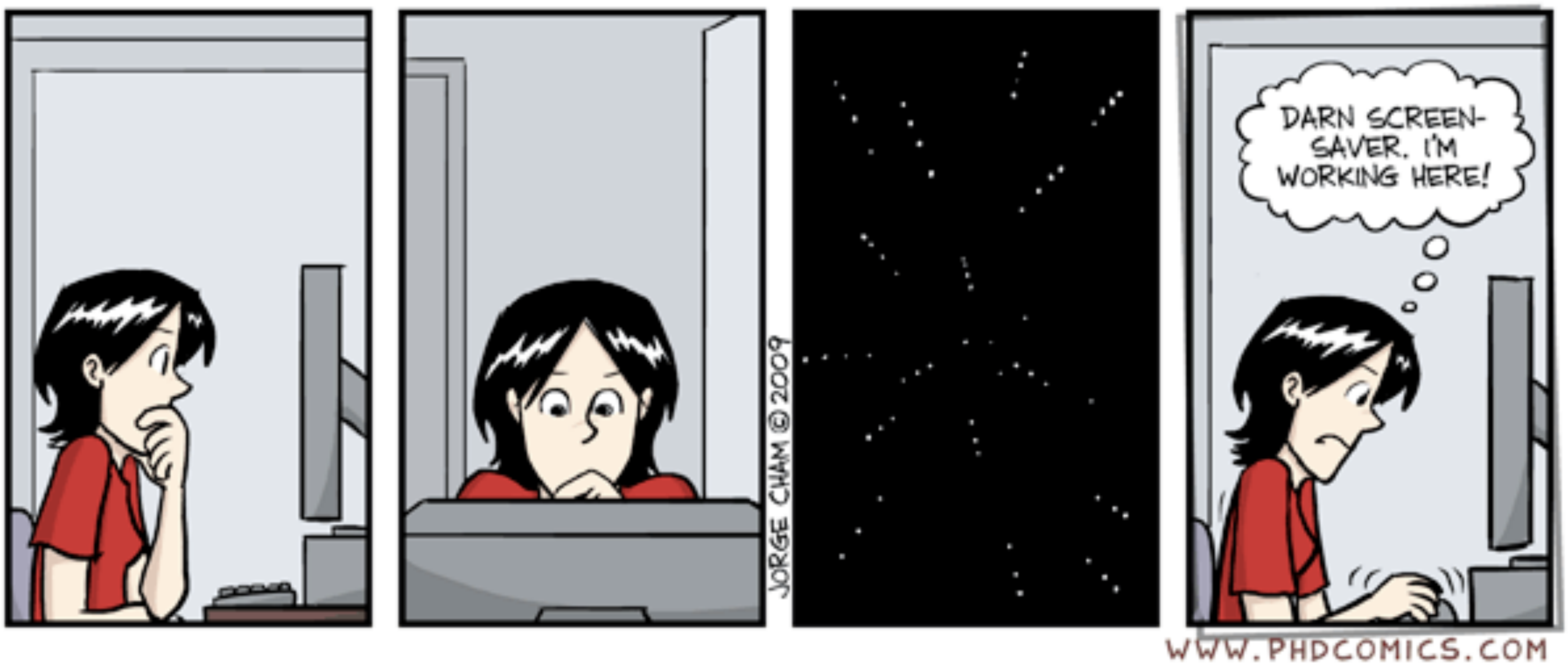


http://www.phdcomics.com/comics.php

Information Extraction

David Kauchak
cs160
Fall 2009

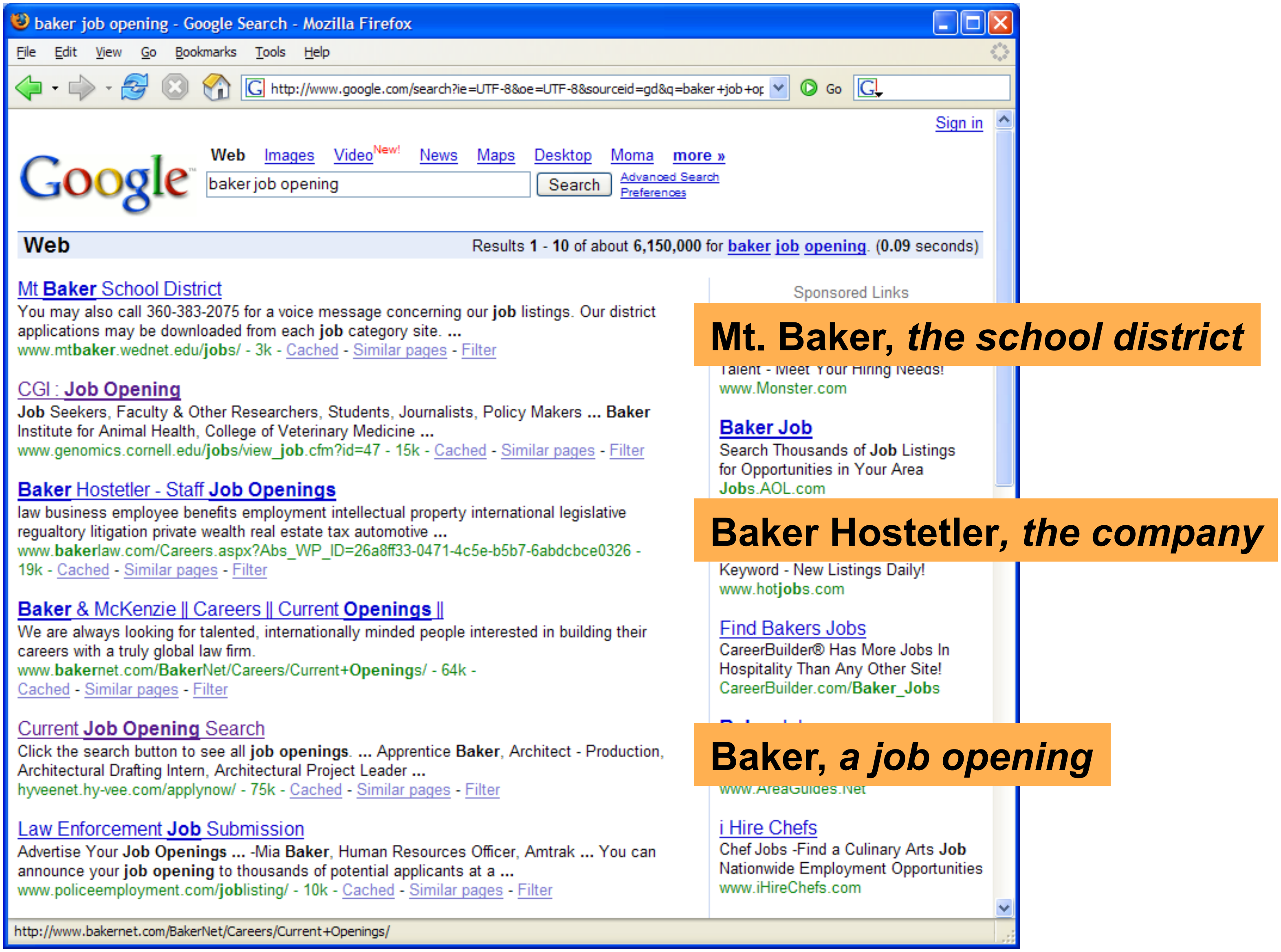
some content adapted from:

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
Sign in
Web Images Video New! 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/ - 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 - 15k - Cached - Similar pages - Filter
Baker Hostetler - Staff Job Openings
law business employee benefits employment intellectual property international legislative regualtory litigation private wealth real estate tax automotive ...
www.bakerlaw.com/Careers.aspx?Abs_WP_ID=26a8ff33-0471-4c5e-b5b7-6abdcbce0326 - 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/ - 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/ - 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/ - 10k - Cached - Similar pages - Filter
Sponsored Links
Talent - Meet Your Hiring Needs!
www.Monster.com
Baker Job
Search Thousands of Job Listings for Opportunities in Your Area
Jobs.AOL.com
Keyword - New Listings Daily!
www.hotjobs.com
Find Bakers Jobs
CareerBuilder® Has More Jobs In Hospitality Than Any Other Site!
CareerBuilder.com/Baker_Jobs
www.AreaGuides.Net
i Hire Chefs
Chef Jobs -Find a Culinary Arts Job Nationwide Employment Opportunities
www.iHireChefs.com
http://www.bakernet.com/BakerNet/Careers/Current+Openings/
Mt. Baker, *the school district*
Baker Hostetler, *the company*
Baker, *a job opening*

A solution

Why is this better? **How does it happen?**

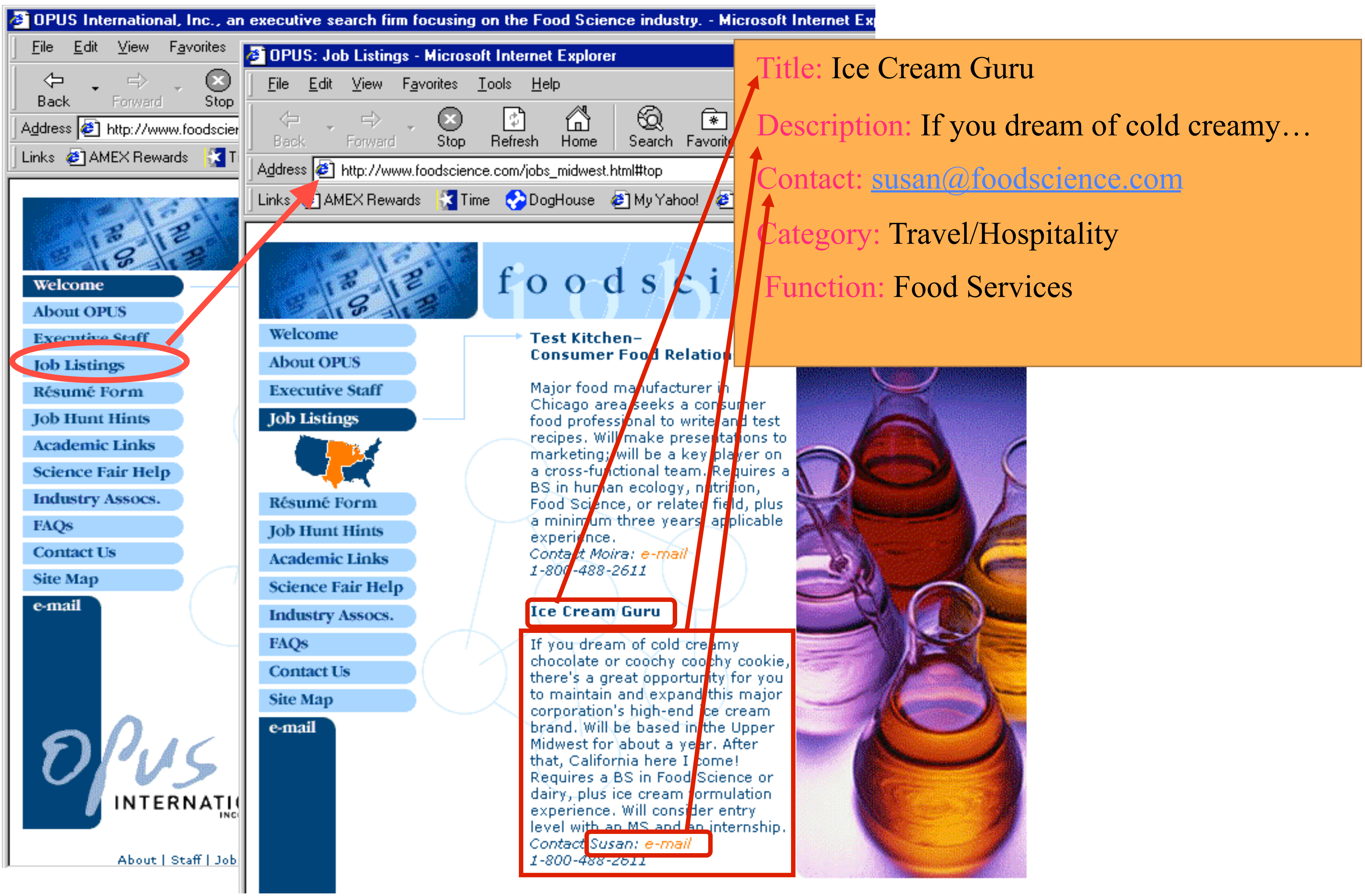
Job Openings:

Category = *Food Services*

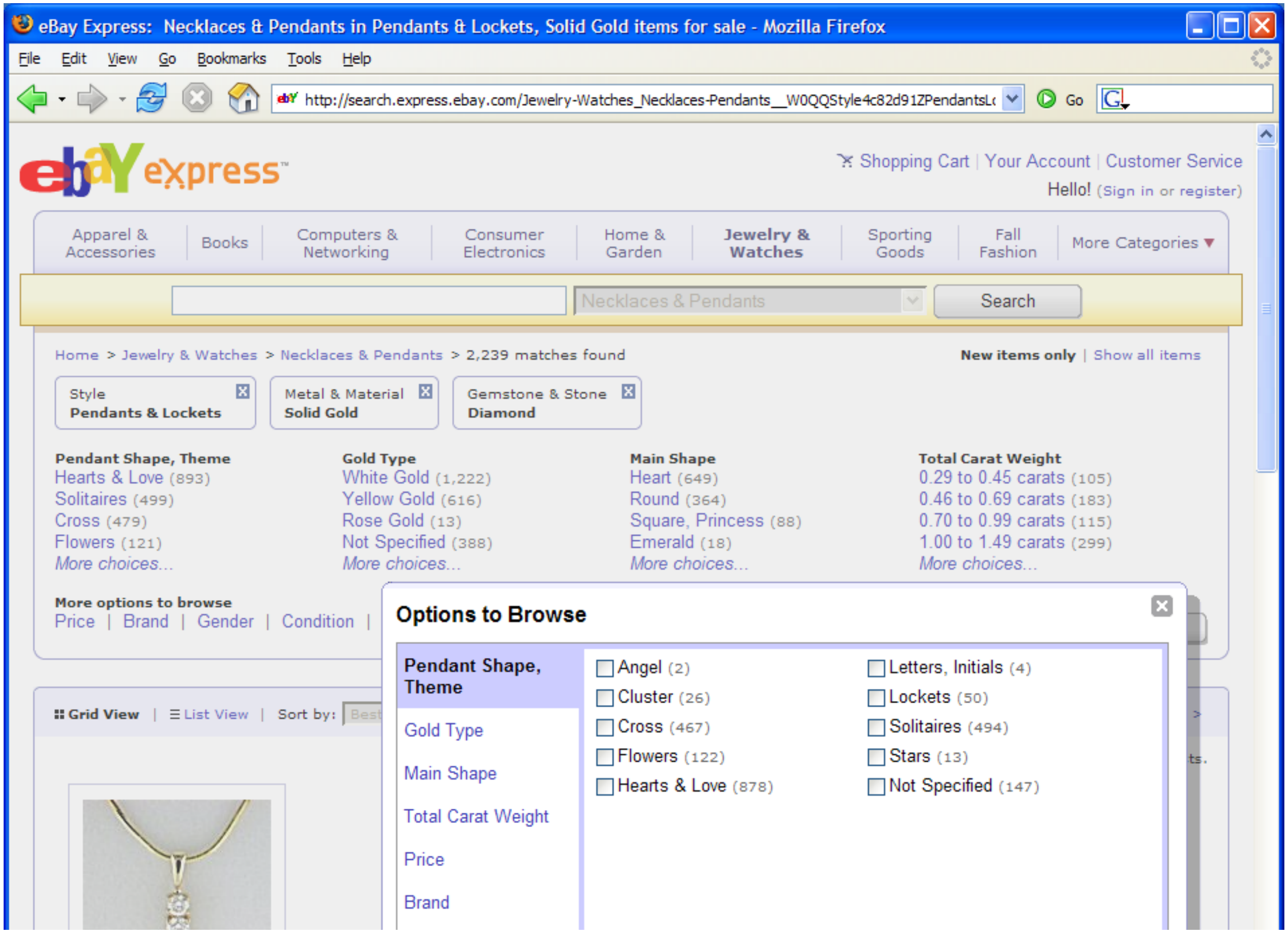
Keyword = *Baker*

Location = *Continental U.S.*

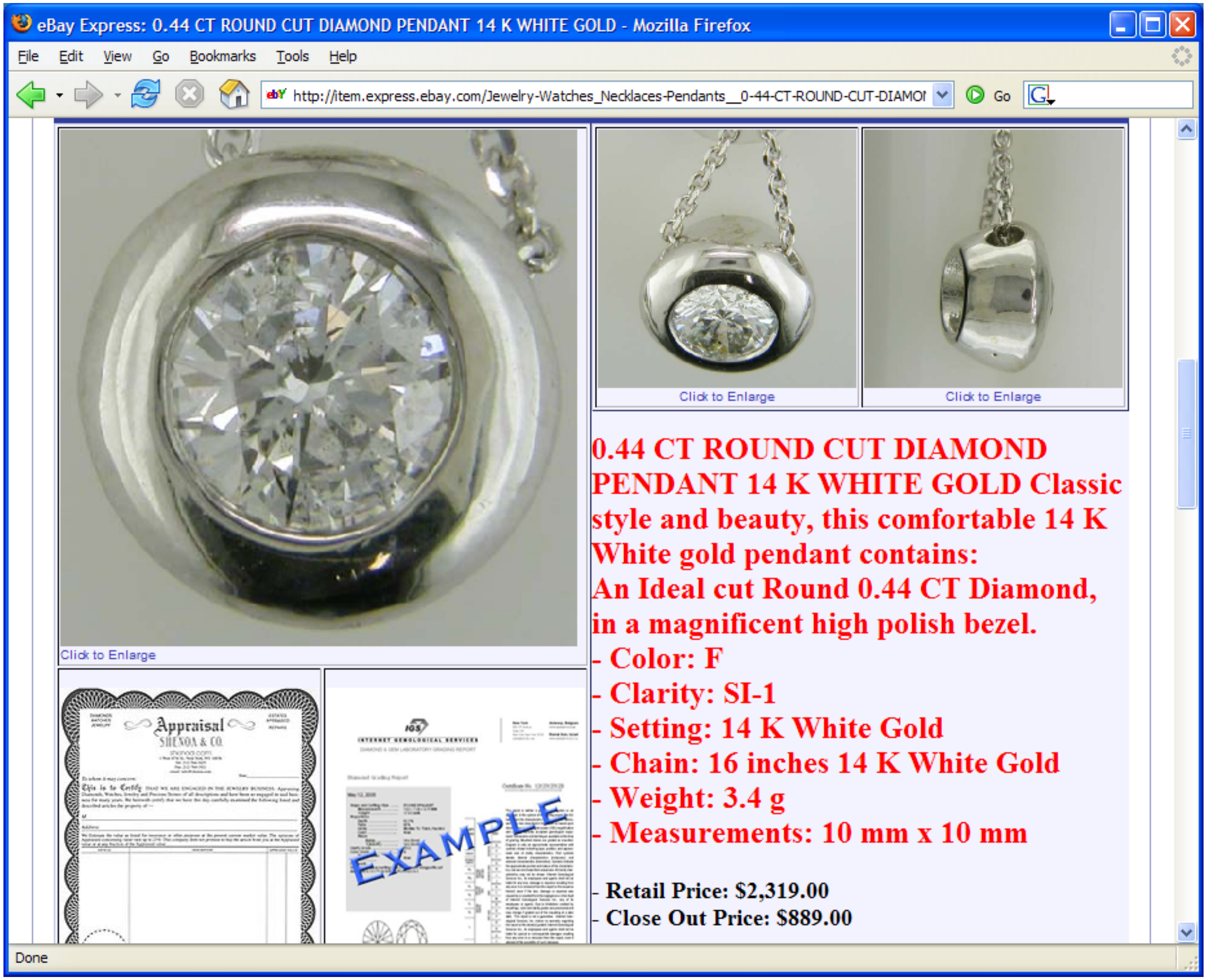
Extracting Job Openings from the Web



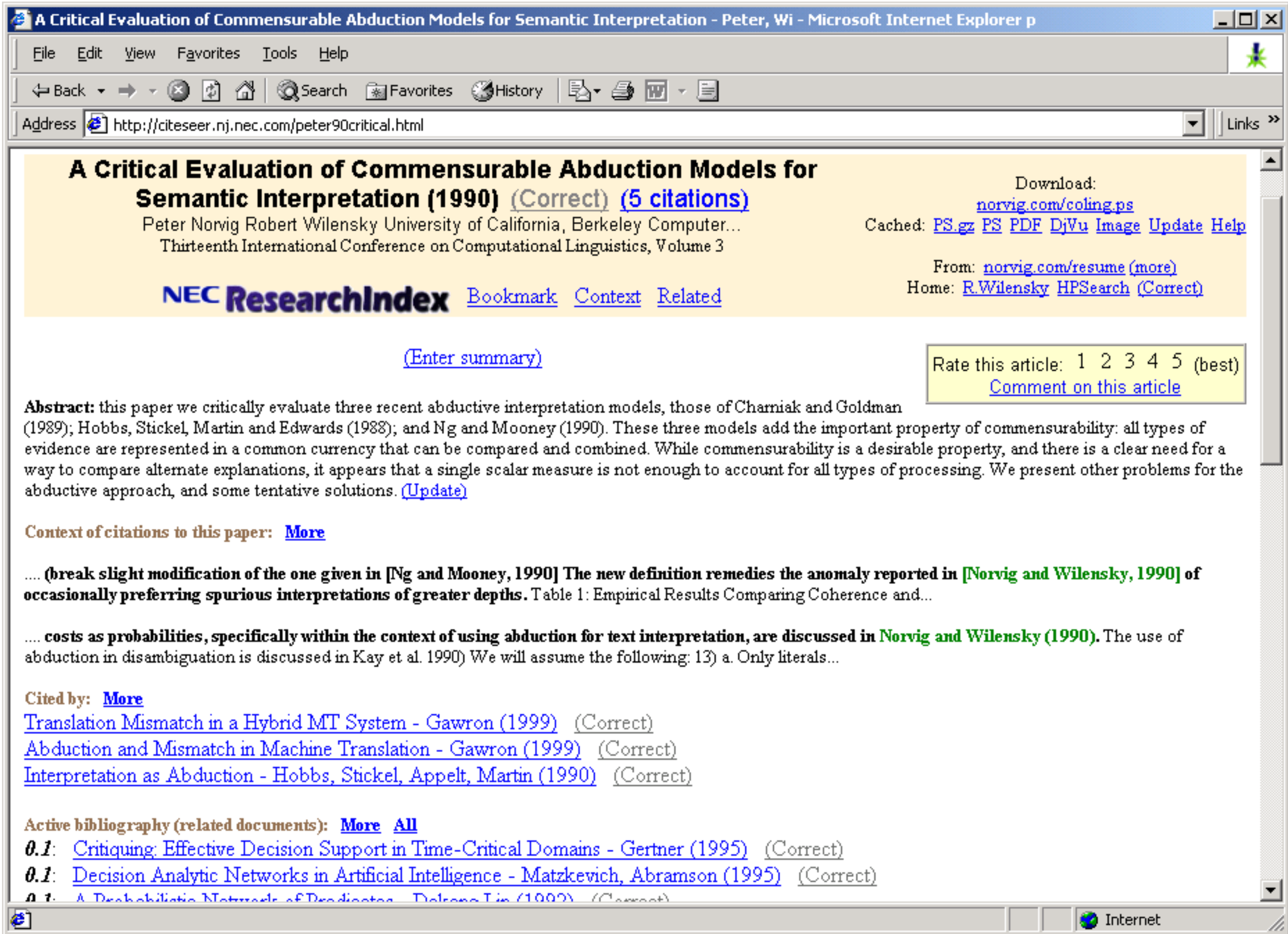
Potential Enabler of Faceted Search



Often structured information in text



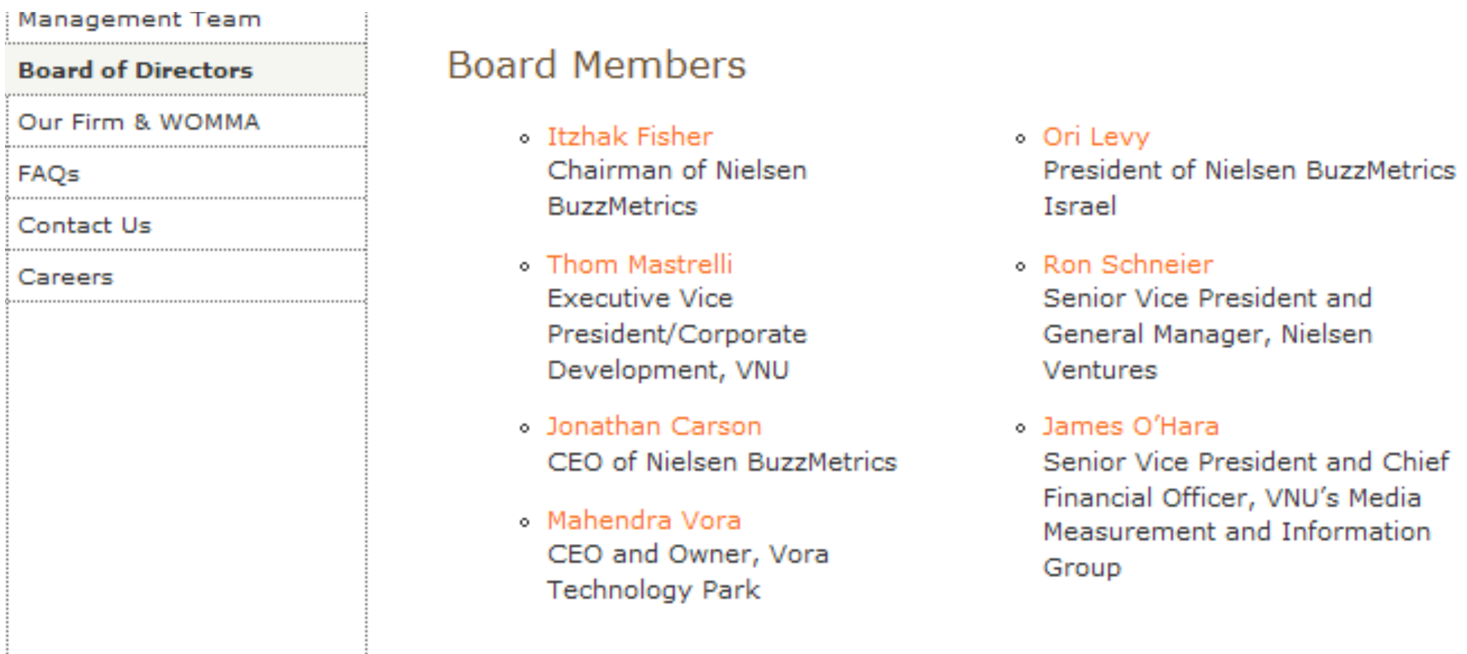
Research papers



What is Information Extraction?

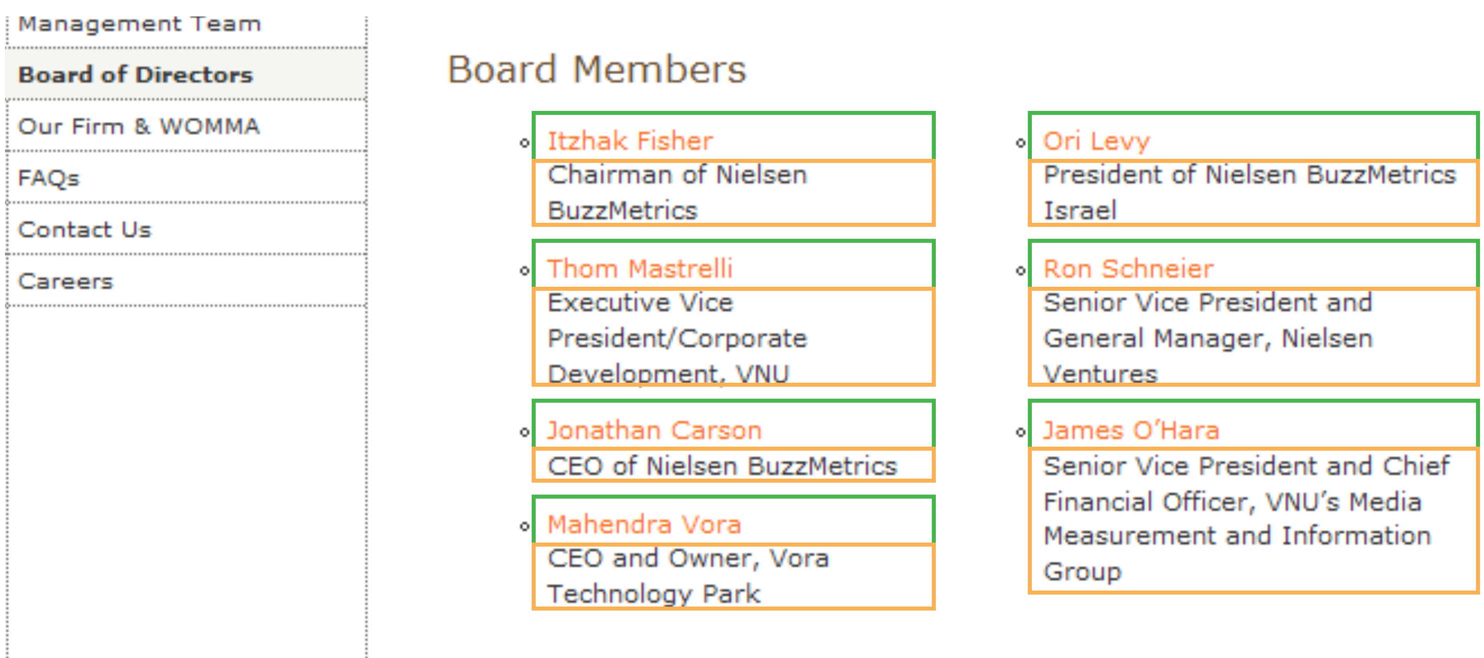
Traditional definition:

- Recovering structured data from formatted text



