

<http://www.phdcomics.com/comics.php>

# Information Extraction

David Kauchak

cs160

Fall 2009

*some content adapted from:*

<http://www.cs.cmu.edu/~knigam/15-505/ie-lecture.ppt>

# Administrative

- Colloquium tomorrow (substitute class)
  - Thursday, 4:15pm in Rose Hill Theater
  - Make-up assignment

# A problem

baker job opening - Google Search - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.google.com/search?ie=UTF-8&oe=UTF-8&sourceid=gd&q=baker+job+op

Google Web Images Video <sup>New!</sup> News Maps Desktop Moma more »

baker job opening Search Advanced Search Preferences

Web Results 1 - 10 of about 6,150,000 for **baker job opening**. (0.09 seconds)

**Mt Baker School District**  
You may also call 360-383-2075 for a voice message concerning our **job** listings. Our district applications may be downloaded from each **job** category site. ...  
[www.mtbaker.wednet.edu/jobs/](http://www.mtbaker.wednet.edu/jobs/) - 3k - [Cached](#) - [Similar pages](#) - [Filter](#)

**CGI : Job Opening**  
**Job** Seekers, Faculty & Other Researchers, Students, Journalists, Policy Makers ... **Baker** Institute for Animal Health, College of Veterinary Medicine ...  
[www.genomics.cornell.edu/jobs/view\\_job.cfm?id=47](http://www.genomics.cornell.edu/jobs/view_job.cfm?id=47) - 15k - [Cached](#) - [Similar pages](#) - [Filter](#)

**Baker Hostetler - Staff Job Openings**  
law business employee benefits employment intellectual property international legislative regulatory litigation private wealth real estate tax automotive ...  
[www.bakerlaw.com/Careers.aspx?Abs\\_WP\\_ID=26a8ff33-0471-4c5e-b5b7-6abdcfce0326](http://www.bakerlaw.com/Careers.aspx?Abs_WP_ID=26a8ff33-0471-4c5e-b5b7-6abdcfce0326) - 19k - [Cached](#) - [Similar pages](#) - [Filter](#)

**Baker & McKenzie || Careers || Current Openings ||**  
We are always looking for talented, internationally minded people interested in building their careers with a truly global law firm.  
[www.bakernet.com/BakerNet/Careers/Current+Openings/](http://www.bakernet.com/BakerNet/Careers/Current+Openings/) - 64k - [Cached](#) - [Similar pages](#) - [Filter](#)

**Current Job Opening Search**  
Click the search button to see all **job openings**. ... Apprentice **Baker**, Architect - Production, Architectural Drafting Intern, Architectural Project Leader ...  
[hyveenet.hy-vee.com/applynow/](http://hyveenet.hy-vee.com/applynow/) - 75k - [Cached](#) - [Similar pages](#) - [Filter](#)

**Law Enforcement Job Submission**  
Advertise Your **Job Openings** ... -Mia **Baker**, Human Resources Officer, Amtrak ... You can announce your **job opening** to thousands of potential applicants at a ...  
[www.policeemployment.com/joblisting/](http://www.policeemployment.com/joblisting/) - 10k - [Cached](#) - [Similar pages](#) - [Filter](#)

<http://www.bakernet.com/BakerNet/Careers/Current+Openings/>

Sponsored Links

Talent - Meet Your Hiring Needs!  
[www.Monster.com](http://www.Monster.com)

**Baker Job**  
Search Thousands of **Job** Listings for Opportunities in Your Area  
[Jobs.AOL.com](http://Jobs.AOL.com)

Keyword - New Listings Daily!  
[www.hotjobs.com](http://www.hotjobs.com)

**Find Bakers Jobs**  
CareerBuilder® Has More Jobs In Hospitality Than Any Other Site!  
[CareerBuilder.com/Baker\\_Jobs](http://CareerBuilder.com/Baker_Jobs)

**Baker, a job opening**  
[www.AreaGuides.net](http://www.AreaGuides.net)

**i Hire Chefs**  
Chef Jobs - Find a Culinary Arts **Job** Nationwide Employment Opportunities  
[www.iHireChefs.com](http://www.iHireChefs.com)

**Mt. Baker, the school district**

**Baker Hostetler, the company**

**Baker, a job opening**

# A solution


job search find employment careers @ FlipDog.com free! - Microsoft Internet Explorer

Address <http://www.flipdog.com/home.html> Go File Edit View Favorites Tools Help Links >>

**FlipDog.com**


Home Find Jobs Your Account Resource Center • Support • Employers

Job Search at FlipDog.com: Employment & Career Management

 **647,514**  
Job Opportunities  
from **53,641** Employers

[Find a Job!](#)

[Post Your Resume](#)

**Employers**  
click here for  
Products & Services 


**Pigskin Places**

- Health Care in NY [2,770](#)
- Health Care in MD [1,262](#)
- Sales in NY [3,751](#)
- Sales in MD [958](#)
- Computing in NY [8,050](#)
- Computing in MD [4,114](#)

**Jobs for Sports Fans**


- [Head Football Coach](#)
- [Football Coach](#)
- [Asst. Football Coach](#)
- [High School Football Coach](#)
- [Univ. Asst. Football Coach](#)

**Showcase Jobs**

  
Management Recruiters  
of Charlotte North

We provide total staffing solutions in the areas of Human Resources, Compensation, Web-based HR self-service, and Customer Management Systems.

[Learn More](#)



Looking for a Vice President of Academic Affairs to oversee planning, operation and evaluation of the college's academic programs.

[Learn More](#)


**Job Seekers: Find your dream job!**


- ▶ Check our 'Best Places to Find a Job' [January report](#).
- ▶ Open your [FREE account](#) and put your [resume online](#).
- ▶ Search 24x7 with our FREE automatic [JobHunters™](#).
- ▶ Research our database of over [50,000 employers](#).
- ▶ Get [expert advice](#) at our new [Resource Center](#).
- ▶ Access [salary surveys/calculators](#), [relocation tools](#), [networking opportunities](#), & [training/testing](#) tools.
- ▶ Use FlipDog.com to search jobs right from your desktop! Download [Snippets](#) today!


**Job Seeker Newsletter**

Enter your e-mail address:

[Sign Me Up!](#)

 "Top 100 Web Sites"  
PC Magazine, Nov. 2000

 "Top 10 Career Web Site"  
Media Metrix, Sept. 2000

 "Top 10 Job Site"

powered by **WhizBang!**

Internet

Why is this better?

How does it happen?



FlipDog.com

Fetch Your Next Job Here™

Home

Find Jobs

Your Account

Resource Center

Employers Support

Return to Results | Modify Search | New Search



Learn While You Earn  
MBA, BA, AA Degrees  
Online & Project Mgt.

Click here to e-mail your resume to 1000's  
of Head Hunters with  
ResumeZapper.com



Breakthrough ebook  
shows why most people  
are WRONG about how  
to apply for jobs.

1 - 25 of 47 jobs shown below

1 2 Next >

Search these results for:



Search tips

Show Jobs Posted:

For all time periods

View: Brief | Detailed

Web Jobs: FlipDog technology has found these jobs on thousands of employer Web sites.

<a href="#">Food Pantry Workers</a> at <a href="#">Lutheran Social Services</a>	October 11, 2002	<a href="#">Archbold, OH</a>
<a href="#">Cooks</a> at <a href="#">Lutheran Social Services</a>	October 11, 2002	<a href="#">Archbold, OH</a>
<a href="#">Bakers Assistants</a> at <a href="#">Fine Catering by Russell Morin</a>	October 11, 2002	<a href="#">Attleboro, MA</a>
<a href="#">Baker's Helper</a> at <a href="#">Bird-in-Hand</a>	October 11, 2002	United States
<a href="#">Assistant Baker</a> at <a href="#">Gourmet To Go</a>	October 11, 2002	<a href="#">Maryland Heights, MO</a>
	October 10, 2002	<a href="#">Beaverton, OR</a>
	October 10, 2002	<a href="#">Alta, UT</a>
	October 10, 2002	<a href="#">Huntsville, UT</a>
<a href="#">ed School District</a>	October 10, 2002	<a href="#">Garden Grove, CA</a>
	October 10, 2002	<a href="#">Houma, LA</a>
	October 10, 2002	<a href="#">Nisswa, MN</a>
<a href="#">Line Cook</a> at <a href="#">Lone Mountain Ranch</a>	October 10, 2002	<a href="#">Big Sky, MT</a>
<a href="#">Production Baker</a> at <a href="#">Whole Foods Market</a>	October 08, 2002	<a href="#">Willowbrook, IL</a>
<a href="#">Cake Decorator/Baker</a> at <a href="#">Mandalay Bay Hotel and Casino</a>	October 08, 2002	<a href="#">Las Vegas, NV</a>
<a href="#">Shift Supervisors</a> at <a href="#">Brueggers Bagels</a>	October 08, 2002	<a href="#">Minneapolis, MN</a>

**Job Openings:**  
**Category = Food Services**  
**Keyword = Baker**  
**Location = Continental U.S.**

# Extracting Job Openings from the Web

OPUS International, Inc., an executive search firm focusing on the Food Science industry. - Microsoft Internet Explorer

File Edit View Favorites

Back Forward Stop

Address <http://www.foodsci...>

Links AMEX Rewards T

OPUS: Job Listings - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites

Address [http://www.foodscience.com/jobs\\_midwest.html#top](http://www.foodscience.com/jobs_midwest.html#top)

Links AMEX Rewards Time DogHouse My Yahoo!

Welcome

About OPUS

Executive Staff

**Job Listings**

Résumé Form

Job Hunt Hints

Academic Links

Science Fair Help

Industry Assocs.

FAQs

Contact Us

Site Map

e-mail

OPUS INTERNATIONAL INC.

About | Staff | Job

Welcome

About OPUS

Executive Staff

**Job Listings**

Résumé Form

Job Hunt Hints

Academic Links

Science Fair Help

Industry Assocs.

