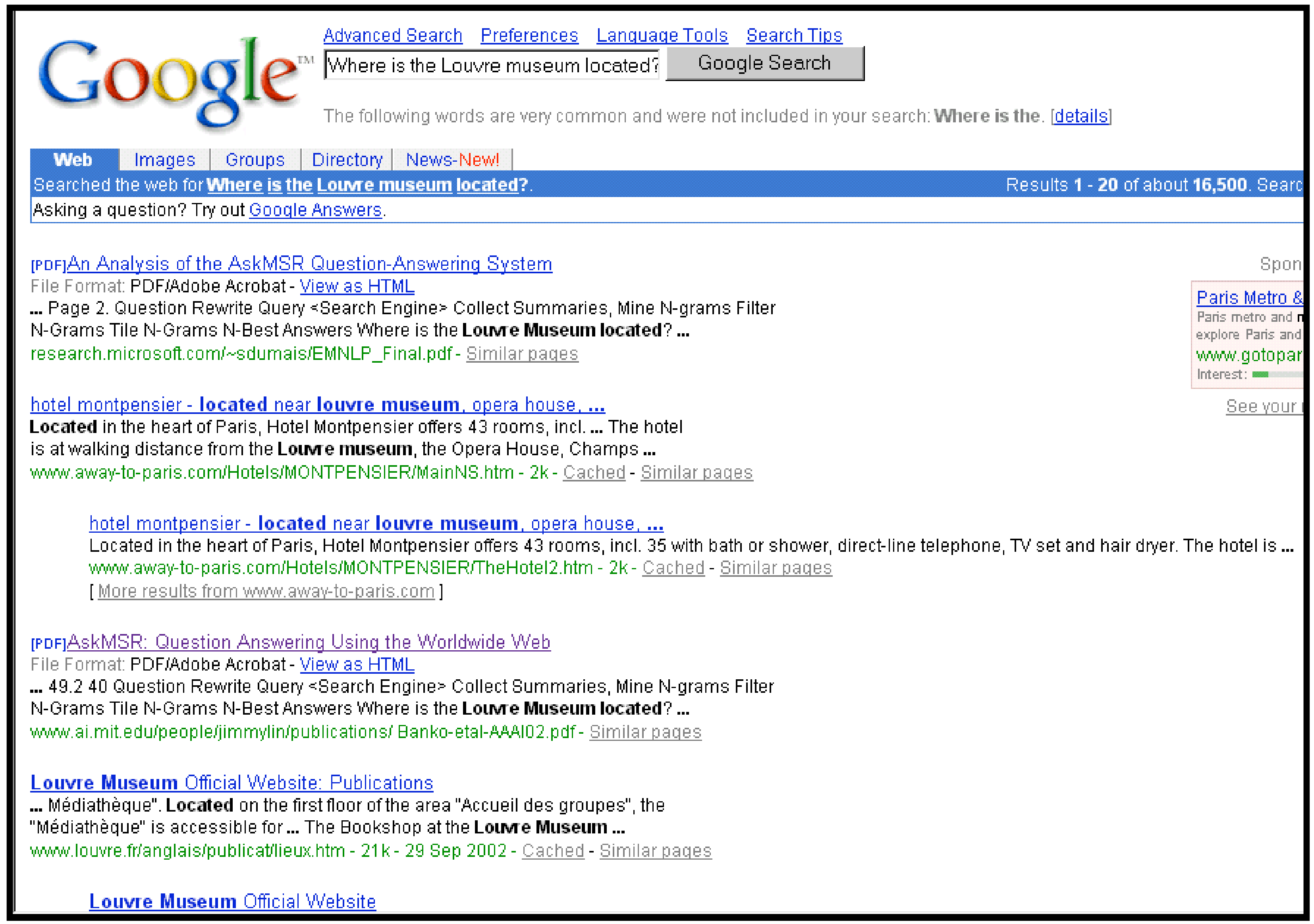
Shortcomings... (cont.)

- Their system uses only one anchor word
  - Doesn't work for Q types requiring multiple words from question to be in answer
    - "In which county does the city of Long Beach lie?”
    - "Long Beach is situated in Los Angeles County”
    - required pattern:
      <Q_TERM_1> is situated in <ANSWER> <Q_TERM_2>
- Does not use case
    - "What is a micron?”
    - "...a spokesman for Micron, <u>a maker of semiconductors</u>, said SIMMs are..."