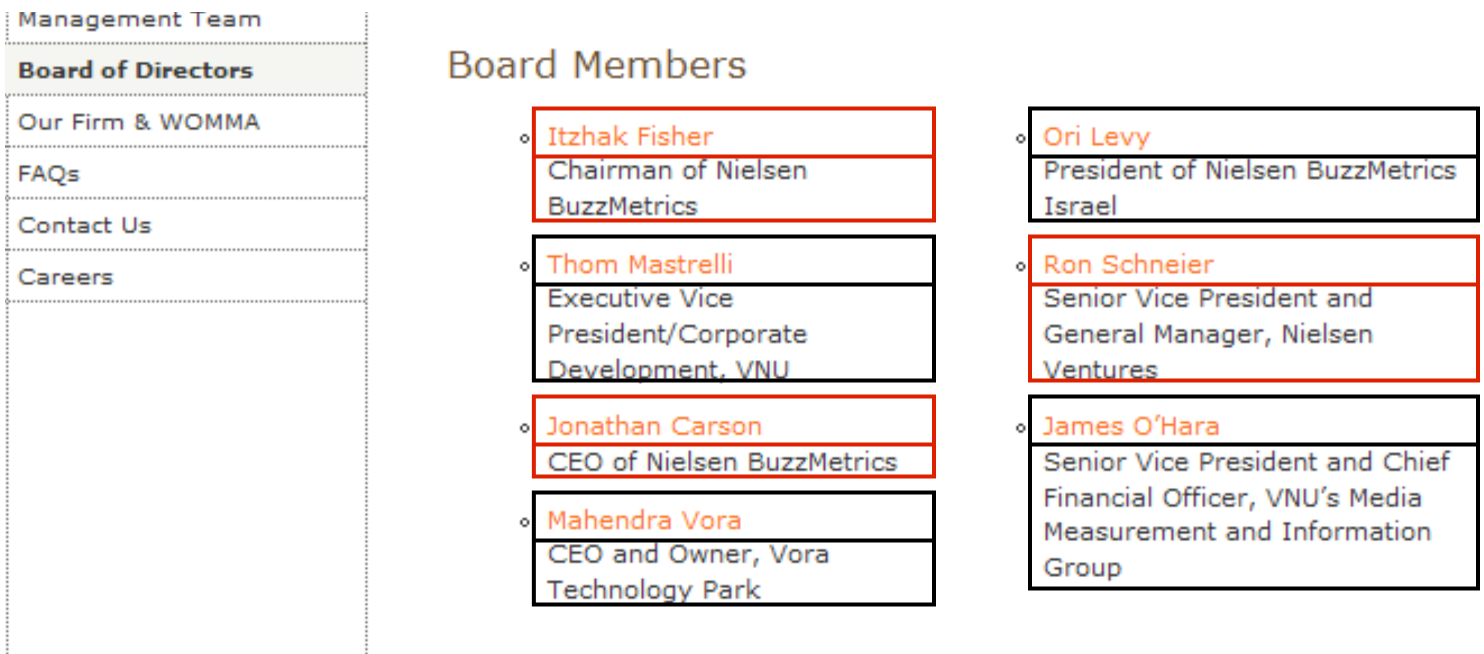
What is Information Extraction?

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



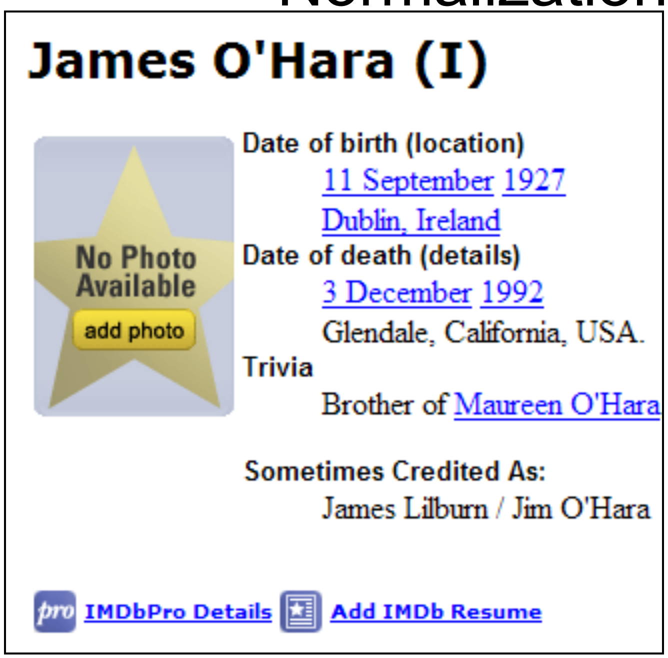
What is Information Extraction?

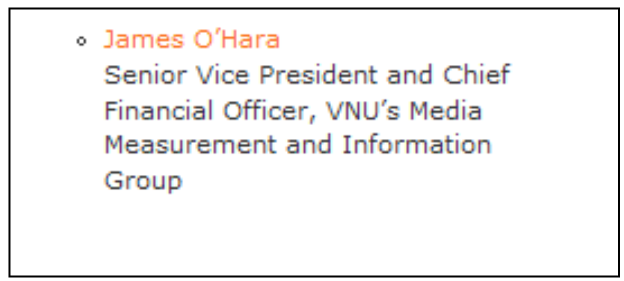
- Recovering structured data from formatted text
  - Identifying fields (e.g. named entity recognition)
  - Understanding relations between fields (e.g. record association)

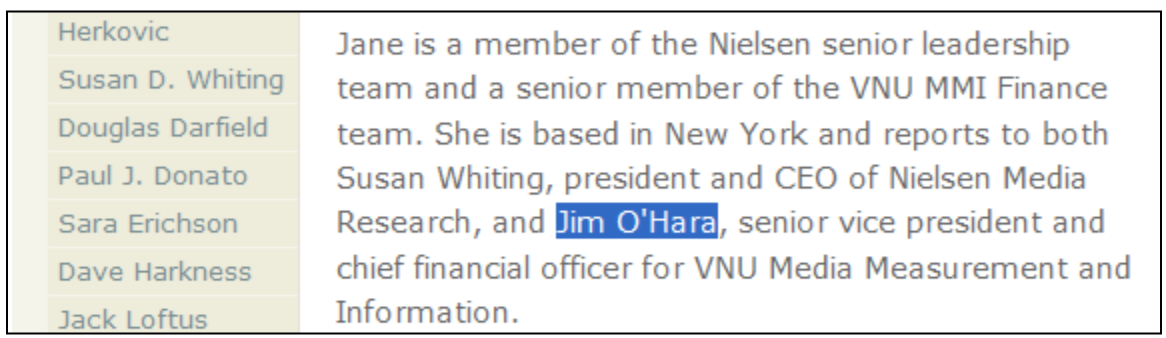


What is Information Extraction?