FAQs

Contact Us

Site Map

e-mail

**Test Kitchen-  
Consumer Food Relations**

Major food manufacturer in Chicago area seeks a consumer food professional to write and test recipes. Will make presentations to marketing; will be a key player on a cross-functional team. Requires a BS in human ecology, nutrition, Food Science, or related field, plus a minimum three years applicable experience.  
Contact Moira: e-mail 1-800-488-2611

**Ice Cream Guru**

If you dream of cold creamy chocolate or coochy coochy cookie, there's a great opportunity for you to maintain and expand this major corporation's high-end ice cream brand. Will be based in the Upper Midwest for about a year. After that, California here I come! Requires a BS in Food Science or dairy, plus ice cream formulation experience. Will consider entry level with an MS and an internship.  
Contact Susan: e-mail 1-800-488-2611

**Title:** Ice Cream Guru

**Description:** If you dream of cold creamy...

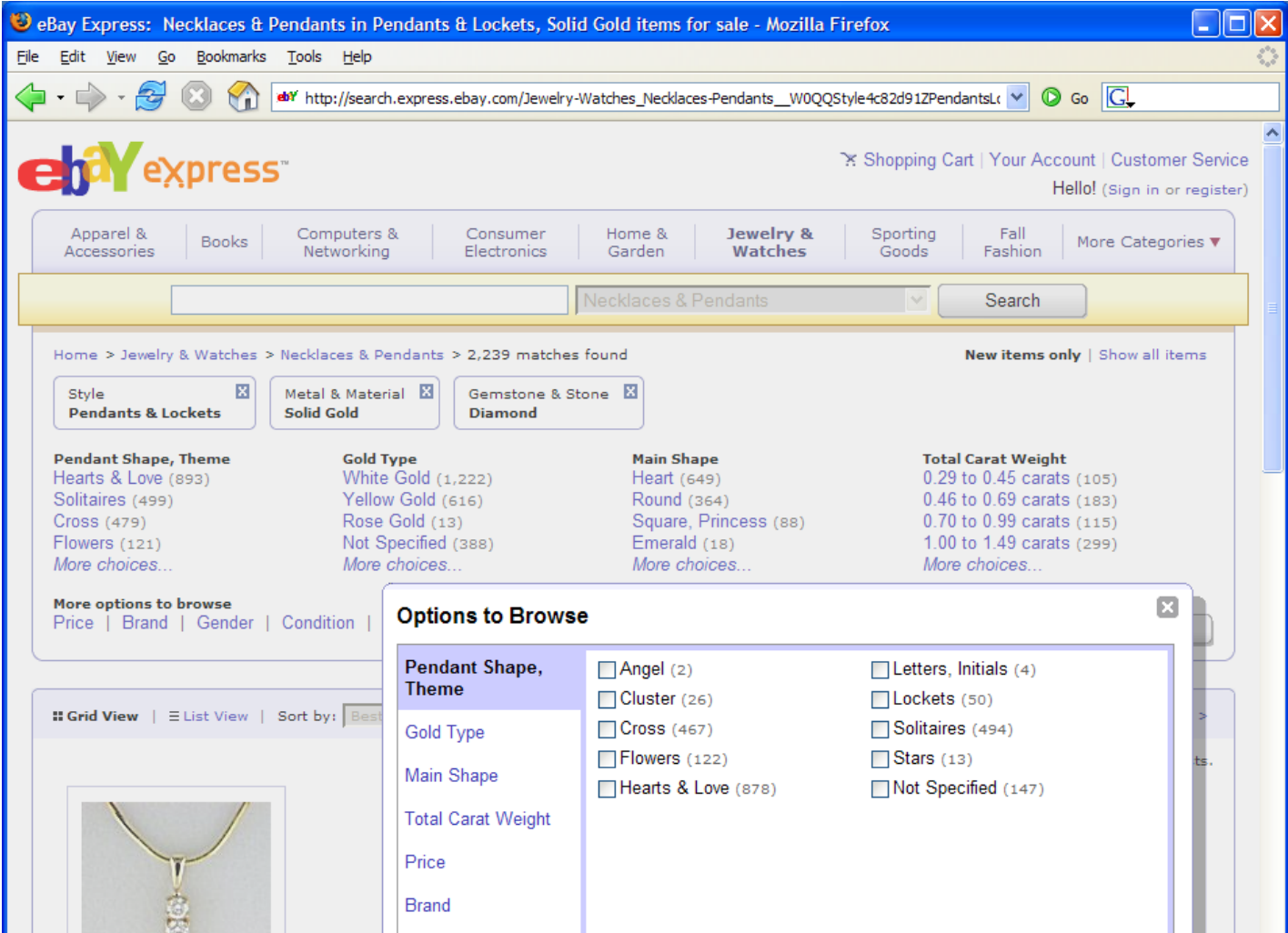
**Contact:** [susan@foodscience.com](mailto:susan@foodscience.com)

**Category:** Travel/Hospitality

**Function:** Food Services



# Potential Enabler of Faceted Search






# Often structured information in text


eBay Express: 0.44 CT ROUND CUT DIAMOND PENDANT 14 K WHITE GOLD - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://item.express.ebay.com/Jewelry-Watches\_Necklaces-Pendants\_\_0-44-CT-ROUND-CUT-DIAMON



Click to Enlarge





Click to Enlarge

**0.44 CT ROUND CUT DIAMOND  
PENDANT 14 K WHITE GOLD** Classic style and beauty, this comfortable 14 K White gold pendant contains:  
**An Ideal cut Round 0.44 CT Diamond, in a magnificent high polish bezel.**

- Color: F
- Clarity: SI-1
- Setting: 14 K White Gold
- Chain: 16 inches 14 K White Gold
- Weight: 3.4 g
- Measurements: 10 mm x 10 mm

- Retail Price: \$2,319.00  
- Close Out Price: \$889.00



Done

# Research papers

**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

Download: [norvig.com/coling.ps](http://norvig.com/coling.ps)  
Cached: [PS.gz](#) [PS](#) [PDF](#) [DjVu](#) [Image](#) [Update](#) [Help](#)

From: [norvig.com/resume](http://norvig.com/resume) (more)  
Home: [R.Wilensky](#) [HPSearch](#) (Correct)

**NEC ResearchIndex** [Bookmark](#) [Context](#) [Related](#)

[\(Enter summary\)](#)

Rate this article: 1 2 3 4 5 (best)  
[Comment on this article](#)

**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 commensurability: all types of evidence are represented in a common currency that can be compared and combined. While commensurability is a desirable property, and there is a clear need for a way to compare alternate explanations, it appears that a single scalar measure is not enough to account for all types of processing. We present other problems for the 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 [Norvig and Wilensky, 1990] of 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 Wilensky \(1990\)](#). The use of 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)  
0.1: [Decision Analytic Networks in Artificial Intelligence - Matzkevich, Abramson \(1995\)](#) (Correct)  
0.1: [A Deshabilitic Network of Deductio... DeLong, Liu \(1992\)](#) (Correct)

# What is Information Extraction?

## Traditional definition:

- Recovering structured data from formatted text

Management Team
<b>Board of Directors</b>
Our Firm & WOMMA
FAQs
Contact Us
Careers

## Board Members

- **Itzhak Fisher**  
Chairman of Nielsen BuzzMetrics
- **Thom Mastrelli**  
Executive Vice President/Corporate Development, VNU
- **Jonathan Carson**  
CEO of Nielsen BuzzMetrics
- **Mahendra Vora**  
CEO and Owner, Vora Technology Park
- **Ori Levy**  
President of Nielsen BuzzMetrics Israel
- **Ron Schneier**  
Senior Vice President and General Manager, Nielsen Ventures
- **James O'Hara**  
Senior Vice President and Chief Financial Officer, VNU's Media Measurement and Information Group

# What is Information Extraction?

- Recovering structured data from formatted text
  - Identifying fields (e.g. named entity recognition)

Management Team
<b>Board of Directors</b>
Our Firm & WOMMA
FAQs
Contact Us
Careers

## Board Members

◦ <b>Itzhak Fisher</b> Chairman of Nielsen BuzzMetrics	◦ <b>Ori Levy</b> President of Nielsen BuzzMetrics Israel
◦ <b>Thom Mastrelli</b> Executive Vice President/Corporate Development, VNU	◦ <b>Ron Schneier</b> Senior Vice President and General Manager, Nielsen Ventures
◦ <b>Jonathan Carson</b> CEO of Nielsen BuzzMetrics	◦ <b>James O'Hara</b> Senior Vice President and Chief Financial Officer, VNU's Media Measurement and Information Group
◦ <b>Mahendra Vora</b> CEO and Owner, Vora Technology Park	

# What is Information Extraction?

- Recovering structured data from formatted text
  - Identifying fields (e.g. named entity recognition)
  - Understanding relations between fields (e.g. record association)

Management Team
<b>Board of Directors</b>
Our Firm & WOMMA
FAQs
Contact Us
Careers

## Board Members

◦ <b>Itzhak Fisher</b> Chairman of Nielsen BuzzMetrics	◦ <b>Ori Levy</b> President of Nielsen BuzzMetrics Israel
◦ <b>Thom Mastrelli</b> Executive Vice President/Corporate Development, VNU	◦ <b>Ron Schneier</b> Senior Vice President and General Manager, Nielsen Ventures
◦ <b>Jonathan Carson</b> CEO of Nielsen BuzzMetrics	◦ <b>James O'Hara</b> Senior Vice President and Chief Financial Officer, VNU's Media Measurement and Information Group
◦ <b>Mahendra Vora</b> CEO and Owner, Vora Technology Park	

# What is Information Extraction?

- Recovering structured data from formatted text
  - Identifying fields (e.g. named entity recognition)
  - Understanding relations between fields (e.g. record association)
  - Normalization and deduplication

**James O'Hara (I)**

 **No Photo Available**  
add photo

Date of birth (location)  
[11 September 1927](#)  
[Dublin, Ireland](#)

Date of death (details)  
[3 December 1992](#)  
Glendale, California, USA.