AskMSR

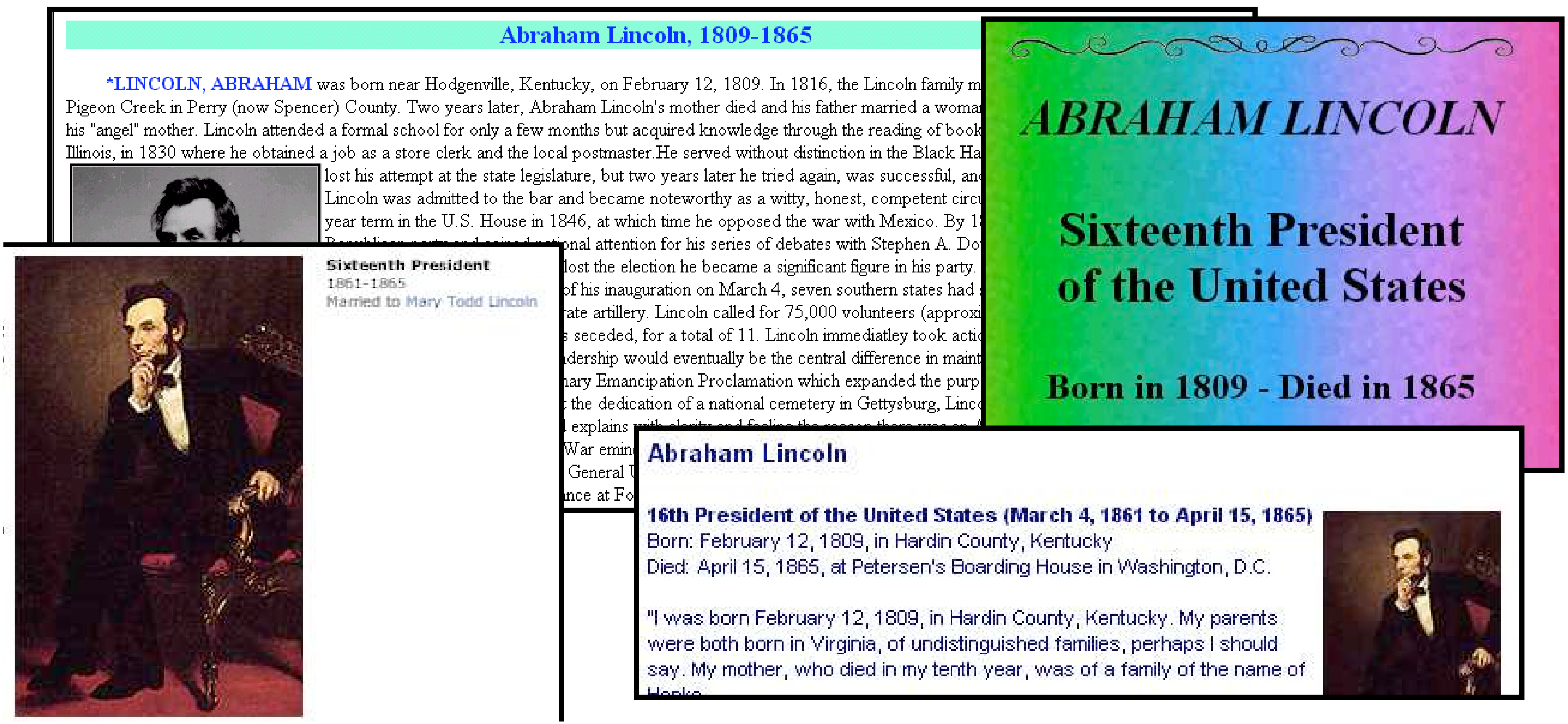
- ***Web Question Answering: Is More Always Better?***
  - Dumais, Banko, Brill, Lin, Ng (Microsoft, MIT, Berkeley)

- **Q: “Where is the Louvre located?”**
- Want “Paris” or “France” or “75058 Paris Cedex 01” or a map
- Don’t just want URLs