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

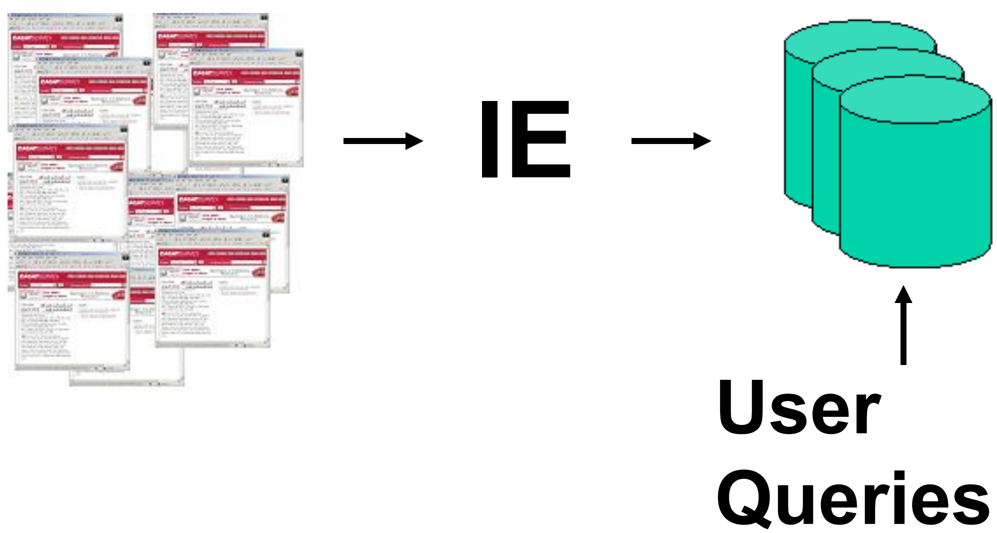






What is information extraction?

- 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 <u>document collections</u>
- Natural or unstructured
  - Little obvious structural information
- Partially structured
  - Contain some canonical formatting
- Highly structured
  - Often, automatically generated

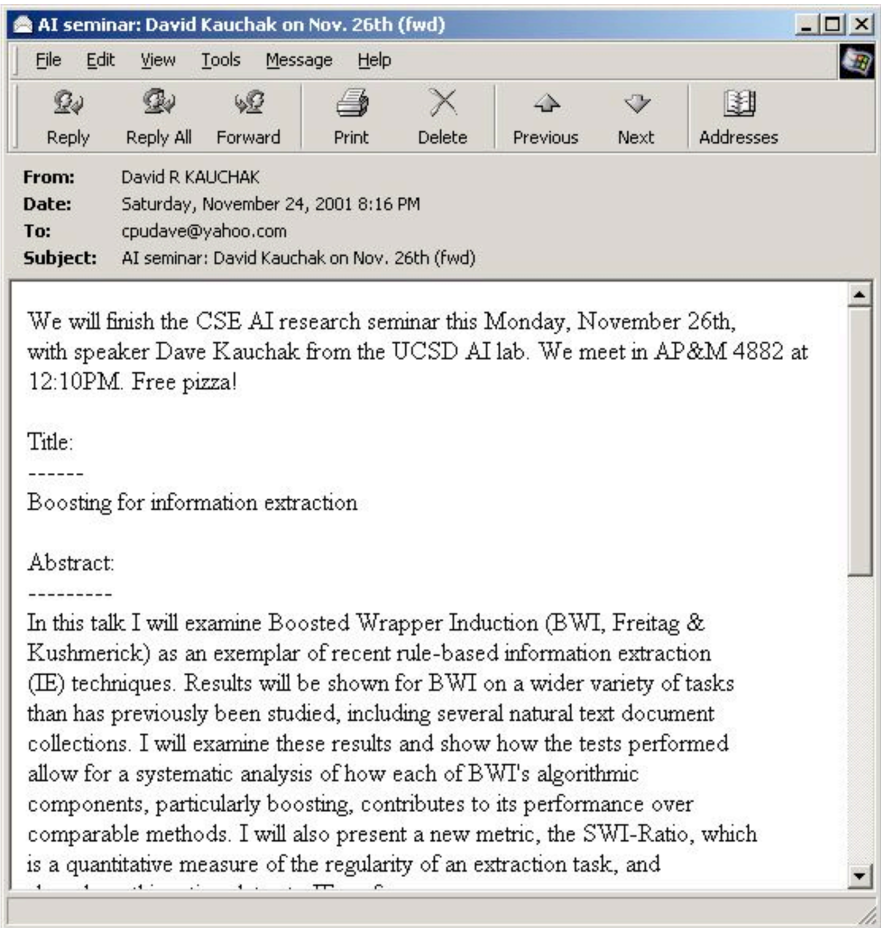
Natural Text: MEDLINE Journal Abstracts

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.



AI seminar: David Kauchak on Nov. 26th (fwd)

File Edit View Tools Message Help

Reply Reply All Forward Print Delete Previous Next Addresses

From: David R KAUCHAK
Date: Saturday, November 24, 2001 8:16 PM
To: cpudave@yahoo.com
Subject: AI seminar: David Kauchak on Nov. 26th (fwd)

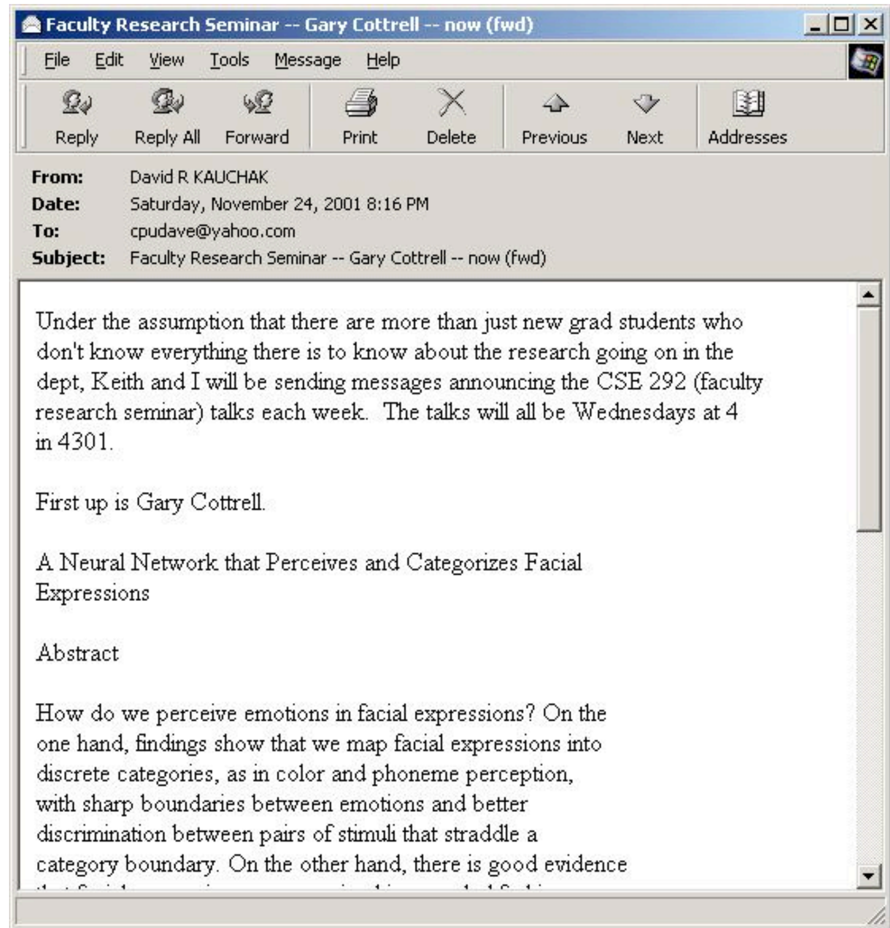
We will finish the CSE AI research seminar this Monday, November 26th, with speaker Dave Kauchak from the UCSD AI lab. We meet in AP&M 4882 at 12:10PM. Free pizza!

Title:

Boosting for information extraction

Abstract:

In this talk I will examine Boosted Wrapper Induction (BWI, Freitag & Kushmerick) as an exemplar of recent rule-based information extraction (IE) techniques. Results will be shown for BWI on a wider variety of tasks than has previously been studied, including several natural text document collections. I will examine these results and show how the tests performed allow for a systematic analysis of how each of BWI's algorithmic components, particularly boosting, contributes to its performance over comparable methods. I will also present a new metric, the SWI-Ratio, which is a quantitative measure of the regularity of an extraction task, and



Faculty Research Seminar -- Gary Cottrell -- now (fwd)

File Edit View Tools Message Help

Reply Reply All Forward Print Delete Previous Next Addresses

From: David R KAUCHAK
Date: Saturday, November 24, 2001 8:16 PM
To: cpudave@yahoo.com
Subject: Faculty Research Seminar -- Gary Cottrell -- now (fwd)

Under the assumption that there are more than just new grad students who don't know everything there is to know about the research going on in the dept, Keith and I will be sending messages announcing the CSE 292 (faculty research seminar) talks each week. The talks will all be Wednesdays at 4 in 4301.

First up is Gary Cottrell.

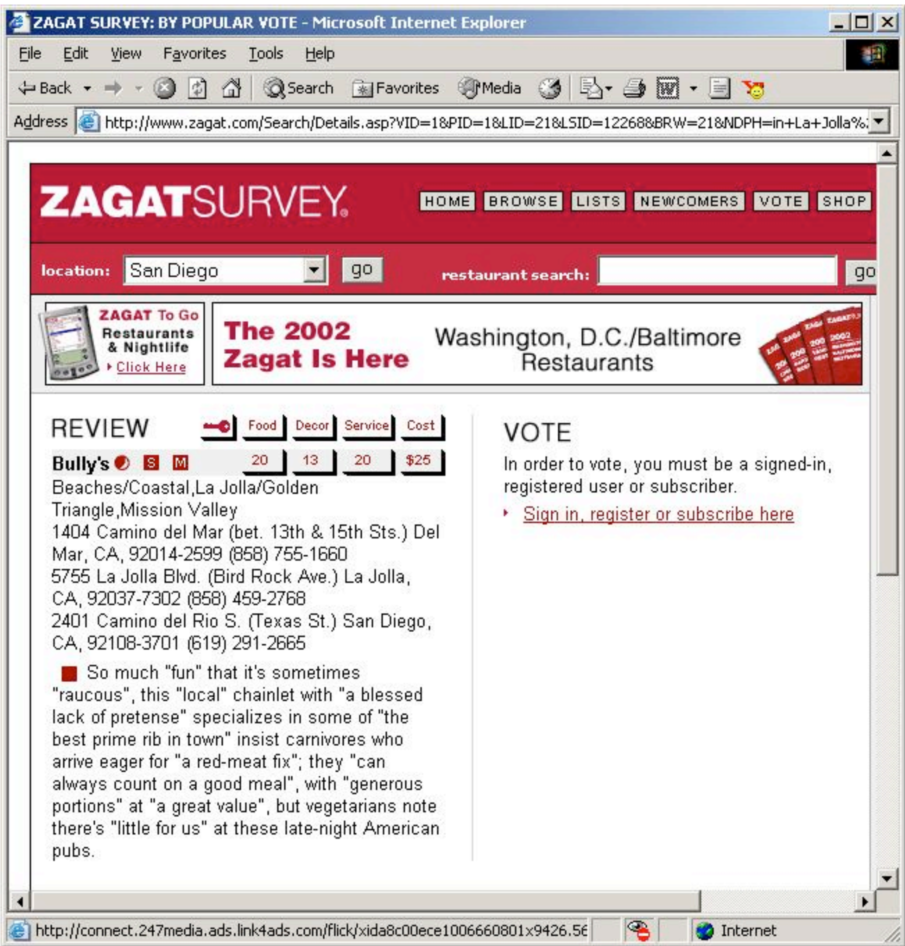
A Neural Network that Perceives and Categorizes Facial Expressions