Trivia  
Brother of [Maureen O'Hara](#)

Sometimes Credited As:  
James Lilburn / Jim O'Hara

 [IMDbPro Details](#)  [Add IMDb Resume](#)

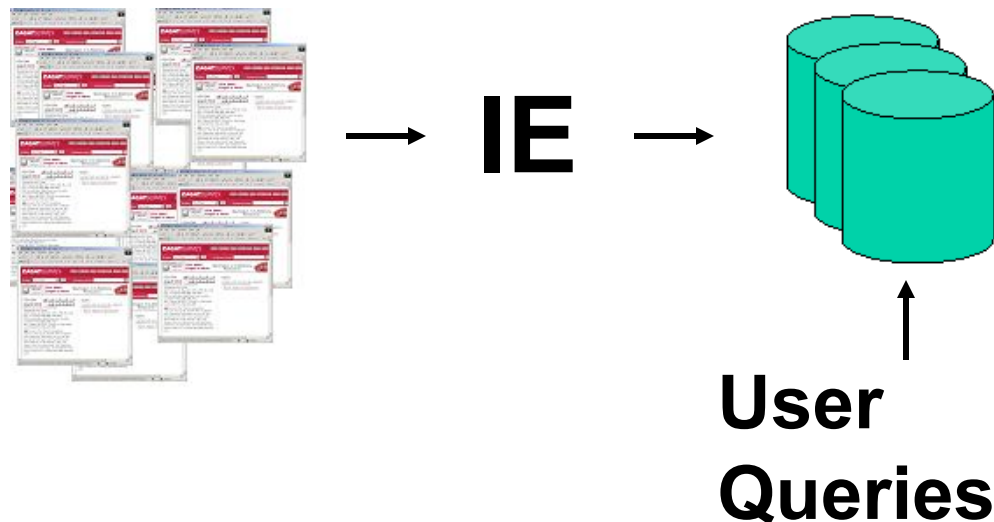
- **James O'Hara**  
Senior Vice President and Chief Financial Officer, VNU's Media Measurement and Information Group

Herkovic  
Susan D. Whiting  
Douglas Darfield  
Paul J. Donato  
Sara Erichson  
Dave Harkness  
Jack Loftus

Jane is a member of the Nielsen senior leadership team and a senior member of the VNU MMI Finance team. She is based in New York and reports to both Susan Whiting, president and CEO of Nielsen Media Research, and [Jim O'Hara](#), senior vice president and chief financial officer for VNU Media Measurement and Information.

# What is information extraction?

- Input: Text Document
  - Various sources: web, e-mail, journals, ...
- Output: Relevant fragments of text, possibly to be processed later in some automated way



# Not all documents are created equal...



- Varying regularity in document collections
- Natural or unstructured
  - Little obvious structural information
- Partially structured
  - Contain some canonical formatting
- Highly structured
  - Often, automatically generated



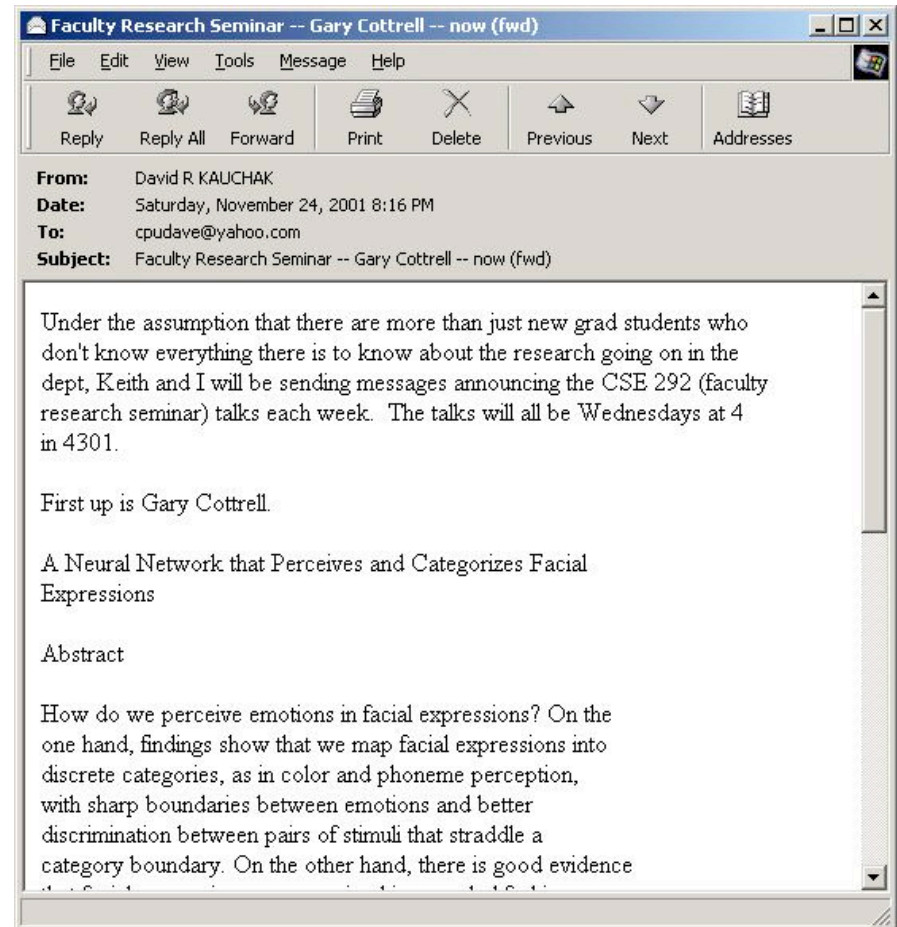
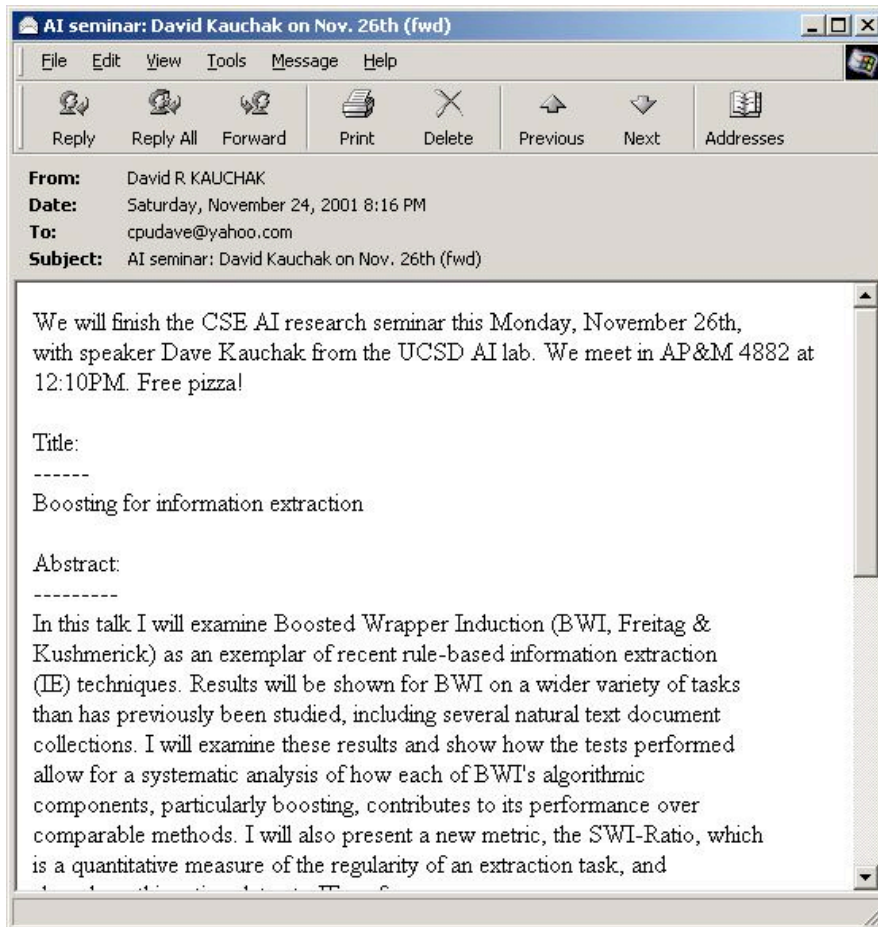
# Natural Text: MEDLINE Journal Abstracts

Extract number of subjects, type of study, conditions, etc.

**BACKGROUND:** The most challenging aspect of revision hip surgery is the management of bone loss. A reliable and valid measure of bone loss is important since it will aid in future studies of hip revisions and in preoperative planning. We developed a measure of femoral and acetabular bone loss associated with failed total hip arthroplasty. The purpose of the present study was to **measure the reliability and the intraoperative validity of this measure** and to determine how it may be useful in preoperative planning. **METHODS:** From July 1997 to December 1998, **forty-five consecutive patients** with a failed hip prosthesis in need of revision surgery were prospectively followed. Three general orthopaedic surgeons were taught the radiographic classification system, and two of them classified standardized preoperative anteroposterior and lateral hip radiographs with use of the system. Interobserver testing was carried out in a **blinded fashion**. These results were then compared with the intraoperative findings of the third surgeon, who was blinded to the preoperative ratings. Kappa statistics (unweighted and weighted) were used to assess correlation. Interobserver reliability was assessed by examining the agreement between the two preoperative raters. Prognostic validity was assessed by examining the agreement between the assessment by either Rater 1 or Rater 2 and the intraoperative assessment (reference standard). **RESULTS:** With regard to the assessments of both the femur and the acetabulum, there was significant agreement ( $p < 0.0001$ ) between the preoperative raters (reliability), with weighted kappa values of  $>0.75$ . There was also significant agreement ( $p < 0.0001$ ) between each rater's assessment and the intraoperative assessment (validity) of both the femur and the acetabulum, with weighted kappa values of  $>0.75$ . **CONCLUSIONS:** With use of the newly developed classification system, preoperative radiographs are reliable and valid for assessment of the severity of bone loss that will be found intraoperatively.

# Partially Structured: Seminar Announcements

Extract time, location, speaker, etc.



# Highly Structured: Zagat's Reviews

Extract restaurant, location, cost, etc.

**ZAGAT SURVEY** HOME BROWSE LISTS NEWCOMERS VOTE SHOP

location: San Diego go restaurant search: go

**ZAGAT To Go Restaurants & Nightlife** [Click Here](#)

**The 2002 Zagat Is Here** Washington, D.C./Baltimore Restaurants [Click Here](#)

**REVIEW** Food Decor Service Cost

**Bully's** S M 20 13 20 \$25

Beaches/Coastal, La Jolla/Golden Triangle, Mission Valley  
1404 Camino del Mar (bet. 13th & 15th Sts.) Del Mar, CA, 92014-2599 (858) 755-1660  
5755 La Jolla Blvd. (Bird Rock Ave.) La Jolla, CA, 92037-7302 (858) 459-2768  
2401 Camino del Rio S. (Texas St.) San Diego, CA, 92108-3701 (619) 291-2665

■ So much "fun" that it's sometimes "raucous", this "local" chainlet with "a blessed lack of pretense" specializes in some of "the best prime rib in town" insist carnivores who arrive eager for "a red-meat fix", they "can always count on a good meal", with "generous portions" at "a great value", but vegetarians note there's "little for us" at these late-night American pubs.