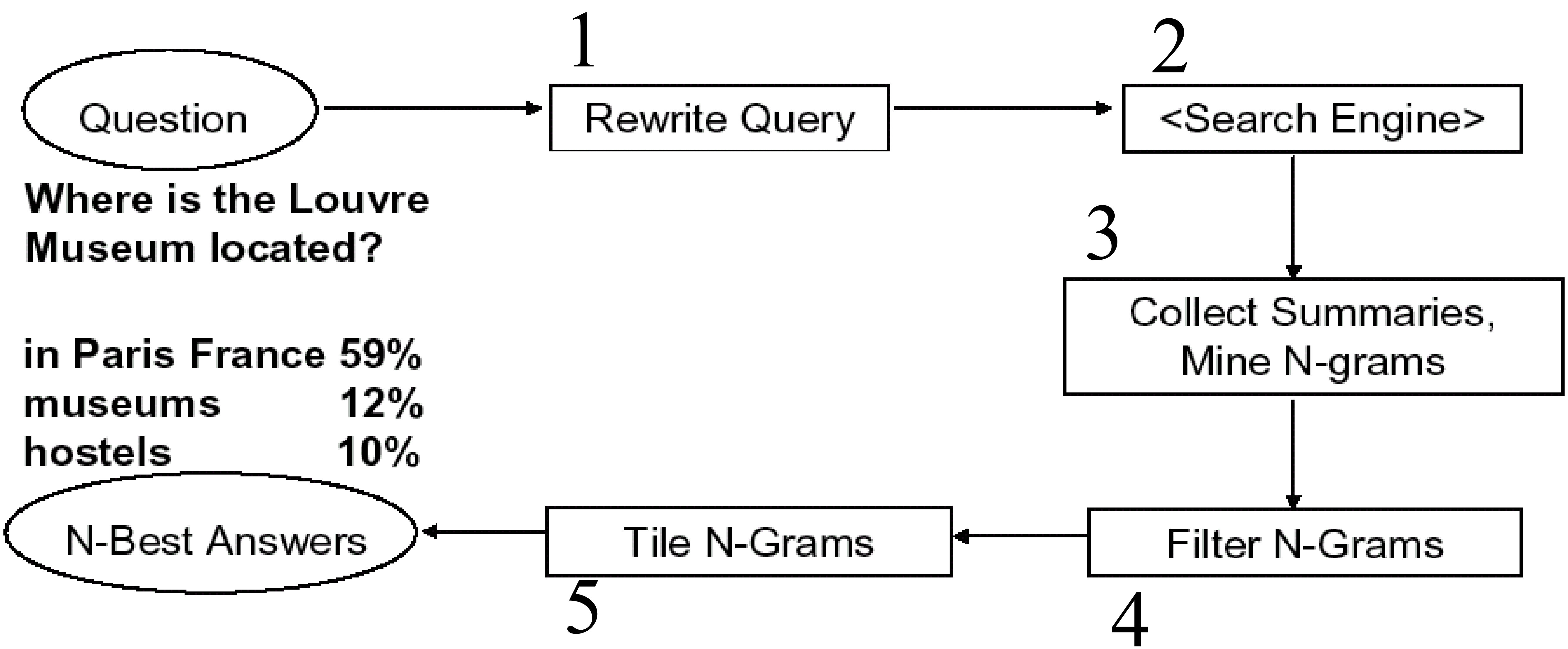


AskMSR: Shallow approach

- *In what year did Abraham Lincoln die?*
- Ignore hard documents and find easy ones



AskMSR: Details



Step 1: Rewrite queries

- Intuition: The user's question is often syntactically quite close to sentences that contain the answer
  - Where is the Louvre Museum located?
  - The Louvre Museum is located in ***Paris***
  - Who created the character of Scrooge?
  - ***Charles Dickens*** created the character of Scrooge.

Query Rewriting: Variations

- Classify question into seven categories
  - **Who** is/was/are/were…?
  - **When** is/did/will/are/were …?
  - **Where** is/are/were …?

  a. Category-specific transformation rules
  eg "For Where questions, move 'is' to all possible locations"
  "Where is the Louvre Museum located"
  → "is the Louvre Museum located"
  → "the is Louvre Museum located"
  → "the Louvre is Museum located"
  → "the Louvre Museum is located"
  → "the Louvre Museum located is"

  Nonsense, but who cares? It's only a few more queries

  b. Expected answer "Datatype" (eg, Date, Person, Location, …)
  When was the French Revolution? → DATE

- Hand-crafted classification/rewrite/datatype rules
  (Could they be automatically learned?)

Query Rewriting: Weights

- One wrinkle: Some query rewrites are more reliable than others

Where is the Louvre Museum located?

Weight 1
Lots of non-answers
could come back too

Weight 5
if we get a match,
it's probably right

+"the Louvre Museum is located"

+Louvre +Museum +located

Step 2: Query search engine

- Send all rewrites to a search engine
- Retrieve top N answers (100?)
- For speed, rely just on search engine's "snippets", not the full text of the actual document

Step 3: Mining N-Grams

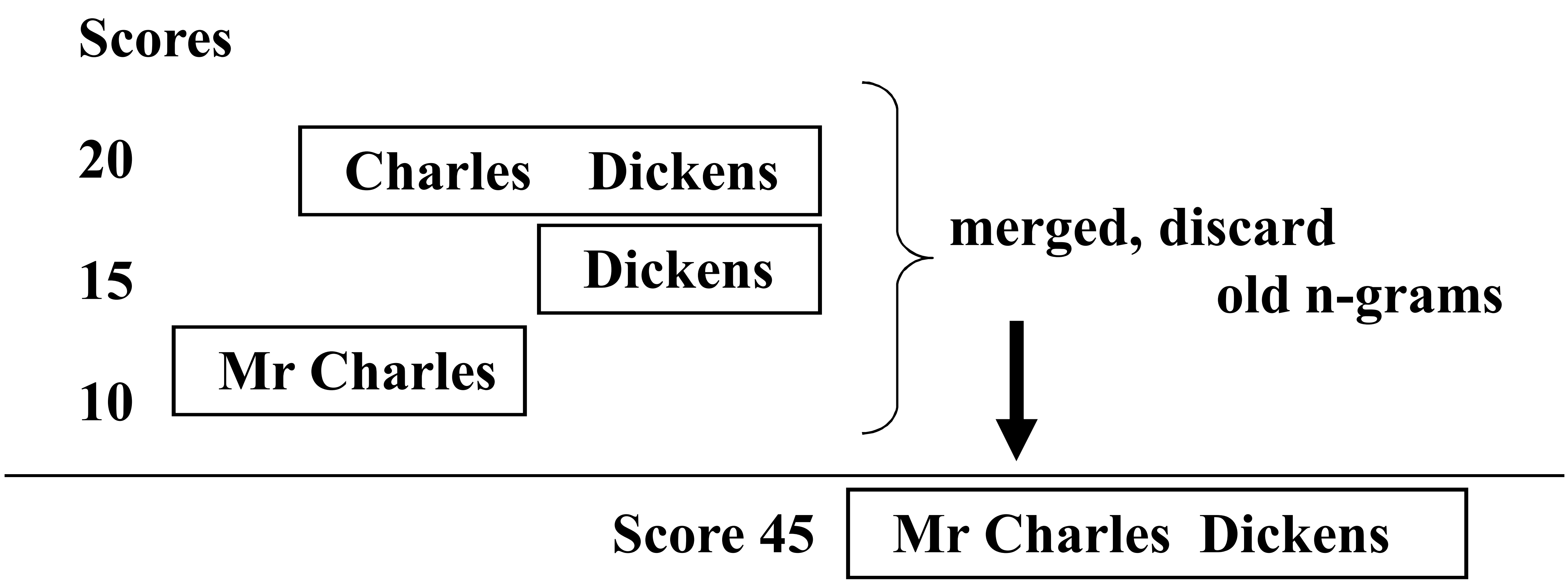
- Simple: Enumerate all N-grams (N=1,2,3 say) in all retrieved snippets
- Weight of an n-gram: occurrence count, each weighted by “reliability” (weight) of rewrite that fetched the document
- Example: “Who created the character of Scrooge?”
  - Dickens - 117
  - Christmas Carol - 78
  - Charles Dickens - 75
  - Disney - 72
  - Carl Banks - 54
  - A Christmas - 41
  - Christmas Carol - 45
  - Uncle - 31