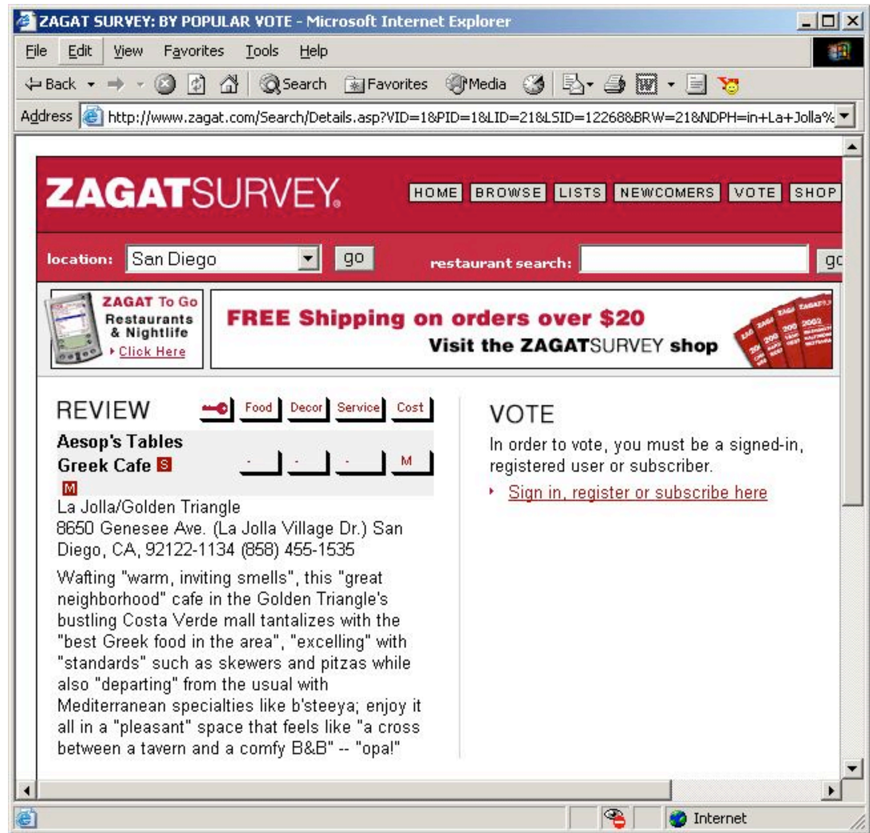
Abstract

How do we perceive emotions in facial expressions? On the one hand, findings show that we map facial expressions into discrete categories, as in color and phoneme perception, with sharp boundaries between emotions and better discrimination between pairs of stimuli that straddle a category boundary. On the other hand, there is good evidence

Highly Structured: Zagat's Reviews

Extract restaurant, location, cost, etc.



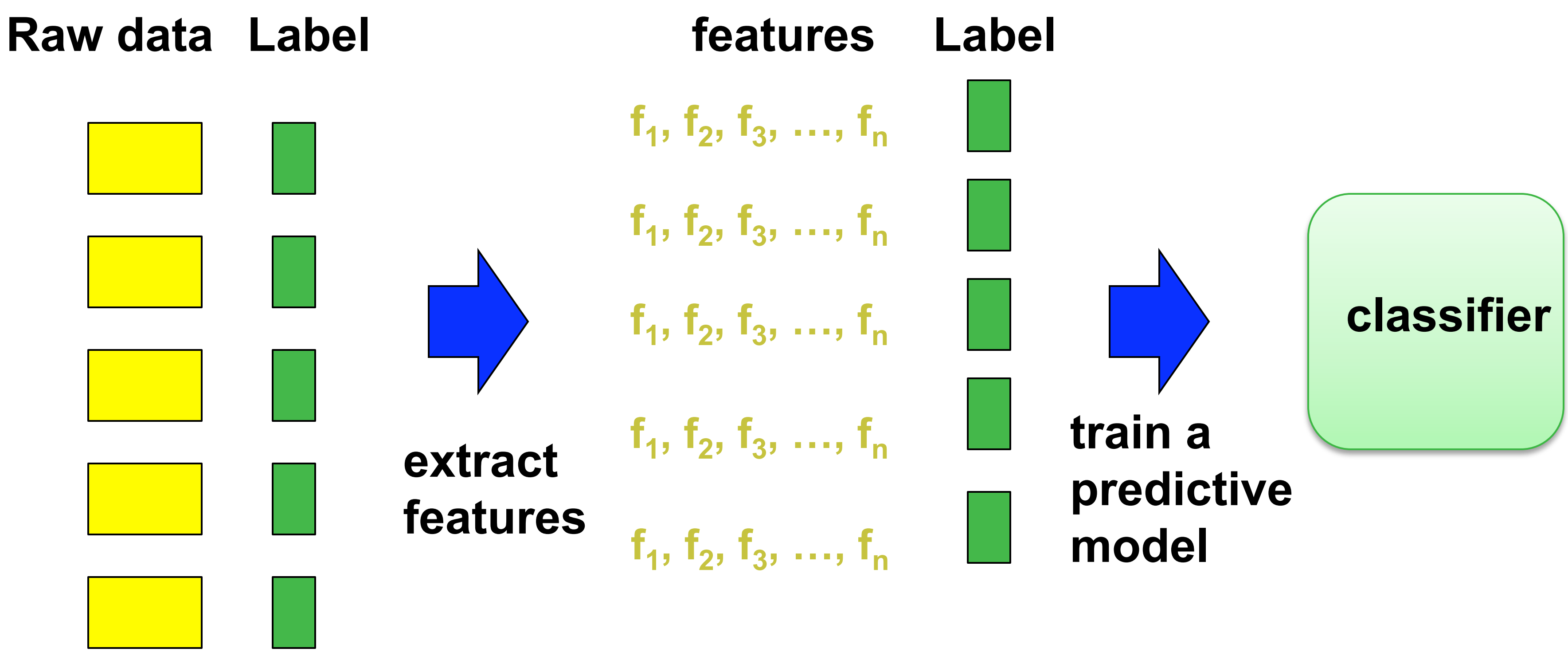


IE for Information Retrieval

- 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”

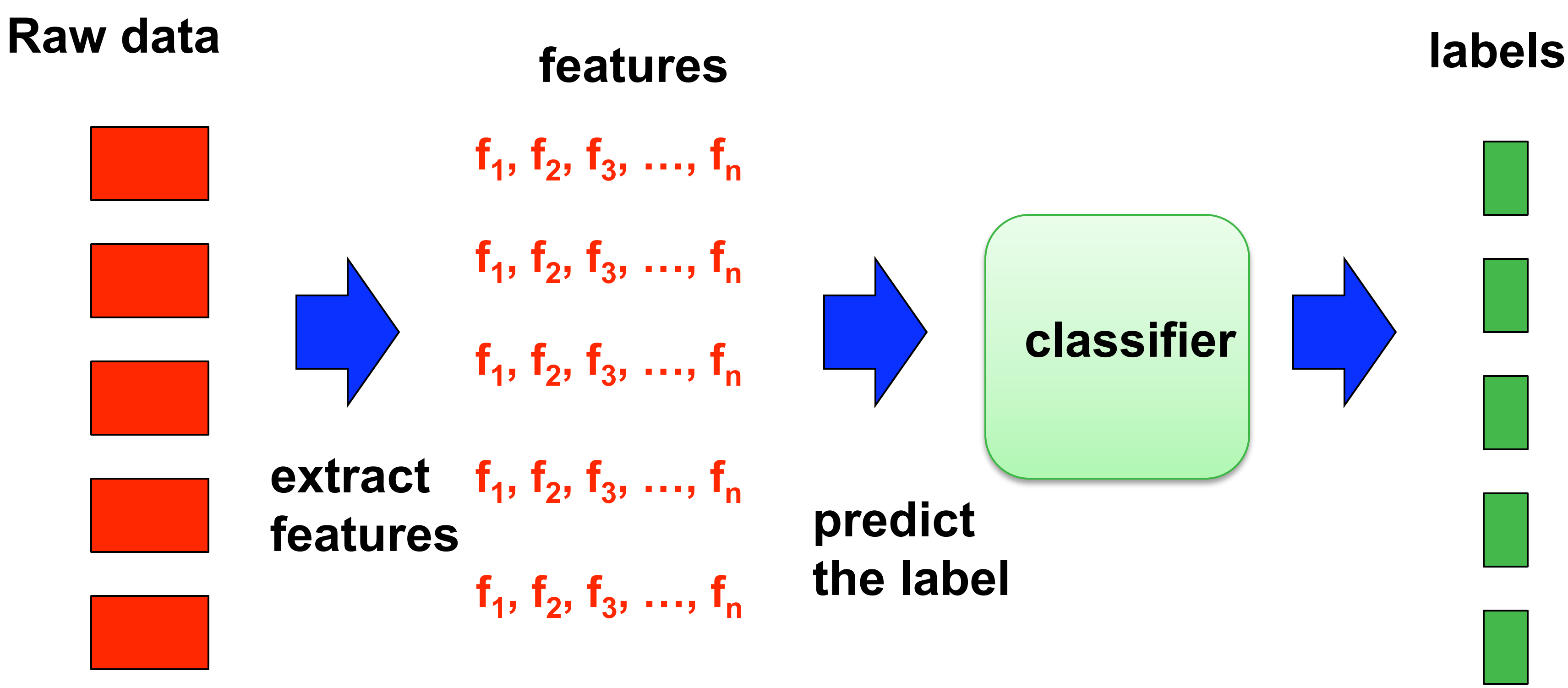
Classifier setup

Training or learning phase



Classifier setup

Testing or classification phase



IE Posed as a Machine Learning Task

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

$$\ldots\ \underbrace{\textbf{00 : pm Place :}}_{\text{prefix}}\ \underbrace{\textbf{Wean Hall Rm 5409}}_{\text{contents}}\ \underbrace{\textbf{Speaker : Sebastian Thrun}}_{\text{suffix}}\ \ldots$$

What features would be useful?

Good Features for Information Extraction

Creativity and Domain Knowledge Required!

begins-with-number
begins-with-ordinal
begins-with-punctuation
begins-with-question-word
begins-with-subject
blank
contains-alphanum
contains-bracketed-number
contains-http
contains-non-space
contains-number
contains-pipe

Example word features:

– identity of word
– is in all caps
– ends in “-ski”
– is part of a noun phrase
– is in a list of city names
– is under node X in WordNet or Cyc
– is in bold font
– is in hyperlink anchor
– *features of past & future*
– last person name was female
– next two words are “and Associates”

contains-question-mark
contains-question-word
ends-with-question-mark
first-alpha-is-capitalized
indented
indented-1-to-4
indented-5-to-10
more-than-one-third-space
only-punctuation
prev-is-blank
prev-begins-with-ordinal
shorter-than-30

Good Features for Information Extraction

Creativity and Domain Knowledge Required!