**VOTE**  
In order to vote, you must be a signed-in, registered user or subscriber.  
[Sign in, register or subscribe here](#)

**ZAGAT SURVEY** HOME BROWSE LISTS NEWCOMERS VOTE SHOP

location: San Diego go restaurant search: go

**ZAGAT To Go Restaurants & Nightlife** [Click Here](#)

**FREE Shipping on orders over \$20**  
Visit the **ZAGAT SURVEY shop**

**REVIEW** Food Decor Service Cost

**Aesop's Tables** S M

**Greek Cafe** S M

La Jolla/Golden Triangle  
8650 Genesee Ave. (La Jolla Village Dr.) San Diego, CA, 92122-1134 (858) 455-1535

Wafting "warm, inviting smells", this "great neighborhood" cafe in the Golden Triangle's bustling Costa Verde mall tantalizes with the "best Greek food in the area", "excelling" with "standards" such as skewers and pizzas while also "departing" from the usual with Mediterranean specialties like b'steeya; enjoy it all in a "pleasant" space that feels like "a cross between a tavern and a comfy B&B" -- "opal"

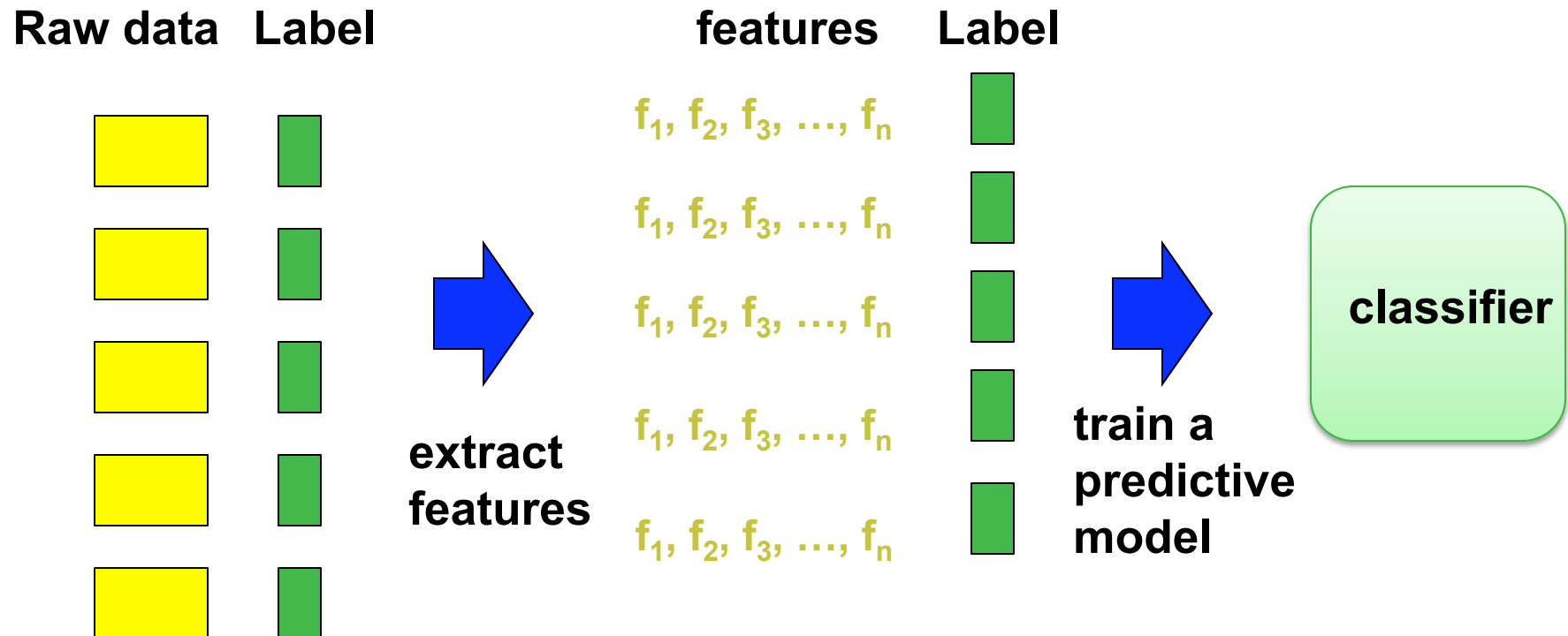
**VOTE**  
In order to vote, you must be a signed-in, registered user or subscriber.  
[Sign in, register or subscribe here](#)

# IE for Information Retrieval

- How is this useful for IR?
  - zone/field based queries
    - explicit: author search
    - implicit: recognizing addresses
  - zone weighting
  - index whole entities “Michael Jackson”
  - understand relationships between entities
    - doc: “X was acquired by Y”
    - query: “Y acquisitions”

# Classifier setup

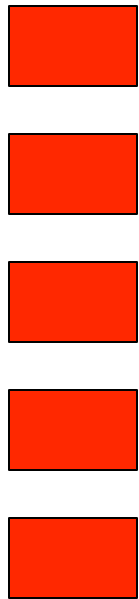
## Training or learning phase



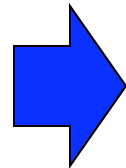
# Classifier setup

## Testing or classification phase

Raw data



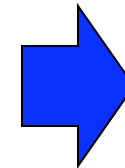
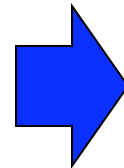
extract  
features



features

$f_1, f_2, f_3, \dots, f_n$   
 $f_1, f_2, f_3, \dots, f_n$   
 $f_1, f_2, f_3, \dots, f_n$   
 $f_1, f_2, f_3, \dots, f_n$   
 $f_1, f_2, f_3, \dots, f_n$   
 $f_1, f_2, f_3, \dots, f_n$

predict  
the label



labels



# IE Posed as a Machine Learning Task

- Training data: documents marked up with ground truth
- Local features crucial

... 00 : pm Place : Wean Hall Rm 5409 Speaker : Sebastian Thrun ...

prefix contents suffix

**What features would be useful?**

# Good Features for Information Extraction

## Creativity and Domain Knowledge Required!

begins-with-number	Example word features: <ul style="list-style-type: none"><li>– identity of word</li><li>– is in all caps</li><li>– ends in “-ski”</li><li>– is part of a noun phrase</li><li>– is in a list of city names</li><li>– is under node X in WordNet or Cyc</li><li>– is in bold font</li><li>– is in hyperlink anchor</li><li>– <i>features of past &amp; future</i></li><li>– last person name was female</li><li>– next two words are “and Associates”</li></ul>	contains-question-mark
begins-with-ordinal		contains-question-word
begins-with-punctuation		ends-with-question-mark
begins-with-question-word		first-alpha-is-capitalized
begins-with-subject		indented
blank		indented-1-to-4
contains-alphanum		indented-5-to-10
contains-bracketed-number		more-than-one-third-space
contains-http		only-punctuation
contains-non-space		prev-is-blank
contains-number		prev-begins-with-ordinal
contains-pipe		shorter-than-30



# Good Features for Information Extraction

## Creativity and Domain Knowledge Required!

Is Capitalized	Character n-gram classifier says string is a person name (80% accurate)	Word Features
Is Mixed Caps		– lists of job titles,
Is All Caps	In stopword list (the, of, their, etc)	– Lists of prefixes
Initial Cap		– Lists of suffixes
Contains Digit	In honorific list (Mr, Mrs, Dr, Sen, etc)	– 350 informative phrases
All lowercase	In person suffix list (Jr, Sr, PhD, etc)	HTML/Formatting Features
Is Initial	In name particle list (de, la, van, der, etc)	– {begin, end, in} x
Punctuation	In Census lastname list; segmented by P(name)	{<b>, <i>, <a>, <hN>} x
Period	In Census firstname list; segmented by P(name)	{lengths 1, 2, 3, 4, or longer}
Comma	In locations lists (states, cities, countries)	– {begin, end} of line
Apostrophe	In company name list ("J. C. Penny")	
Dash	In list of company suffixes (Inc, & Associates, Foundation)	
Preceded by HTML tag		

**How can we pose this as a  
classification (or learning)  
problem?**

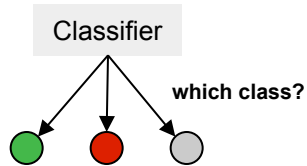


**classifier**

# Landscape of ML Techniques for IE:

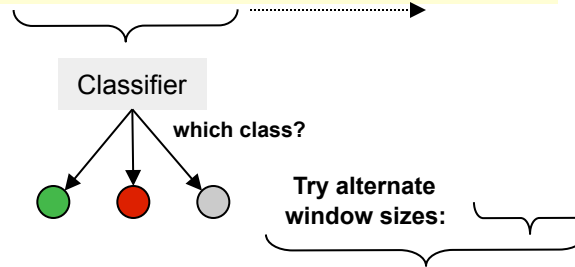
## Classify Candidates

Abraham Lincoln was born in Kentucky.

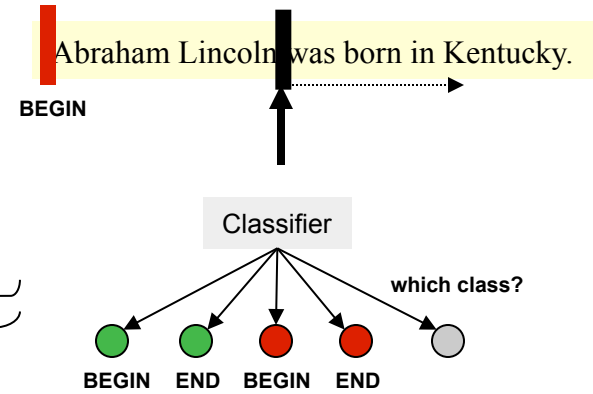


## Sliding Window

Abraham Lincoln was born in Kentucky.

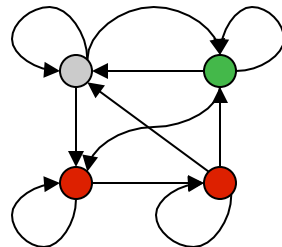


## Boundary Models



## Finite State Machines

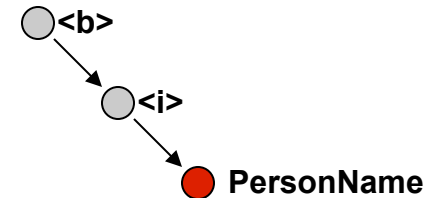
Abraham Lincoln was born in Kentucky.



## Wrapper Induction

**<b><i>Abraham Lincoln</i></b>** was born in Kentucky.

Learn and apply pattern for a website



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

# Sliding Windows & Boundary Detection

# Information Extraction by Sliding Windows

**E.g.  
Looking for  
seminar  
location**

GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell  
School of Computer Science  
Carnegie Mellon University

3:30 pm  
7500 Wean Hall

Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g. analogy, explanation-based learning), learning theory (e.g. PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.