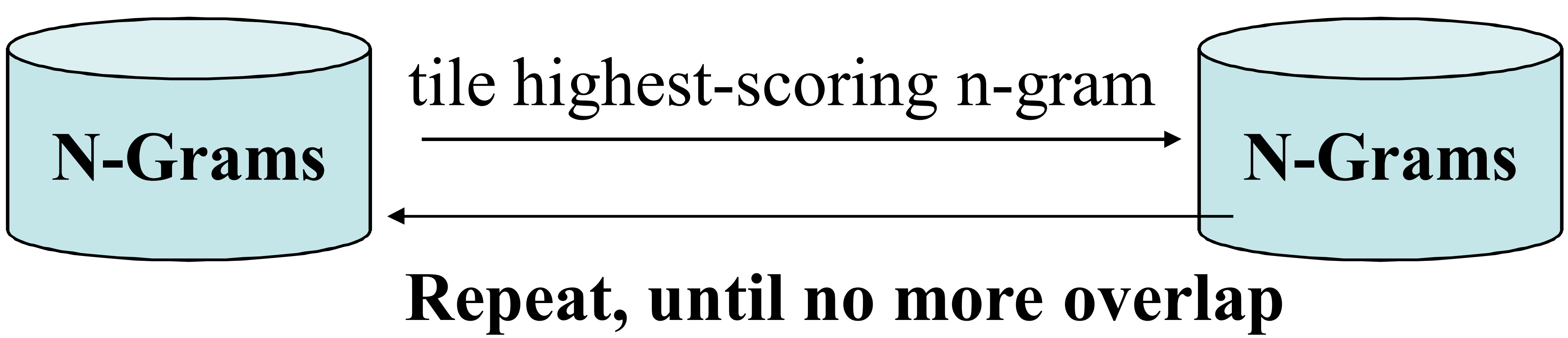
Step 4: Filtering N-Grams

- Each question type is associated with one or more "**data-type filters**" = regular expression
- When… Date
- Where… Location
- What …
- Who … Person

- Boost score of n-grams that do match regexp
- Lower score of n-grams that don't match regexp
- Details omitted from paper….

Step 5: Tiling the Answers





Results

- Standard TREC contest test-bed:
  ~1M documents; 900 questions

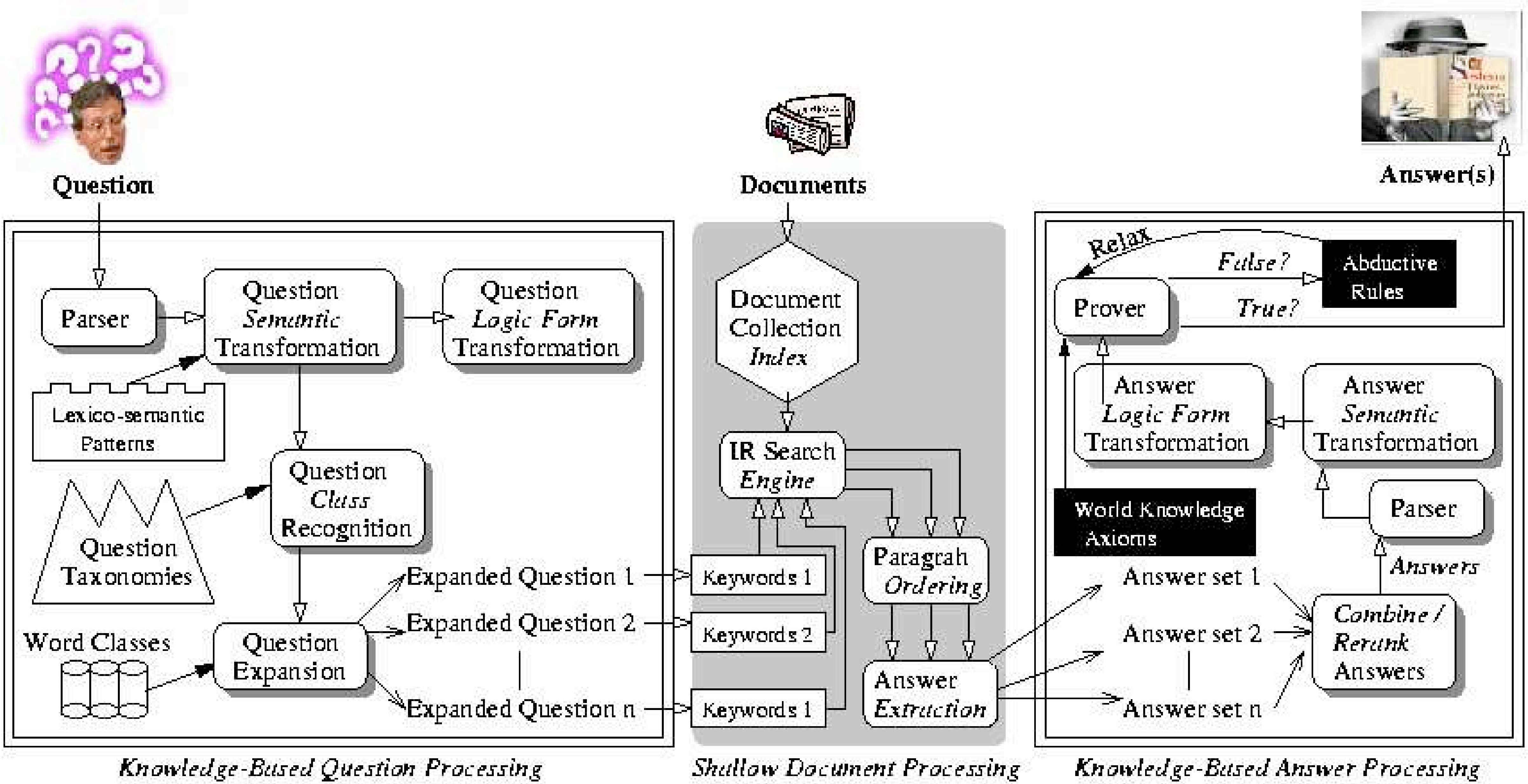
- Technique doesn't do too well (though would have placed in top 9 of ~30 participants!)
  - MRR = 0.262 (ie, right answered ranked about #4-#5 on average)
  - Why? Because it relies on the redundancy of the Web