Is Capitalized
Is Mixed Caps
Is All Caps
Initial Cap
Contains Digit
All lowercase
Is Initial
Punctuation
Period
Comma
Apostrophe
Dash
Preceded by HTML tag

Character n-gram classifier
 says string is a person
 name (80% accurate)
In stopword list
 (the, of, their, etc)
In honorific list
 (Mr, Mrs, Dr, Sen, etc)
In person suffix list
 (Jr, Sr, PhD, etc)
In name particle list
 (de, la, van, der, etc)
In Census lastname list;
 segmented by P(name)
In Census firstname list;
 segmented by P(name)
In locations lists
 (states, cities, countries)
In company name list
 ("J. C. Penny")
In list of company suffixes
 (Inc, & Associates, Foundation)

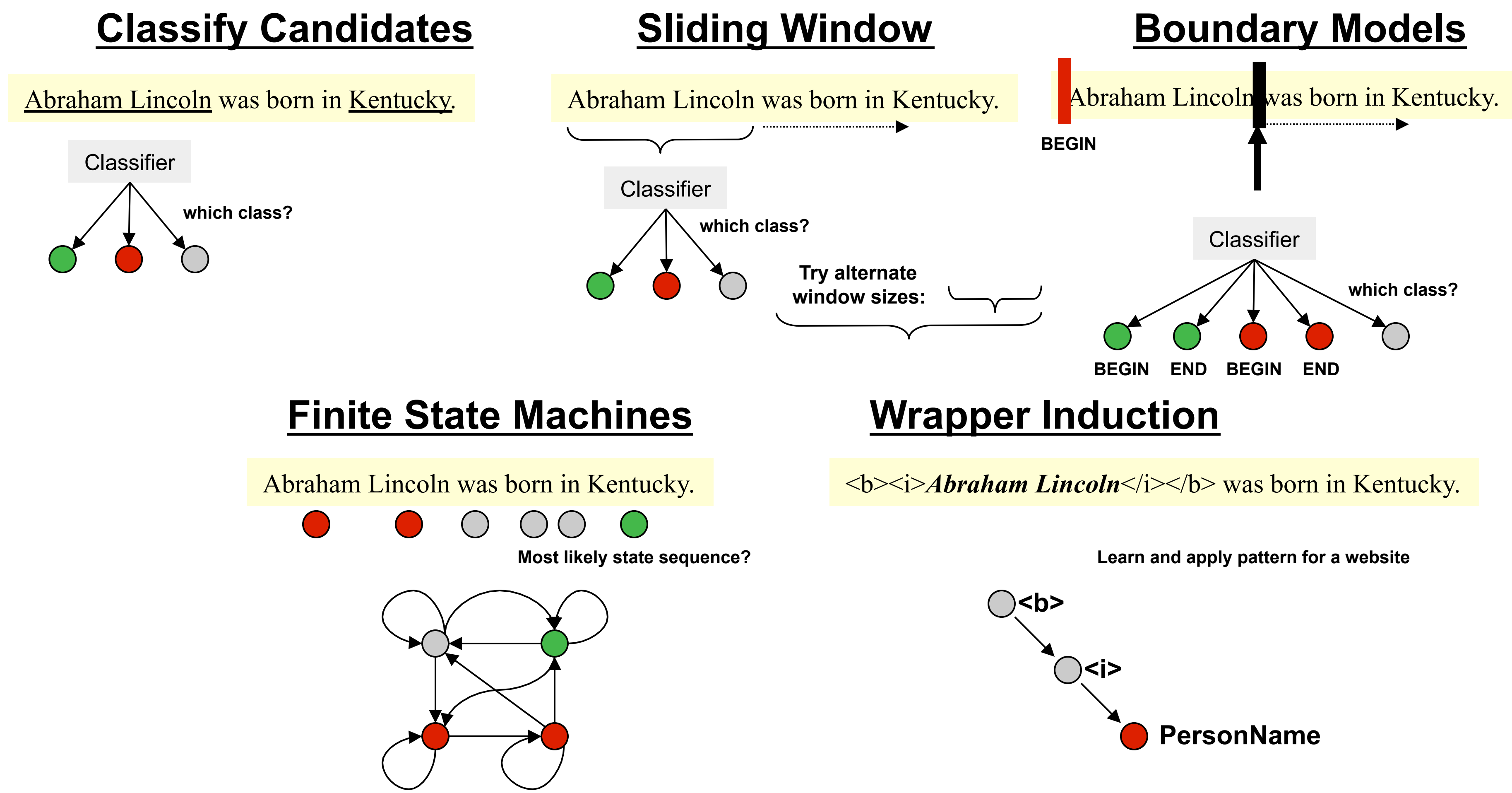
Word Features
- lists of job titles,
- Lists of prefixes
- Lists of suffixes
- 350 informative phrases

HTML/Formatting Features
- {begin, end, in} x {<b>, <i>, <a>, <hN>} x {lengths 1, 2, 3, 4, or longer}
- {begin, end} of line

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

classifier

Landscape of ML Techniques for IE:



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]