**CMU UseNet Seminar Announcement**

# Information Extraction by Sliding Windows

**E.g.  
Looking for  
seminar  
location**

GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell  
School of Computer Science  
Carnegie Mellon University

3:30 pm  
7500 Wean Hall

Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g. analogy, explanation-based learning), learning theory (e.g. PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.

**CMU UseNet Seminar Announcement**

# Information Extraction by Sliding Window

**E.g.  
Looking for  
seminar  
location**

GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell  
School of Computer Science  
Carnegie Mellon University

3:30 pm  
7500 Wean Hall

Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g. analogy, explanation-based learning), learning theory (e.g. PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.

**CMU UseNet Seminar Announcement**

# Information Extraction by Sliding Window

**E.g.  
Looking for  
seminar  
location**

GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell  
School of Computer Science  
Carnegie Mellon University

3:30 pm  
7500 Wean Hall

Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g. analogy, explanation-based learning), learning theory (e.g. PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.

**CMU UseNet Seminar Announcement**



# Information Extraction by Sliding Window

**E.g.  
Looking for  
seminar  
location**

GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell  
School of Computer Science  
Carnegie Mellon University

3:30 pm

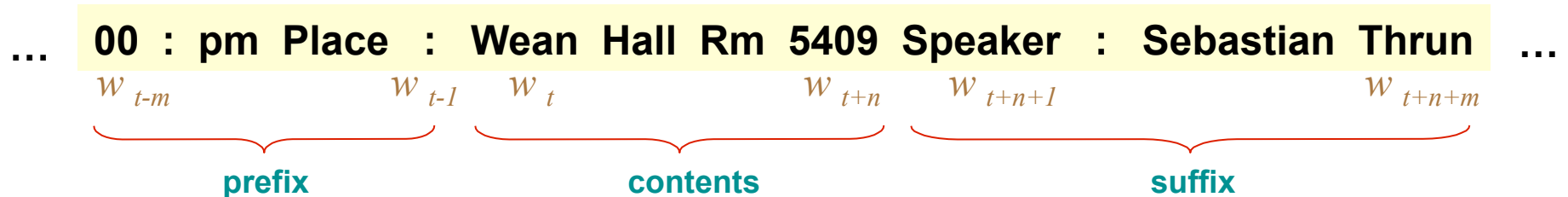
7500 Wean Hall

Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g. analogy, explanation-based learning), learning theory (e.g. PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.

**CMU UseNet Seminar Announcement**

# Information Extraction with Sliding Windows

[Freitag 97, 98; Soderland 97; Califf 98]



- Standard supervised learning setting
  - Positive instances: Windows with real label
  - Negative instances: All other windows
  - Features based on candidate, prefix and suffix

# IE by Boundary Detection

**E.g.  
Looking for  
seminar  
location**

GRAND CHALLENGES FOR MACHINE LEARNING



Jaime Carbonell  
School of Computer Science  
Carnegie Mellon University

3:30 pm  
7500 Wean Hall

Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g. analogy, explanation-based learning), learning theory (e.g. PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.

**CMU UseNet Seminar Announcement**

# IE by Boundary Detection

**E.g.  
Looking for  
seminar  
location**

GRAND CHALLENGES FOR MACHINE LEARNING



Jaime Carbonell  
School of Computer Science  
Carnegie Mellon University

3:30 pm  
7500 Wean Hall

Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g. analogy, explanation-based learning), learning theory (e.g. PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.

**CMU UseNet Seminar Announcement**

# IE by Boundary Detection

GRAND CHALLENGES FOR MACHINE LEARNING



Jaime Carbonell  
School of Computer Science  
Carnegie Mellon University

3:30 pm  
7500 Wean Hall

Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g. analogy, explanation-based learning), learning theory (e.g. PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.

**E.g.  
Looking for  
seminar  
location**

**CMU UseNet Seminar Announcement**

# IE by Boundary Detection

**E.g.  
Looking for  
seminar  
location**

GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell  
School of Computer Science  
Carnegie Mellon University

3:30 pm

7500 Wean Hall



Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g. analogy, explanation-based learning), learning theory (e.g. PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.

**CMU UseNet Seminar Announcement**

# IE by Boundary Detection

**E.g.  
Looking for  
seminar  
location**

GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell  
School of Computer Science  
Carnegie Mellon University

3:30 pm  
7500 Wean Hall



Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g. analogy, explanation-based learning), learning theory (e.g. PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.

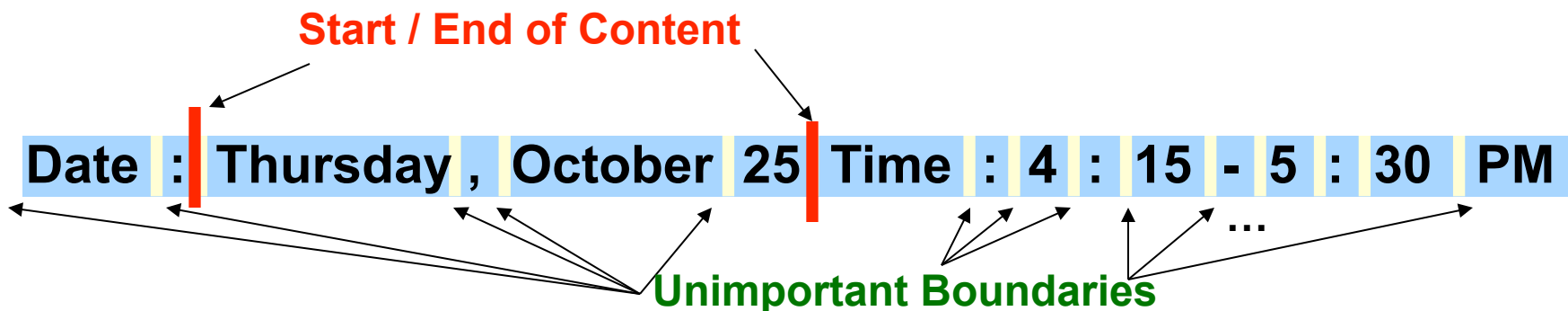
**CMU UseNet Seminar Announcement**

# IE by Boundary Detection

Input: Linear Sequence of Tokens

Date : Thursday , October 25 Time : 4 : 15 - 5 : 30 PM

Method: Identify start and end Token Boundaries



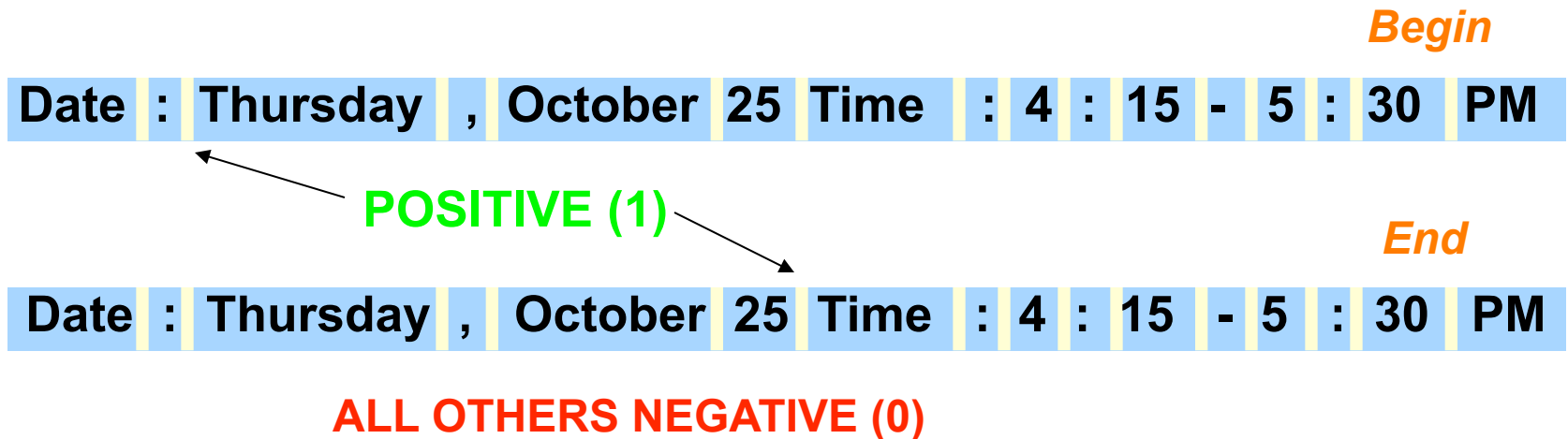
Output: Tokens Between Identified Start / End Boundaries

Date : Thursday , October 25 Time : 4 : 15 - 5 : 30 PM



# Learning: IE as Classification

- The set of training examples is all of the boundaries in a document




- The goal is to approximate two extraction functions *Begin* and *End*:

$$\begin{aligned} \textit{Begin}(i) = & \quad 1 \text{ if } i \text{ begins a field} \\ & \quad 0 \text{ otherwise} \end{aligned}$$

# Boundary Detectors

- “Boundary Detectors” are pairs of token sequences  $\langle p, s \rangle$ 
  - A detector matches a boundary iff  $p$  matches text before boundary and  $s$  matches text after boundary
  - Detectors can contain wildcards, e.g. “capitalized word”, “number”, etc.
- Example:
  - $\langle \text{Date:}, [\textit{CapitalizedWord}] \rangle$  matches:

Date: Thursday, October 25



# Another Detector Example

*Begin* boundary detector:

*End* boundary detector:


Prefix	Suffix
<code>&lt; a href = "</code>	<code>http</code>
<code>empty</code>	<code>"&gt;</code>

Input Text:

```
text<b><a href=http://www.cs.ucsd.edu>
```

Matches:

```
text<b><a href="http://www.cs.ucsd.edu">
```



# BWI: Learning to detect boundaries

*[Freitag & Kushmerick, AAAI 2000]*

- Another formulation: learn **three** probabilistic classifiers:
  - $Begin(i) = \text{Prob}(\text{ position } i \text{ starts a field})$
  - $End(j) = \text{Prob}(\text{ position } j \text{ ends a field})$
  - $Len(k) = \text{Prob}(\text{ an extracted field has length } k)$
- Then score a possible extraction  $(i,j)$  by  $Begin(i) * End(j) * Len(j-i)$
- $Len(k)$  is estimated from a histogram
- $Begin(i)$  and  $End(j)$  learned by boosting over simple boundary patterns and features

# Problems with Sliding Windows and Boundary Finders

- Decisions in neighboring parts of the input are made independently from each other.
  - Sliding Window may predict a “seminar end time” before the “seminar start time”.
  - It is possible for two *overlapping* windows to both be above threshold.
  - In a Boundary-Finding system, left boundaries are laid down independently from right boundaries, and their pairing happens as a separate step.