- Using the Web as a whole, not just TREC's 1M documents… MRR = 0.42 (ie, on average, right answer is ranked about #2-#3)

Issues

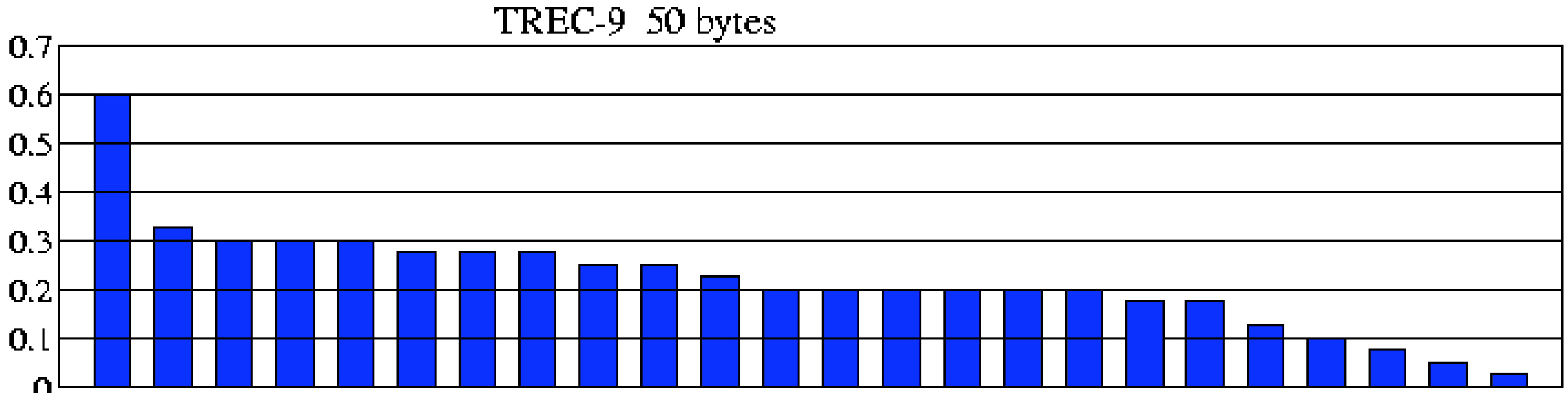
- In many scenarios (e.g., monitoring an individuals email…) we only have a small set of documents
- Works best/only for “Trivial Pursuit”-style fact-based questions
- Limited/brittle repertoire of
  - question categories
  - answer data types/filters
  - query rewriting rules

LCC: Harabagiu, Moldovan et al.



Value from Sophisticated NLP
Pasca and Harabagiu (2001)

- Good IR is needed: SMART paragraph retrieval
- Large taxonomy of question types and expected answer types is crucial
- Statistical parser used to parse questions and relevant text for answers, and to build KB
- Query expansion loops (morphological, lexical synonyms, and semantic relations) important
- Answer ranking by simple ML method