… **00 : pm Place : Wean Hall Rm 5409 Speaker : Sebastian Thrun** …

w_{t-m} w_{t-1} w_{t} w_{t+n} w_{t+n+1} w_{t+n+m}

prefix **contents** **suffix**

- 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

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

**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.
```

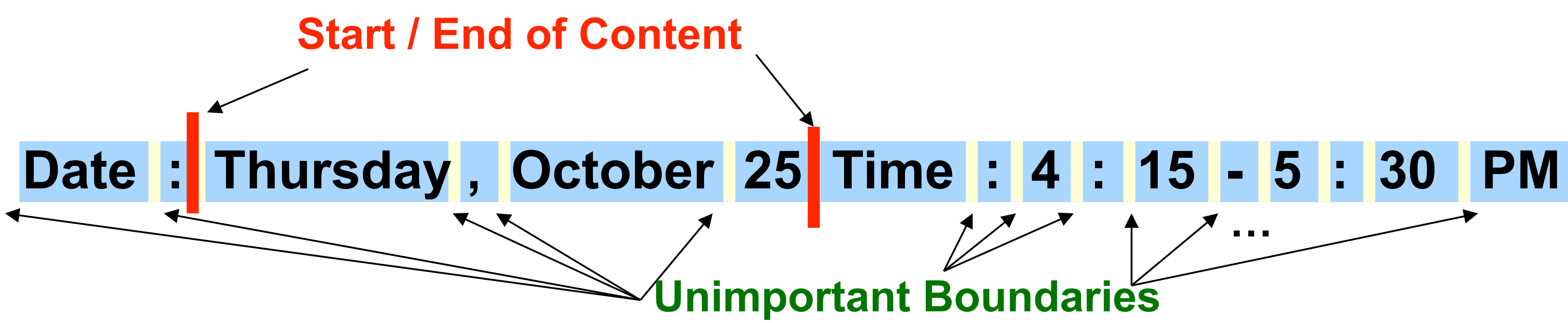
CMU UseNet Seminar Announcement

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

Begin

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

POSITIVE (1)

End

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

ALL OTHERS NEGATIVE (0)

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

***Begin*(*i*)=** 1 if *i* begins a field
0 otherwise

Boundary Detectors

- "Boundary Detectors" are pairs of token sequences ‹p,s›
  - 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:
  - <Date:,[*CapitalizedWord*]> matches:

Date: Thursday, October 25

Another Detector Example

	Prefix	Suffix
***Begin* boundary detector:**	< a href = "	http
***End* boundary detector:**	*empty*	">

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)* = Prob(position *i* starts a field)
  - *End(j)* = Prob(position *j* ends a field)
  - *Len(k)* = Prob(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 to 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

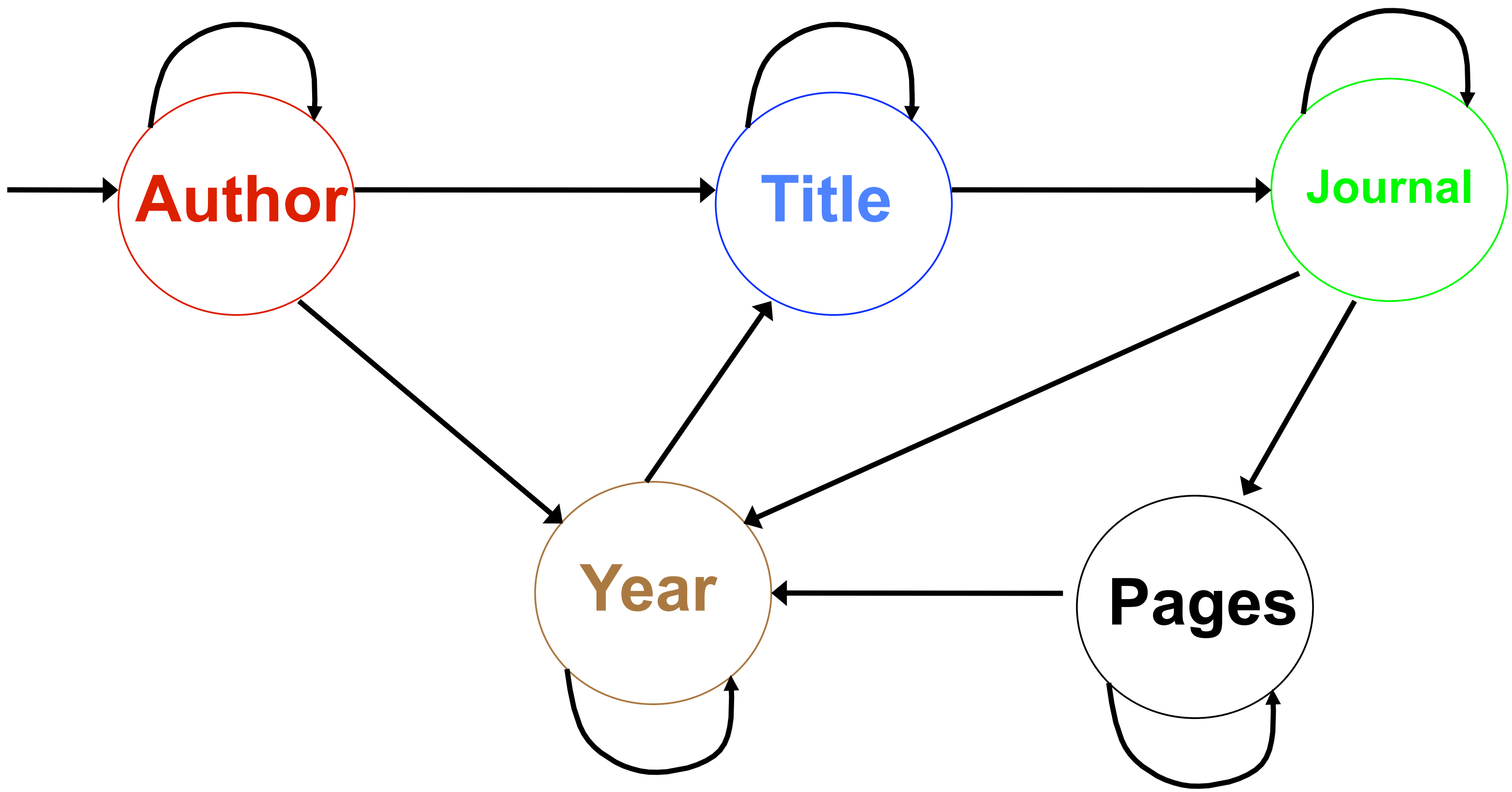
Fahlman, S. E. (1991) The recurrent cascade

title title title journal pages

correlation learning architecture. NIPS 3, 190-205.

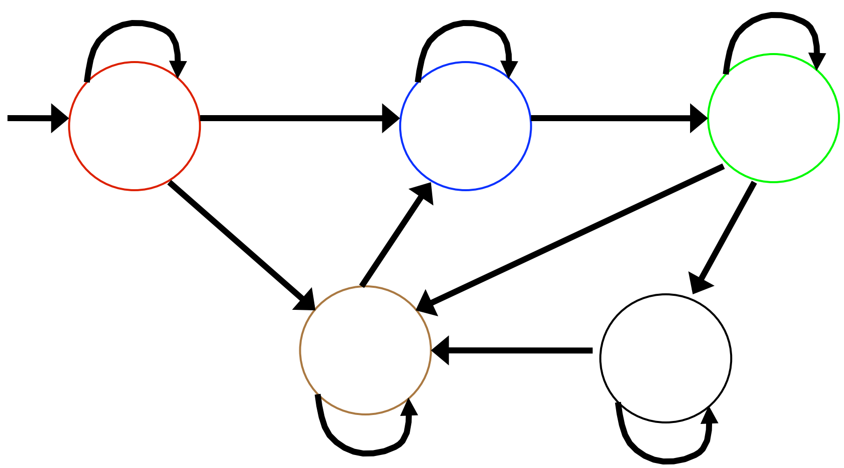
Ideas?