# Modeling the sequential nature of data: citation parsing

- Fahlman, Scott & Lebiere, Christian (1989). The cascade-correlation learning architecture. *Advances in Neural Information Processing Systems*, pp. 524-532.
- Fahlman, S.E. and Lebiere, C., “The Cascade Correlation Learning Architecture,” *Neural Information Processing Systems*, pp. 524-532, 1990.
- Fahlman, S. E. (1991) The recurrent cascade-correlation learning architecture. *NIPS 3*, 190-205.

**What patterns do you see here?**

**Ideas?**

## Some sequential patterns

- Something interesting in the sequence of fields that we'd like to capture
  - Authors come first
  - Title comes before journal
  - Page numbers come near the end
  - All types of things generally contain multiple words

# Predict a sequence of tags

author

author

year

title

title

title

**Fahlman, S. E. (1991) The recurrent cascade**

title

title

title

journal

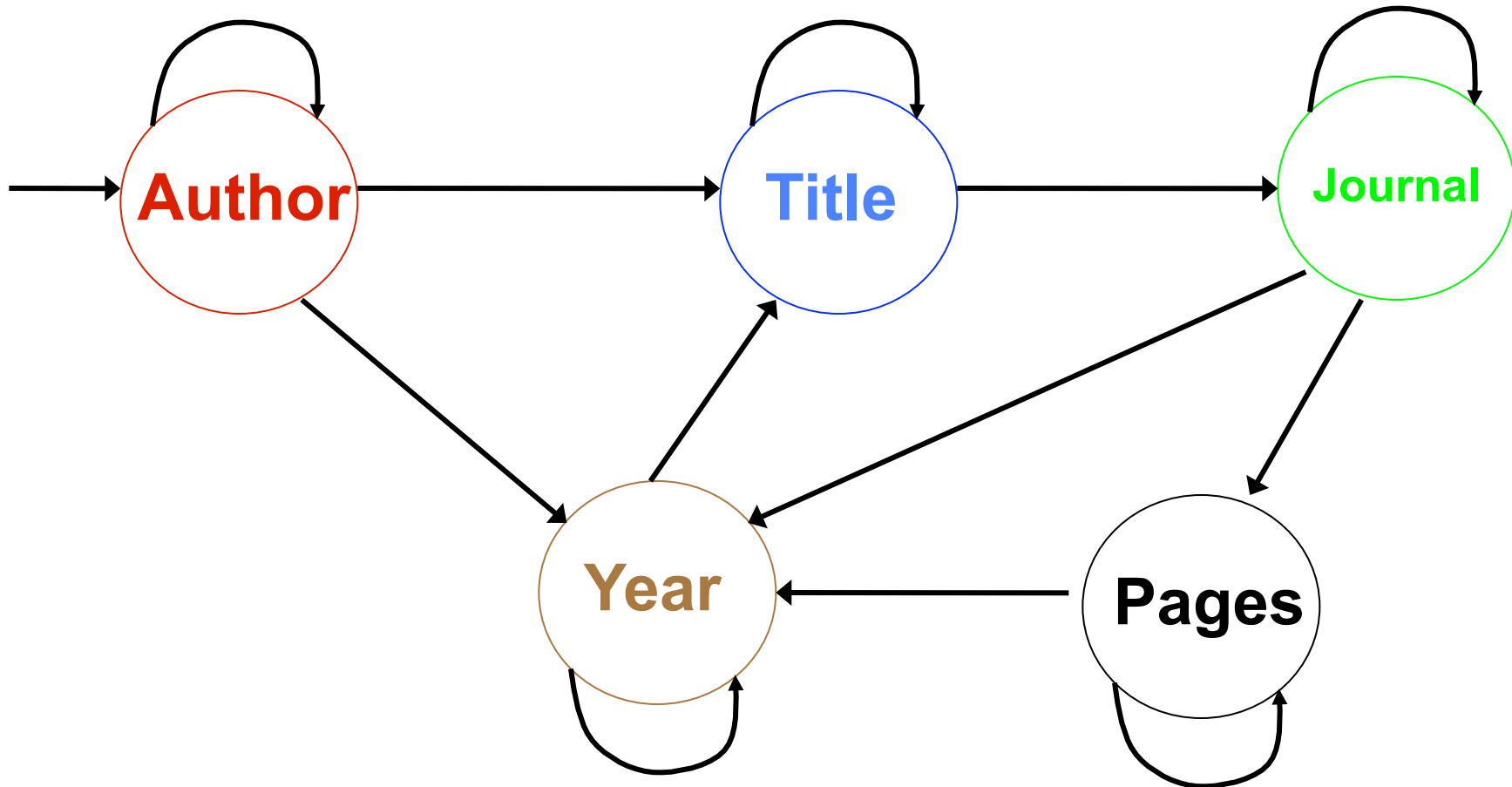
pages

**correlation learning architecture. NIPS 3, 190-205.**

**Ideas?**

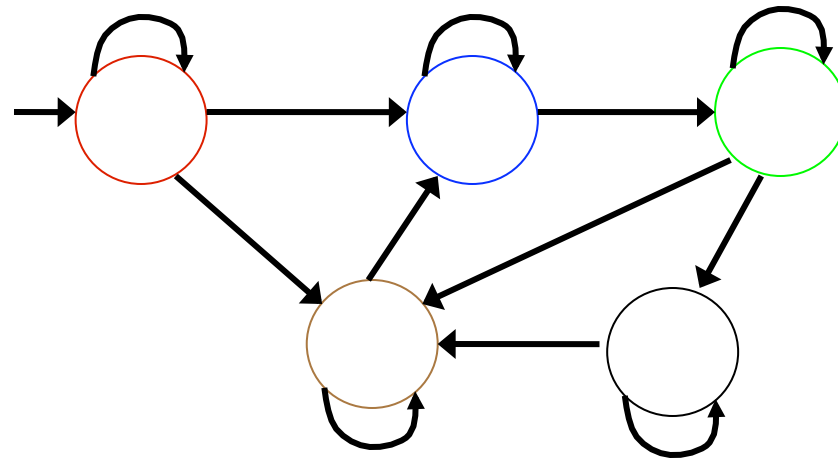


# Hidden Markov Models(HMMs)



## HMM: Generative Model (3)

- States:  $x_i$
- State transitions:  $P(x_i|x_j) = a[x_i|x_j]$
- Output probabilities:  $P(o_i|x_j) = b[o_i|x_j]$



- Markov independence assumption

# HMMs: Performing Extraction

- Given output words:
  - fahlman s e 1991 the recurrent cascade correlation learning architecture nips 3 190 205
- Find state sequence that maximizes:

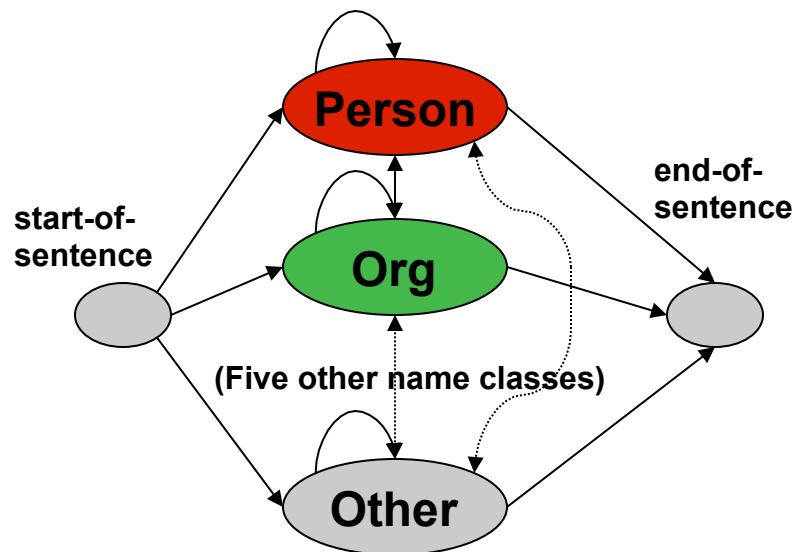
$$\prod_i a[x_i | x_{i-1}] b[o_i | x_i]$$

- Lots of possible state sequences to test ( $5^{14}$ )

# HMM Example: Nymble

[Bikel, et al 97]

Task: Named Entity Extraction



- Bigram within classes
- Backoff to unigram
- Special capitalization and number features...

Train on 450k words of news wire text.

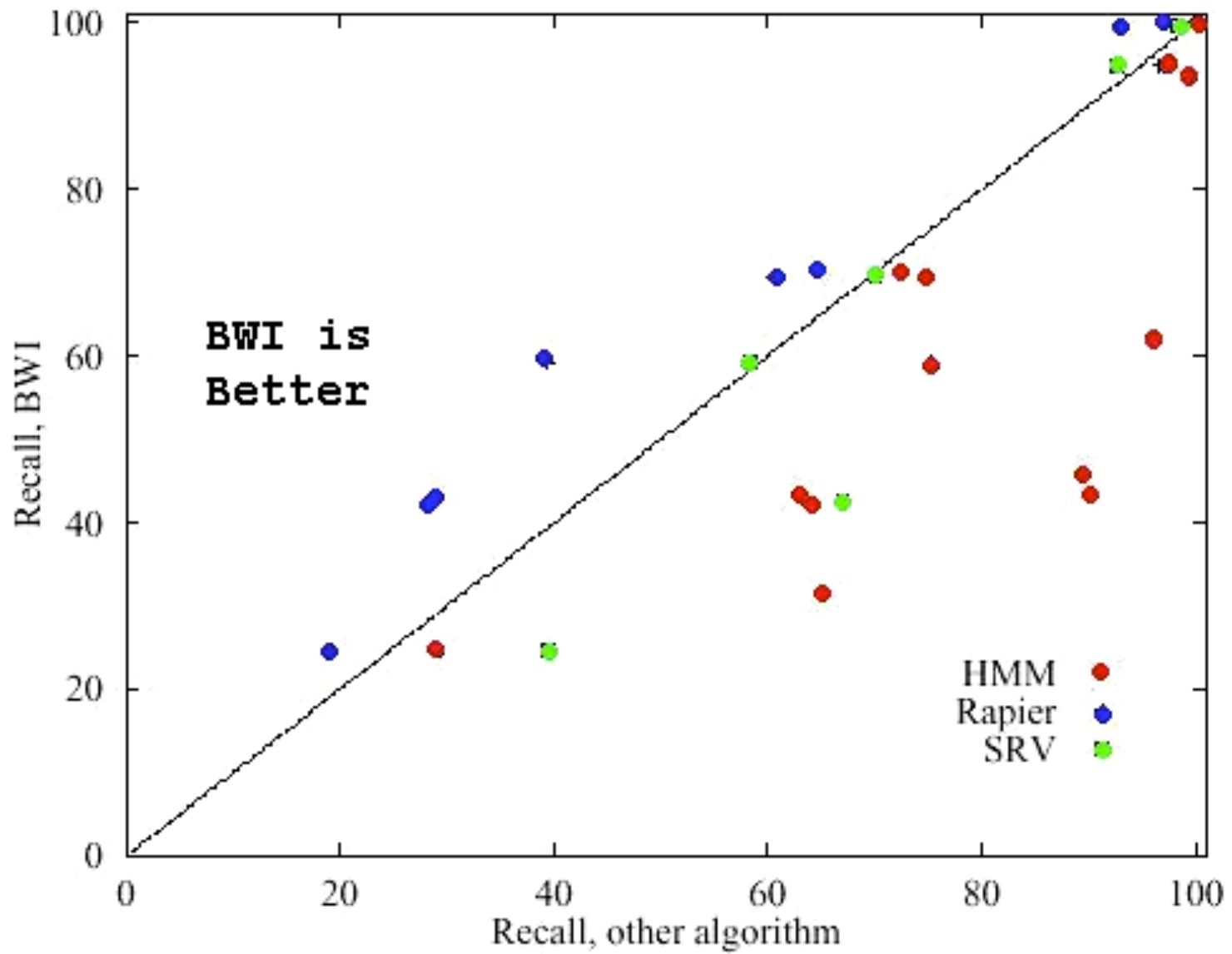
Results:

<b>Case</b>	<b>Language</b>	<b>F1 .</b>
Mixed	English	93%
Upper	English	91%
Mixed	Spanish	90%

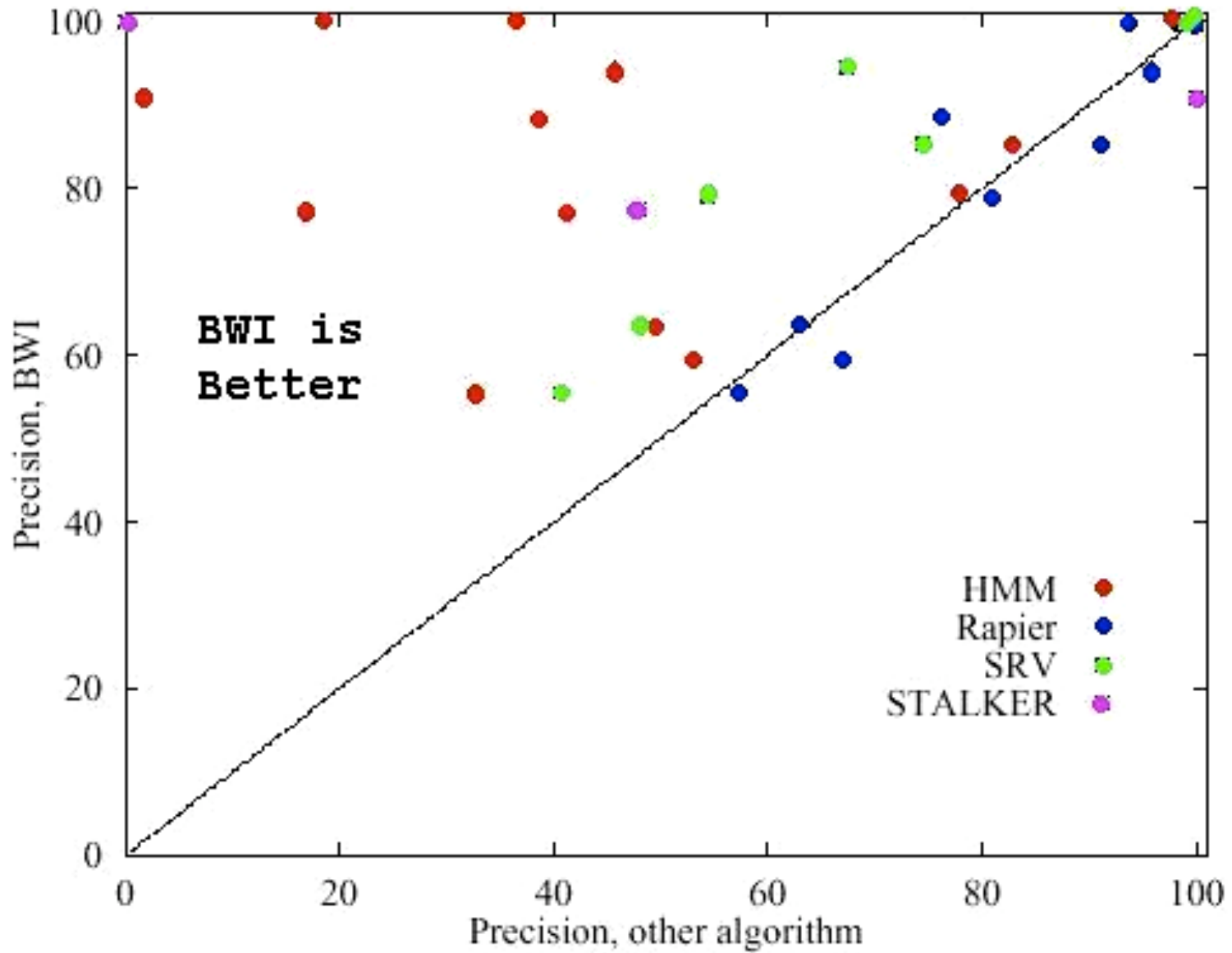
# System Comparison

- 16 different extraction tasks
- Algorithms
  - BWI vs.
  - Two rule learners: SRV and Rapier
  - One algorithm based on hidden Markov models
  - One wrapper induction algorithm: Stalker