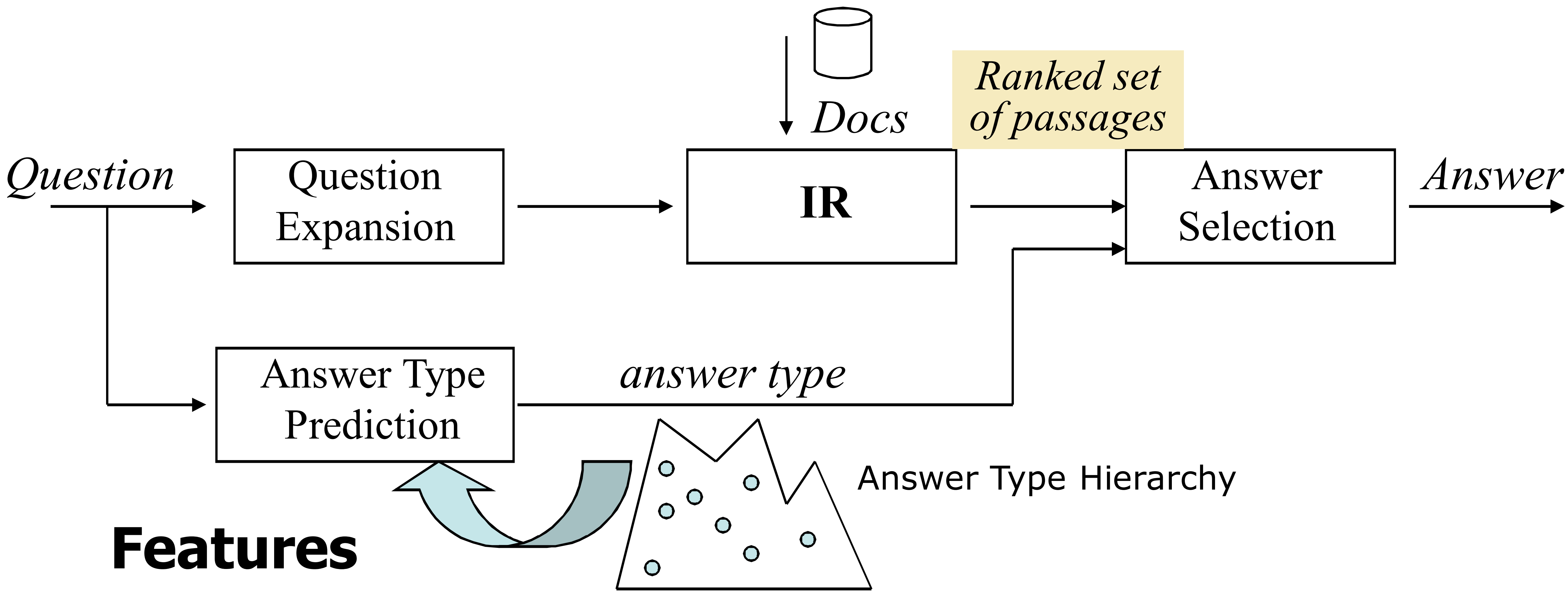
Abductive inference

- System attempts inference to justify an answer (often following lexical chains)
- Their inference is a kind of funny middle ground between logic and pattern matching
- But quite effective: 30% improvement
- *Q: When was the internal combustion engine invented?*
- *A: The first internal-combustion engine was built in 1867.*
- invent -> create_mentally -> create -> build

Question Answering Example

- How hot does the inside of an active volcano get?
- get(TEMPERATURE, inside(volcano(active)))
- “lava fragments belched out of the mountain were as hot as 300 degrees Fahrenheit”
- fragments(lava, TEMPERATURE(degrees(300)), belched(out, mountain))
  - volcano ISA mountain
  - lava ISPARTOF volcano
  - lava inside volcano
  - fragments of lava HAVEPROPERTIESOF lava
- The needed semantic information is in WordNet definitions, and was successfully translated into a form that was used for rough ‘proofs’

Answer types in SOA QA systems



Features

- Answer type
  - Labels questions with answer type based on a taxonomy
  - Classifies questions (e.g. by using a maximum entropy model)

QA Typology (from ISI USC)