Hiddent 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:

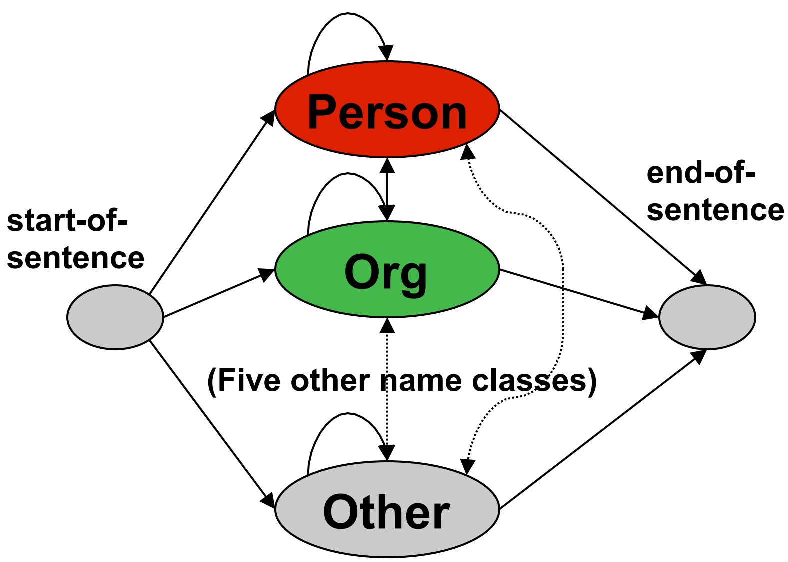
$$\prod_i a[x_i \mid x_{i-1}]\, b[o_i \mid x_i]$$

- 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:

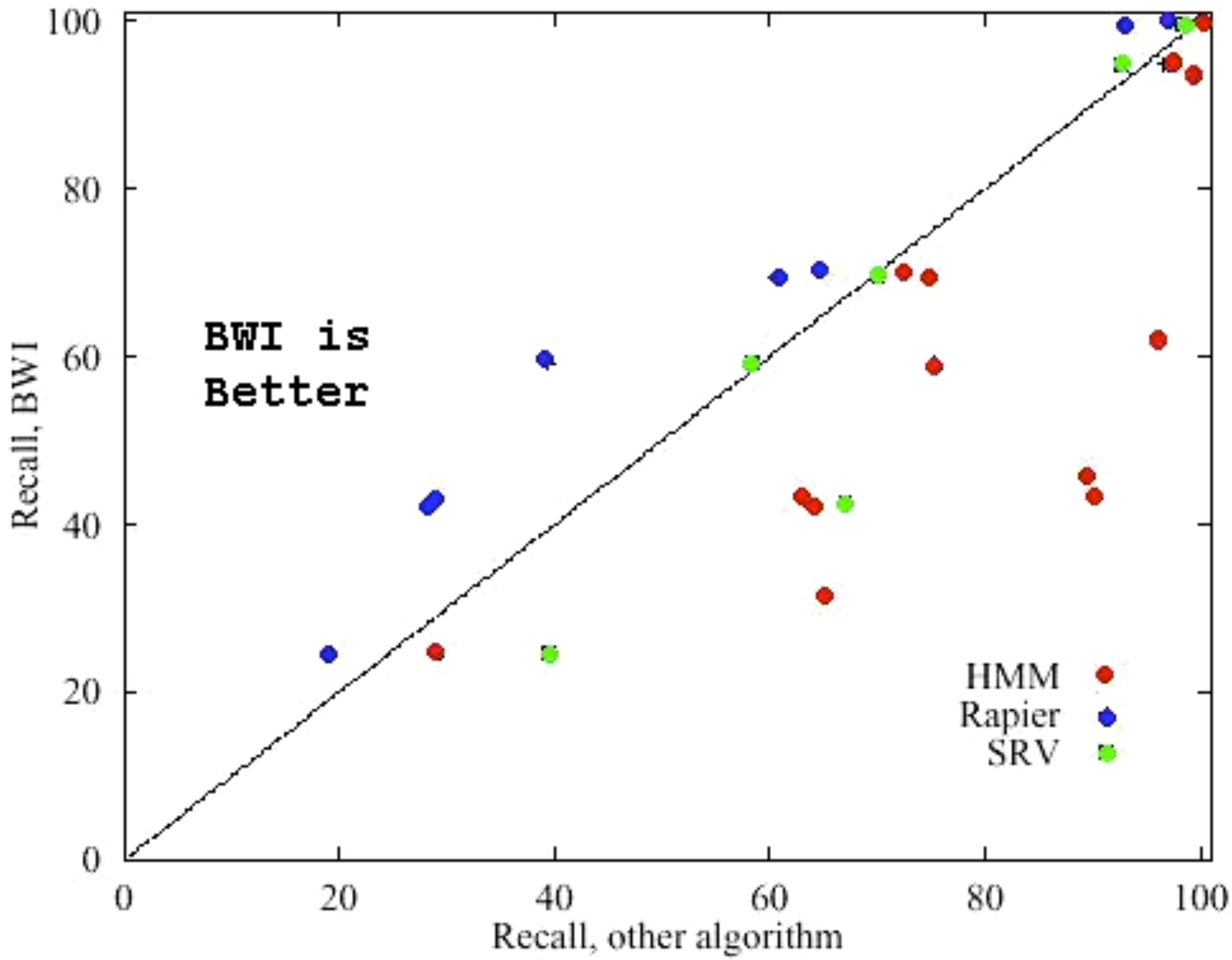
Case	Language	F1 .
Mixed	English	93%
Upper	English	91%
Mixed	Spanish	90%

System Comparison

- 16 different extraction tasks

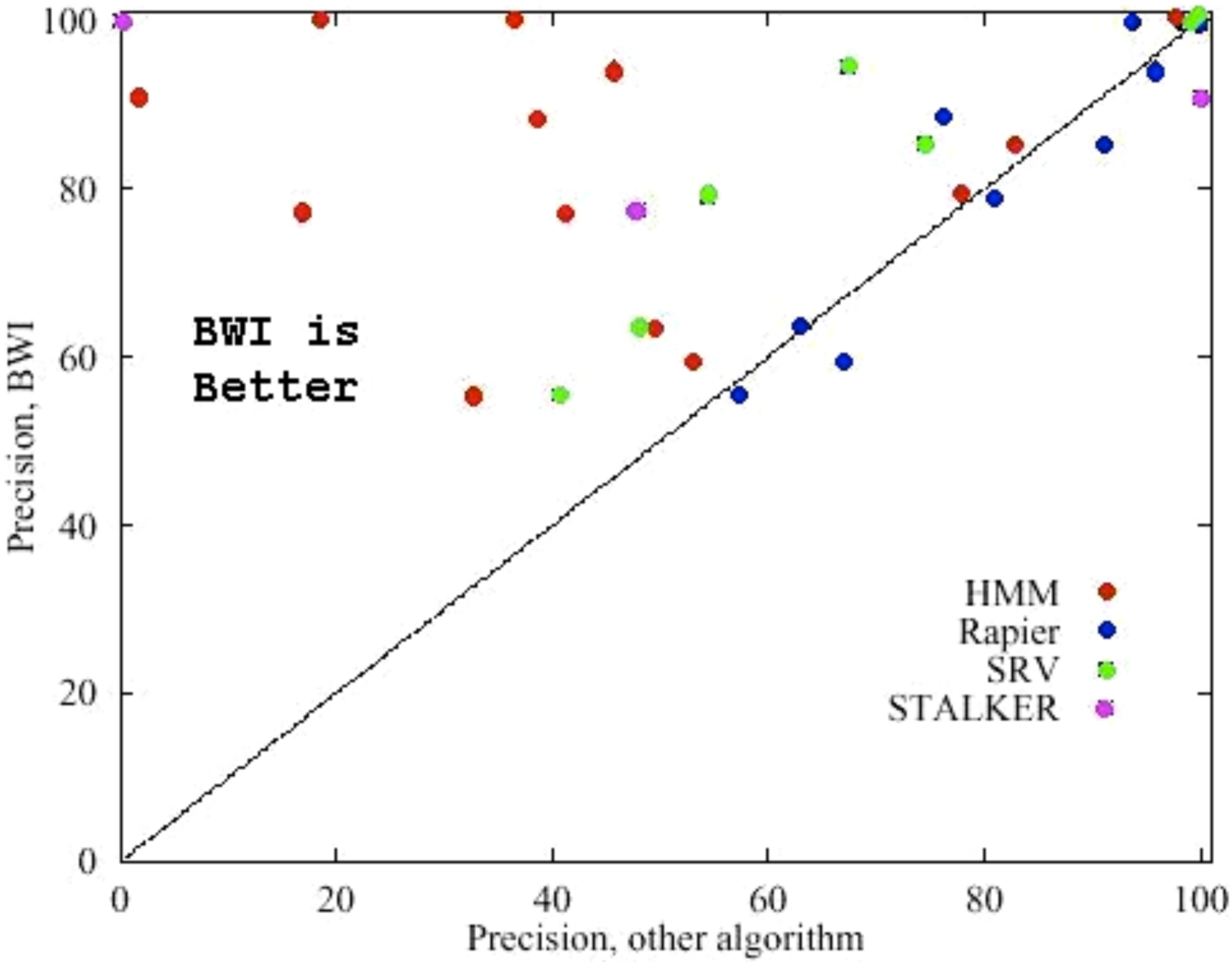
- Algorithms
  - BWI vs.
  - Two rule learners: SRV and Rapier
  - One algorithm based on hidden Markov models
  - One wrapper induction algorithm: Stalker

Recall



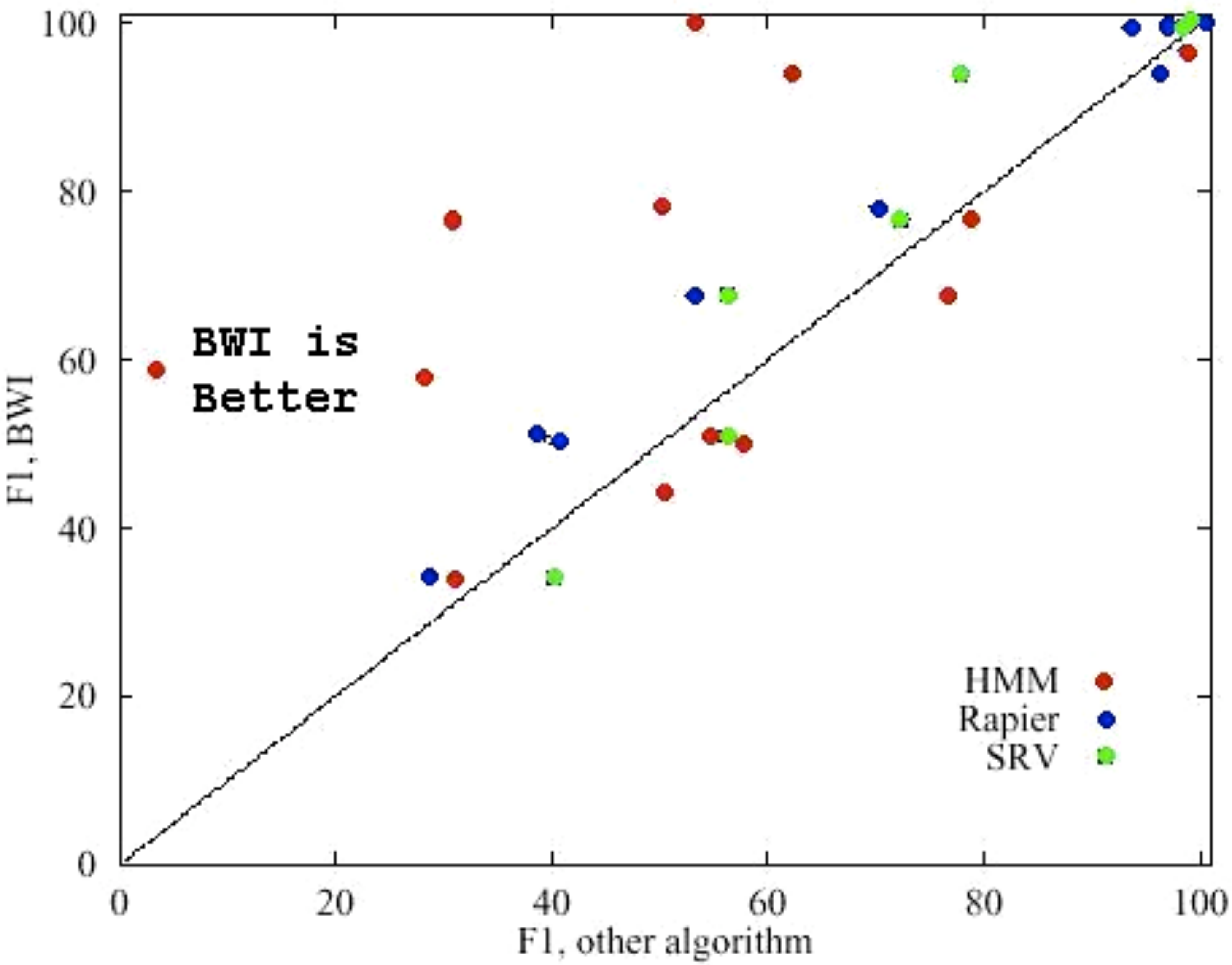
BWI is
Better
Recall, BWI
Recall, other algorithm
HMM
Rapier
SRV
0
20
40
60
80
100

Precision



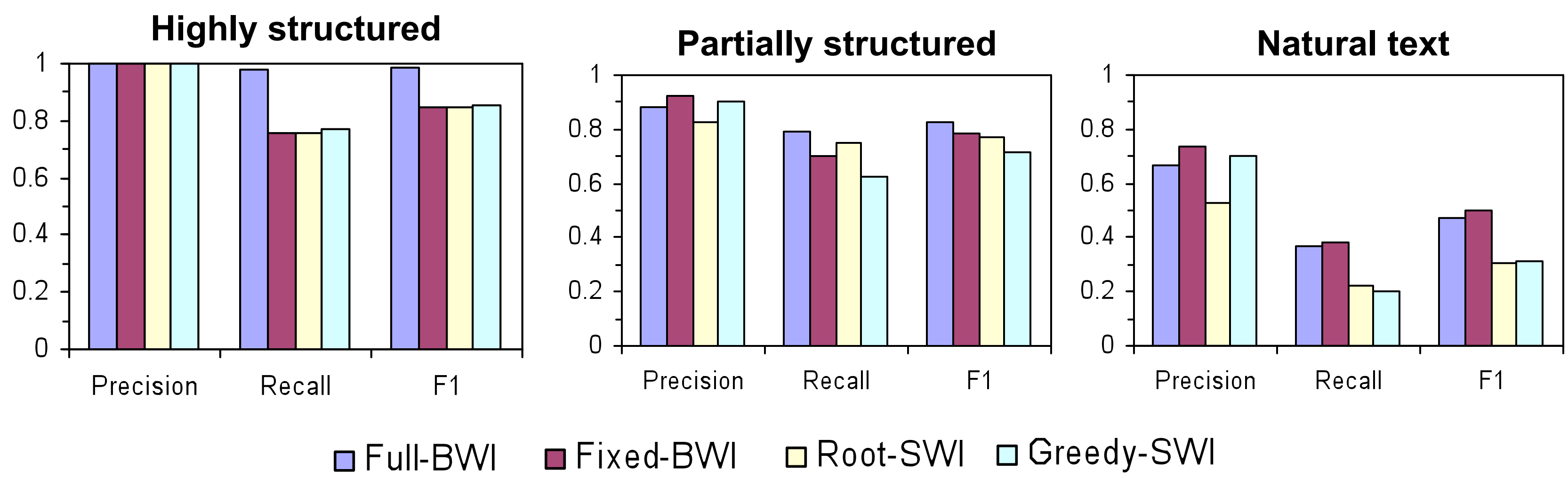
100
80
60
40
20
0
Precision, BWI
BWI is
Better
HMM
Rapier
SRV
STALKER
0
20
40
60
80
100
Precision, other algorithm

F1



BWI is Better
F1, BWI
F1, other algorithm
HMM
Rapier
SRV
0
20
40
60
80
100

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	95.0
default		67.7	76.7	99.4	94.6
lexical		73.5	-	-	-

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

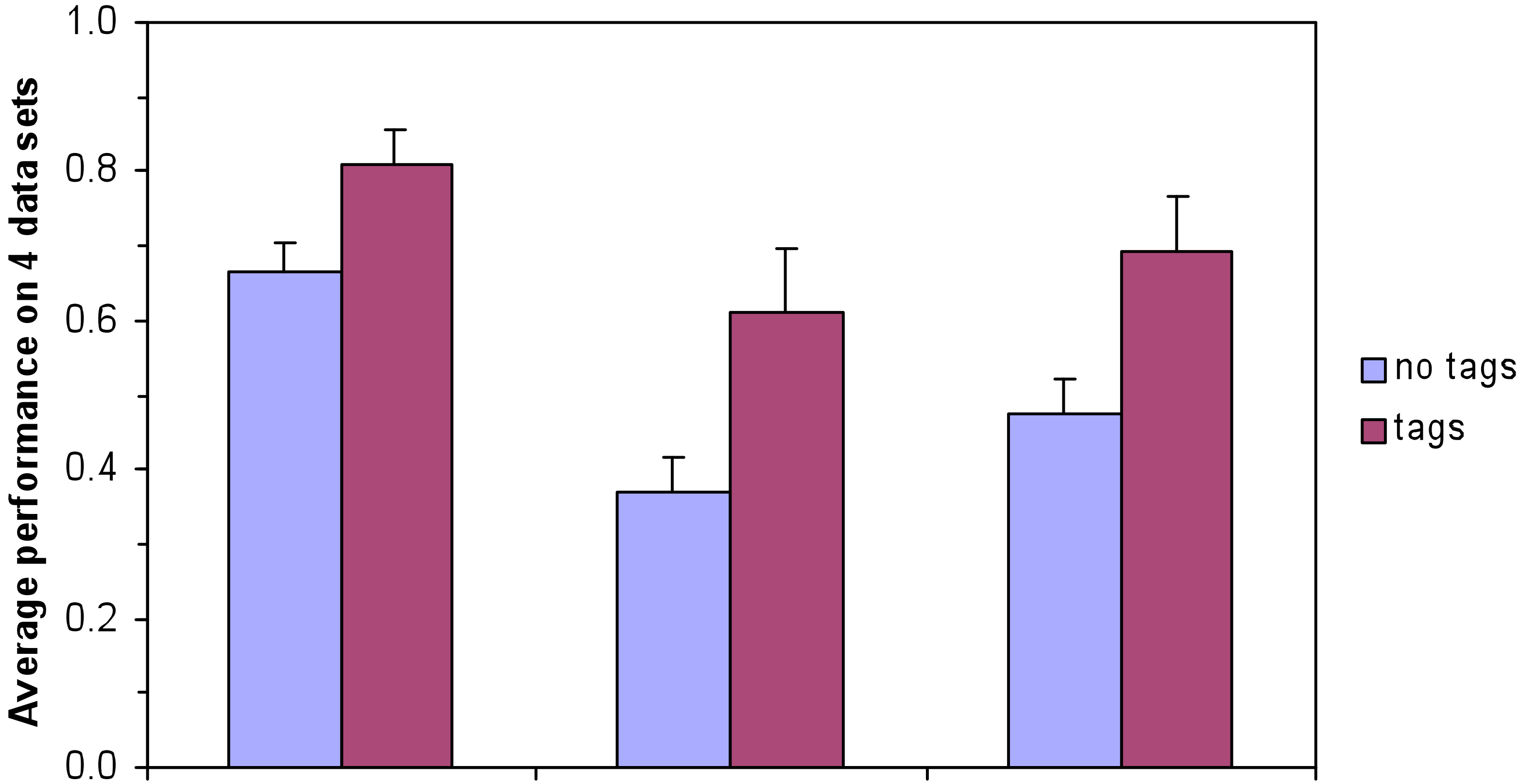
- Instead of altering methods, alter text
- Idea: Add limited grammatical information
  - Run shallow parser over text
  - Flatten parse tree and insert as tags

Example of Tagged Sentence:

Uba2p is located largely in the nucleus.

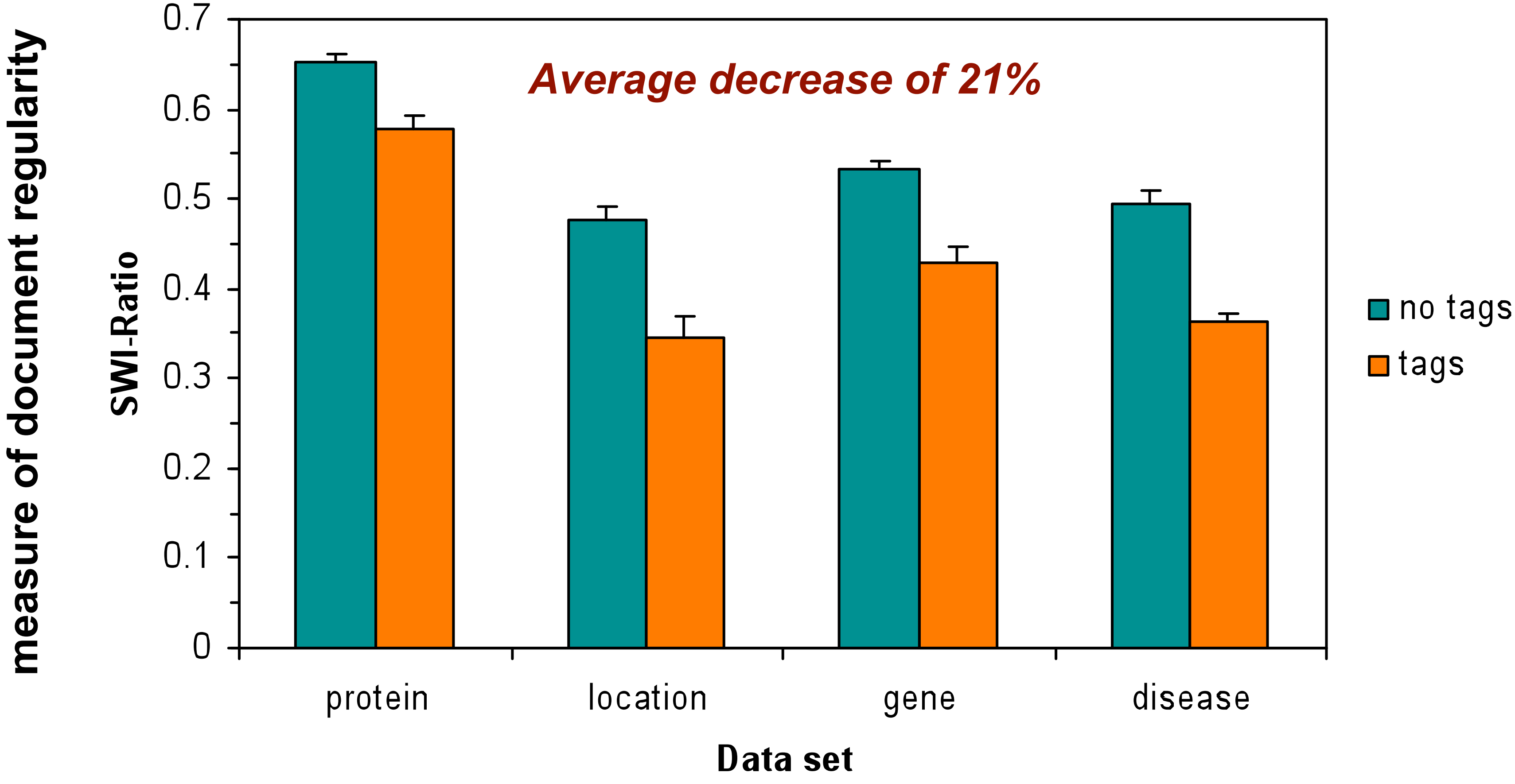
NP_SEG VP_SEG PP_SEG NP_SEG

Tagging Results on Natural Domain



21% increase ***65% increase*** ***45% increase***

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?