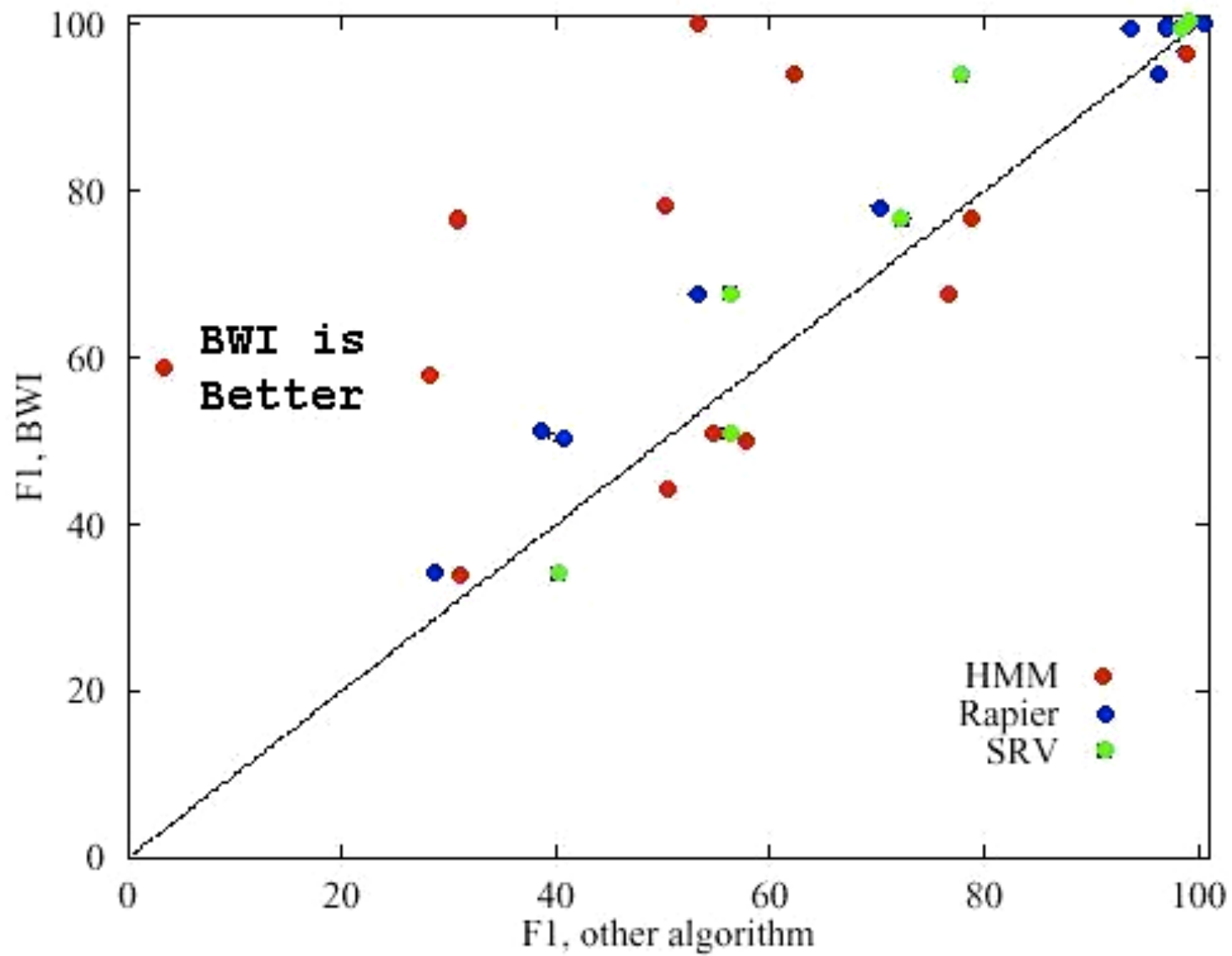
# Recall



# Precision

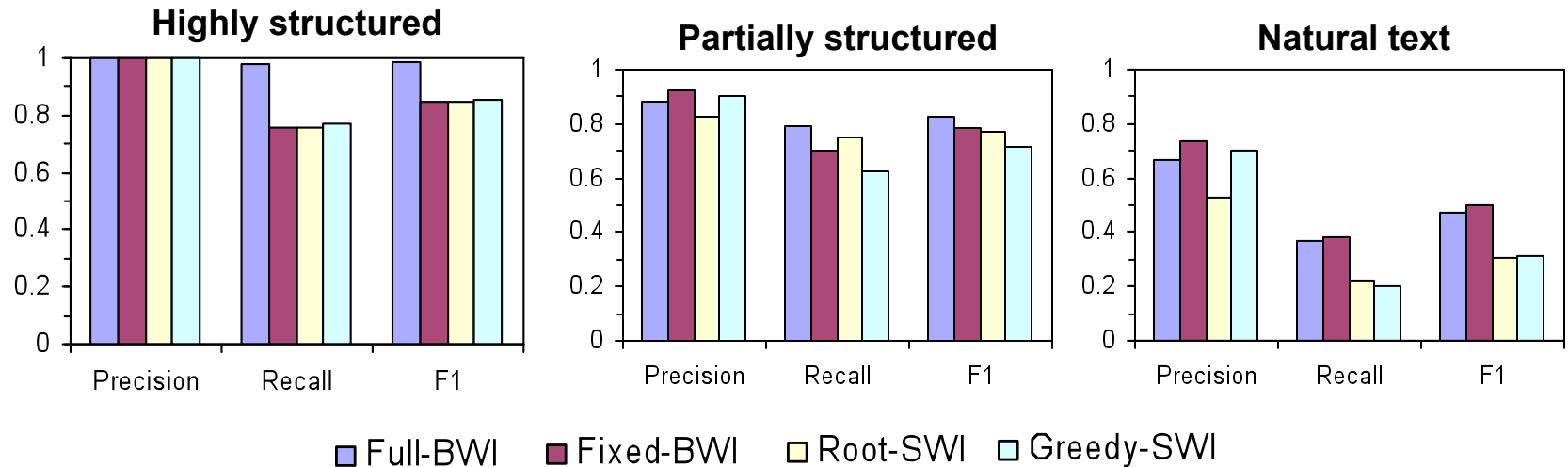


# F1





# Data regularity is important!



- As the regularity decreases, so does the performance
- Algorithms interact differently at with different levels of regularity

# How important are features?

- One of the challenges for IE methods is generalizability
- Wildcards can help with this

wildcards	speaker	location	stime	etime
none	15.1	69.2	95.7	83.4
just <*>	49.4	73.5	99.3	<u>95.0</u>
default	67.7	<u>76.7</u>	<u>99.4</u>	94.6
lexical	<u>73.5</u>	-	-	-

default: a set of eight wildcards

lexical: task specific lexical resources:

- <FName>: common first names released by U.S. Census Bureau.
- <LName>: common last names
- <NEW>: tokens not found in /usr/dict/words on Unix

# Improving task regularity

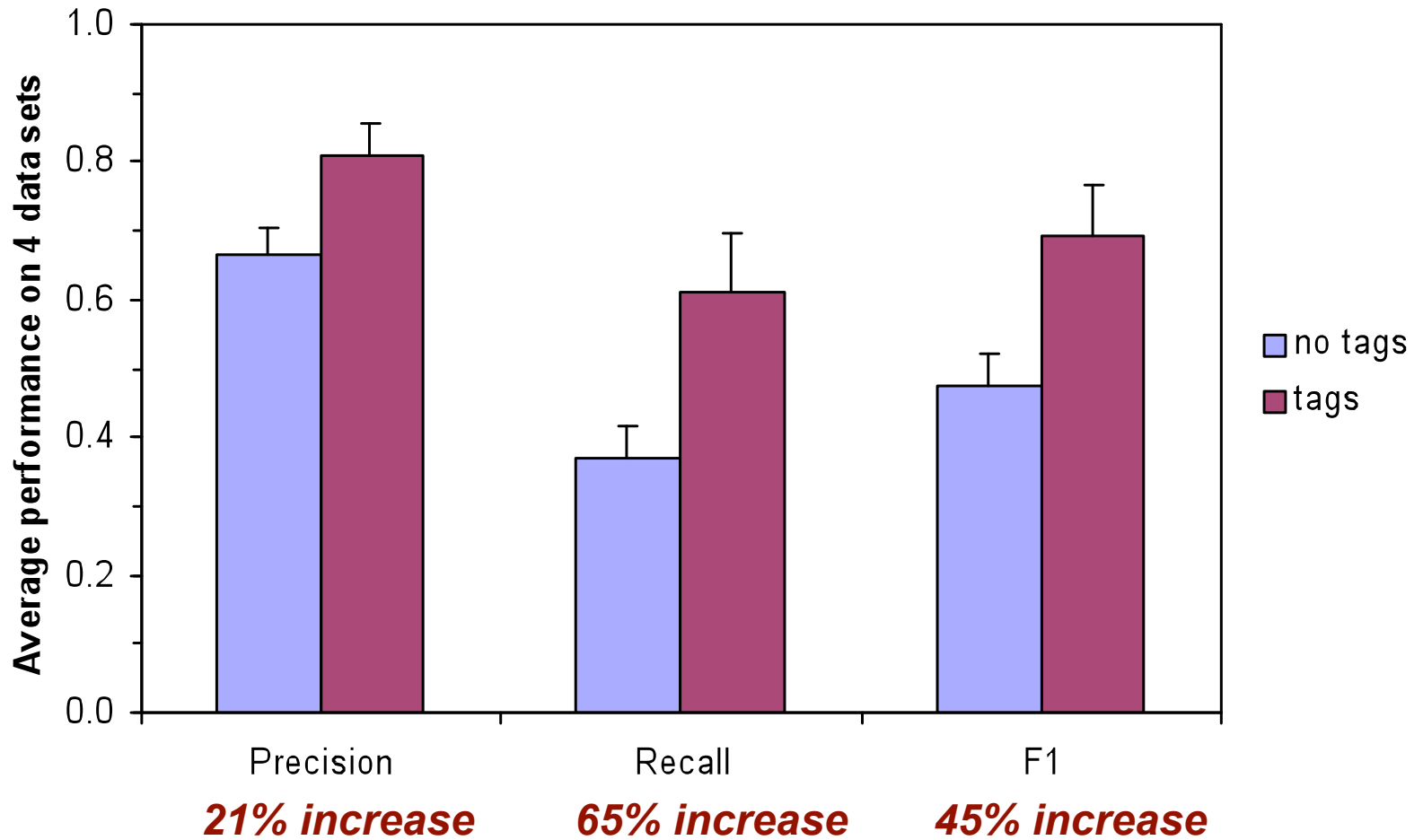
- Instead of altering methods, alter text
- Idea: Add limited grammatical information
  - Run shallow parser over text
  - Flatten parse tree and insert as tags

## Example of Tagged Sentence:

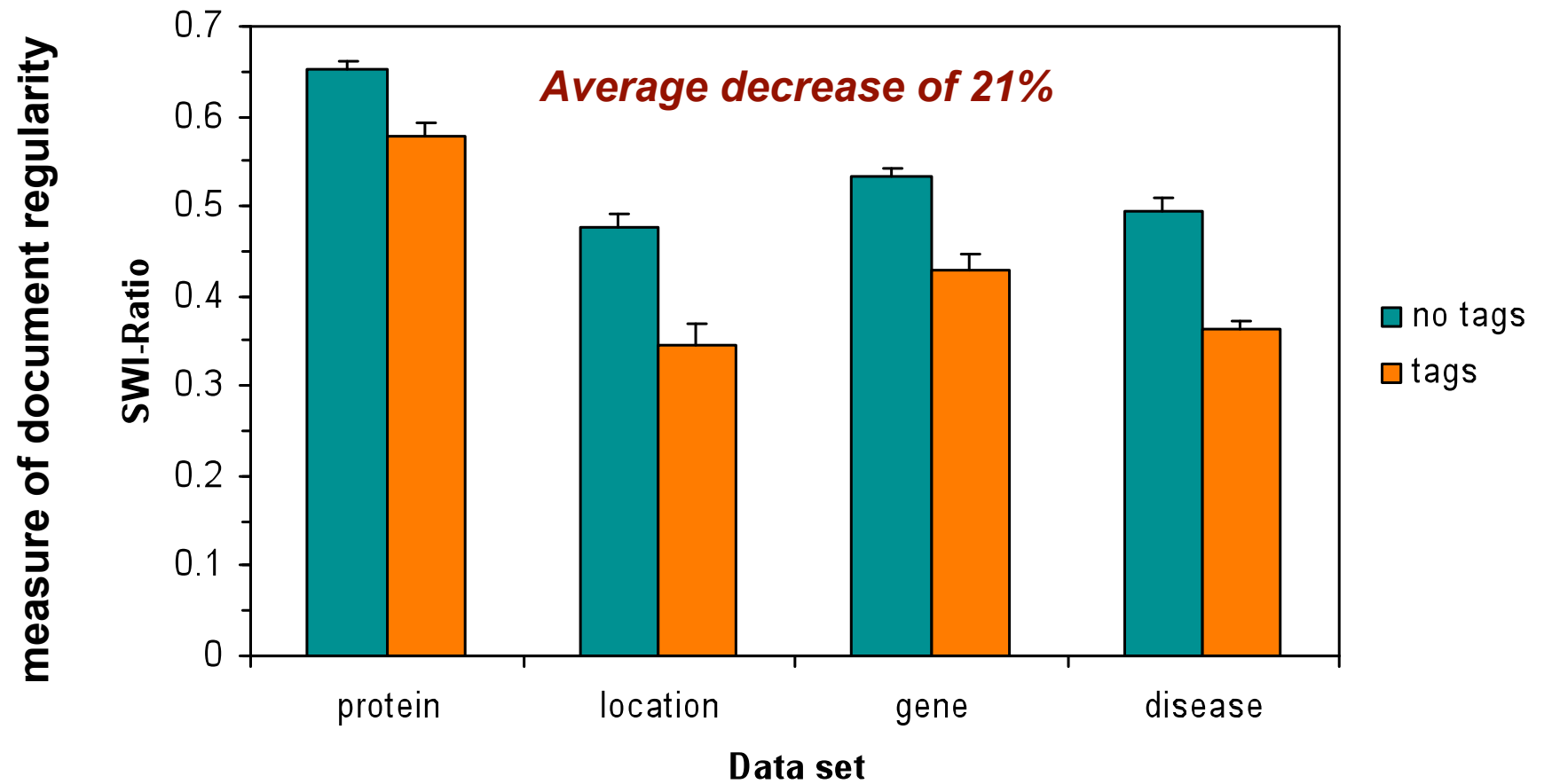
**Uba2p is located largely in the nucleus.**

**NP\_SEG**      **VP\_SEG**      **PP\_SEG**      **NP\_SEG**

# Tagging Results on Natural Domain



# Tagging increases regularity



# Collaborative Searching

- What are other gains that can be achieved through collaborative searching?
- What are cons to collaborative searching?
- Who do you think will be the primary users of collaborative searching sites?