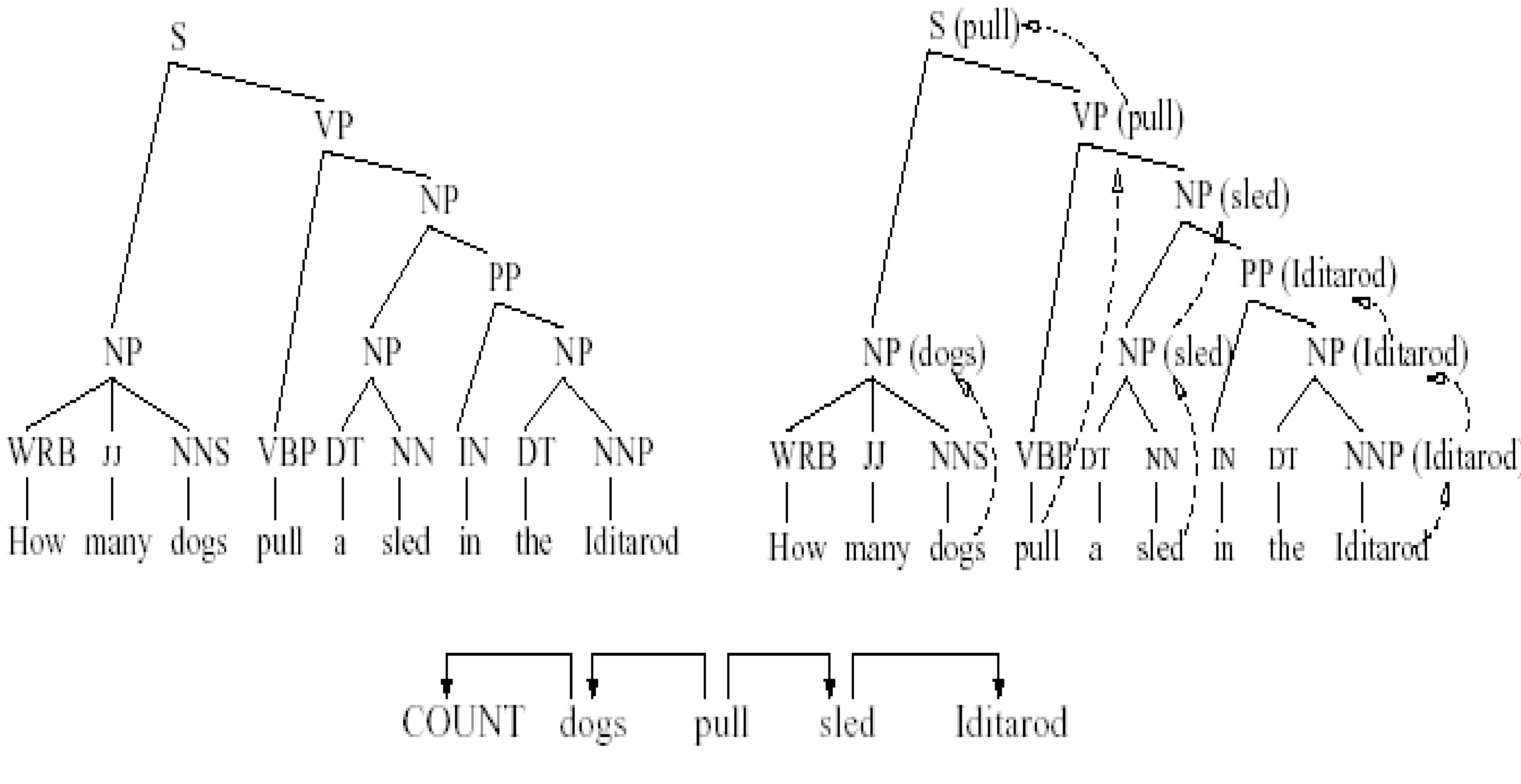
- Typology of typical Q forms—94 nodes (47 leaf nodes)
- Analyzed 17,384 questions (from answers.com)

```
(THING
 ((AGENT
   (NAME (FEMALE-FIRST-NAME (EVE MARY ...))
         (MALE-FIRST-NAME (LAWRENCE SAM ...))))
         (COMPANY-NAME (BOEING AMERICAN-EXPRESS))
         JESUS ROMANOFF ...)
   (ANIMAL-HUMAN (ANIMAL (WOODCHUCK YAK ...))
                 PERSON)
   (ORGANIZATION (SQUADRON DICTATORSHIP ...))
   (GROUP-OF-PEOPLE (POSSE CHOIR ...))
   (STATE-DISTRICT (TIROL MISSISSIPPI ...))
   (CITY (ULAN-BATOR VIENNA ...))
   (COUNTRY (SULTANATE ZIMBABWE ...))))
 (PLACE
   (STATE-DISTRICT (CITY COUNTRY...))
   (GEOLOGICAL-FORMATION (STAR CANYON...))
   AIRPORT COLLEGE CAPITOL ...)
 (ABSTRACT
  (LANGUAGE (LETTER-CHARACTER (A B ...)))
  (QUANTITY
   (NUMERICAL-QUANTITY INFORMATION-QUANTITY
   MASS-QUANTITY MONETARY-QUANTITY
   TEMPORAL-QUANTITY ENERGY-QUANTITY
   TEMPERATURE-QUANTITY ILLUMINATION-QUANTITY
   (SPATIAL-QUANTITY
    (VOLUME-QUANTITY AREA-QUANTITY DISTANCE-QUANTITY)) ...
     PERCENTAGE)))
 (UNIT
  ((INFORMATION-UNIT (BIT BYTE ... EXABYTE))
   (MASS-UNIT (OUNCE ...)) (ENERGY-UNIT (BTU ...))
   (CURRENCY-UNIT (ZLOTY PESO ...))
   (TEMPORAL-UNIT (ATTOSECOND ... MILLENIUM))
   (TEMPERATURE-UNIT (FAHRENHEIT KELVIN CELCIUS))
   (ILLUMINATION-UNIT (LUX CANDELA))
   (SPATIAL-UNIT
    ((VOLUME-UNIT (DECILITER ...))
     (DISTANCE-UNIT (NANOMETER ...))))
     (AREA-UNIT (ACRE)) ... PERCENT))
 (TANGIBLE-OBJECT
  ((FOOD (HUMAN-FOOD (FISH CHEESE ...)))
   (SUBSTANCE
    ((LIQUID (LEMONADE GASOLINE BLOOD ...))
     (SOLID-SUBSTANCE (MARBLE PAPER ...))
     (GAS-FORM-SUBSTANCE (GAS AIR)) ...))
   (INSTRUMENT (DRUM DRILL (WEAPON (ARM GUN)) ...)
   (BODY-PART (ARM HEART ...))
   (MUSICAL-INSTRUMENT (PIANO)))
   ... *GARMENT *PLANT DISEASE)
```

Named Entity Recognition for QA

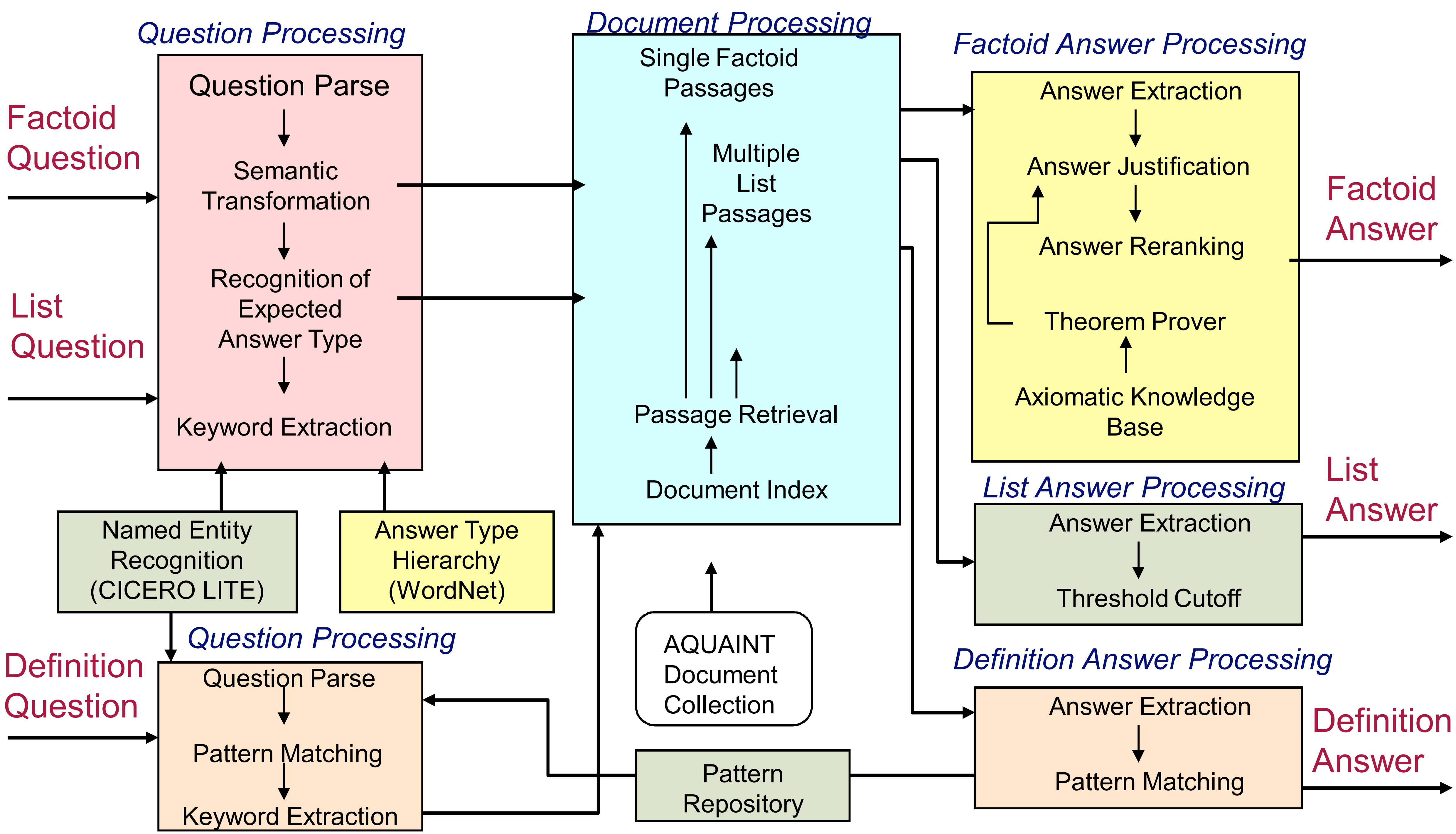
- The results of the past 5 TREC evaluations of QA systems indicate that current state-of-the-art QA is determined by the recognition of Named Entities:
  - *Precision of recognition*
  - *Coverage of name classes*
  - *Mapping into concept hierarchies*
  - *Participation into semantic relations (e.g. predicate-argument structures or frame semantics)*

Syntax to Logical Forms



- Syntactic analysis plus semantic => logical form
- Mapping of question and potential answer LFs to find the best match

The Architecture of LCC's QA System around 2003



Answering definition questions

- Most QA systems use between 30-60 patterns
- The most popular patterns:

Id	Pattern	Freq.	Usage	Question
25	person-hyponym QP	0.43%	The doctors also consult with former Italian Olympic skier Alberto Tomba, along with other Italian athletes	1907: Who is Alberto Tomba?
9	QP, the AP	0.28%	Bausch Lomb, the company that sells contact lenses, among hundreds of other optical products, has come up with a new twist on the computer screen magnifier	1917: What is Bausch & Lomb?
11	QP, a AP	0.11%	ETA, a Basque language acronym for Basque Homeland and Freedom _ has killed nearly 800 people since taking up arms in 1968	1987: What is ETA in Spain?
13	QA, an AP	0.02%	The kidnappers claimed they are members of the Abu Sayaf, an extremist Muslim group, but a leader of the group denied that	2042: Who is Abu Sayaf?
21	AP such as QP	0.02%	For the hundreds of Albanian refugees undergoing medical tests and treatments at Fort Dix, the news is mostly good: Most are in reasonable good health, with little evidence of infectious diseases such as TB	2095: What is TB?

Example of Complex Question

How have thefts impacted on the safety of Russia's nuclear navy, and has the theft problem been increased or reduced over time?

Need of domain knowledge

To what degree do different thefts put nuclear or radioactive materials at risk?

Question decomposition

Definition questions:
- *What is meant by nuclear navy?*
- *What does 'impact' mean?*
- *How does one define the increase or decrease of a problem?*

Factoid questions:
- *What is the number of thefts that are likely to be reported?*
- *What sort of items have been stolen?*

Alternative questions:
- *What is meant by Russia? Only Russia, or also former Soviet facilities in non-Russian republics?*

Complex questions

- Characterized by the need of domain knowledge
- There is no single answer type that can be identified, but rather an answer structure needs to be recognized
- Answer selection becomes more complicated, since inference based on the semantics of the answer type needs to be activated
- Complex questions need to be decomposed into a set of simpler questions