Information Extraction from the World Wide Web

Andrew McCallum
University of Massachusetts Amherst

William Cohen
Carnegie Mellon University

Example: The Problem

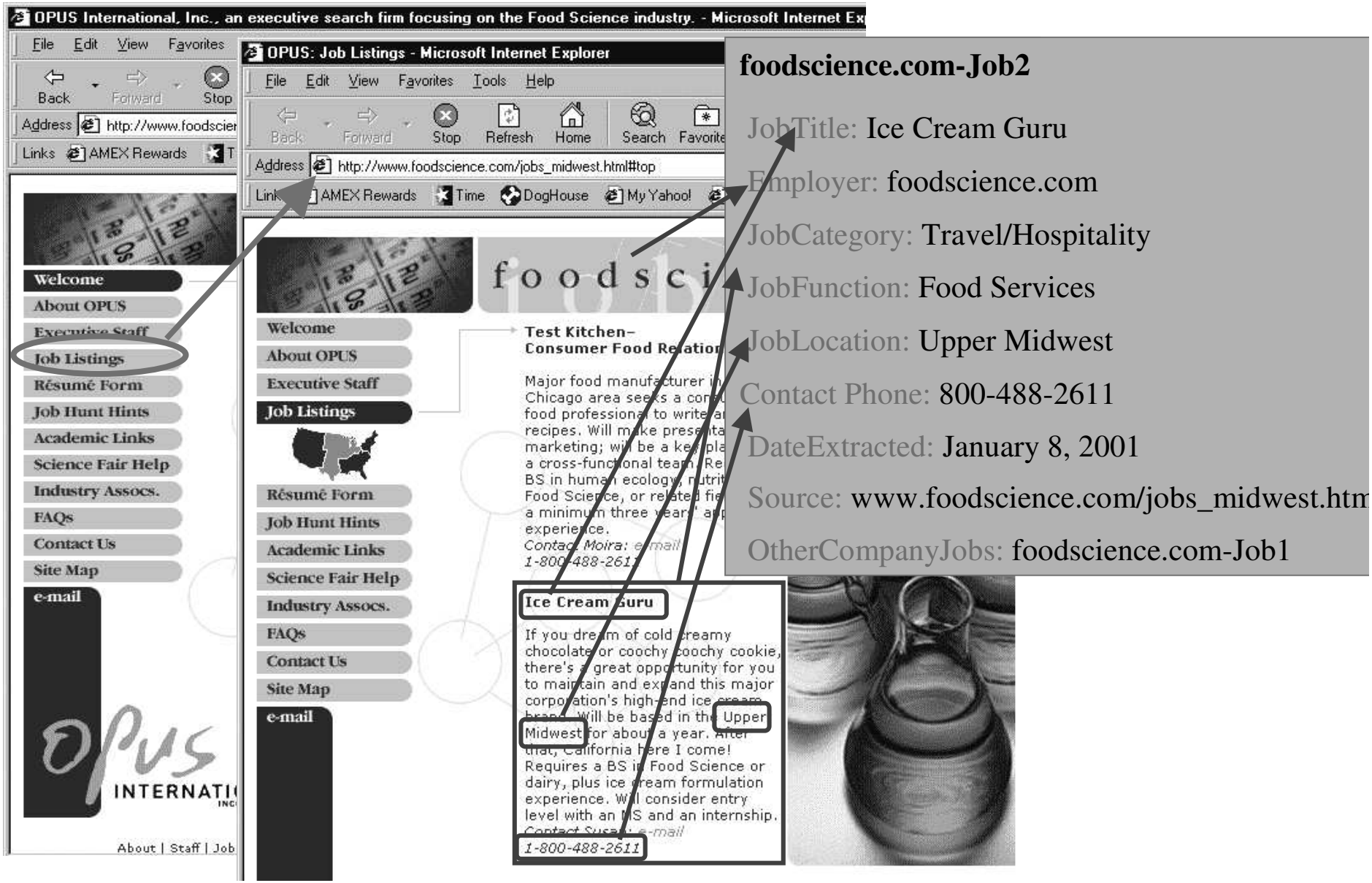
***Martin Baker*, a person**

Genomics job

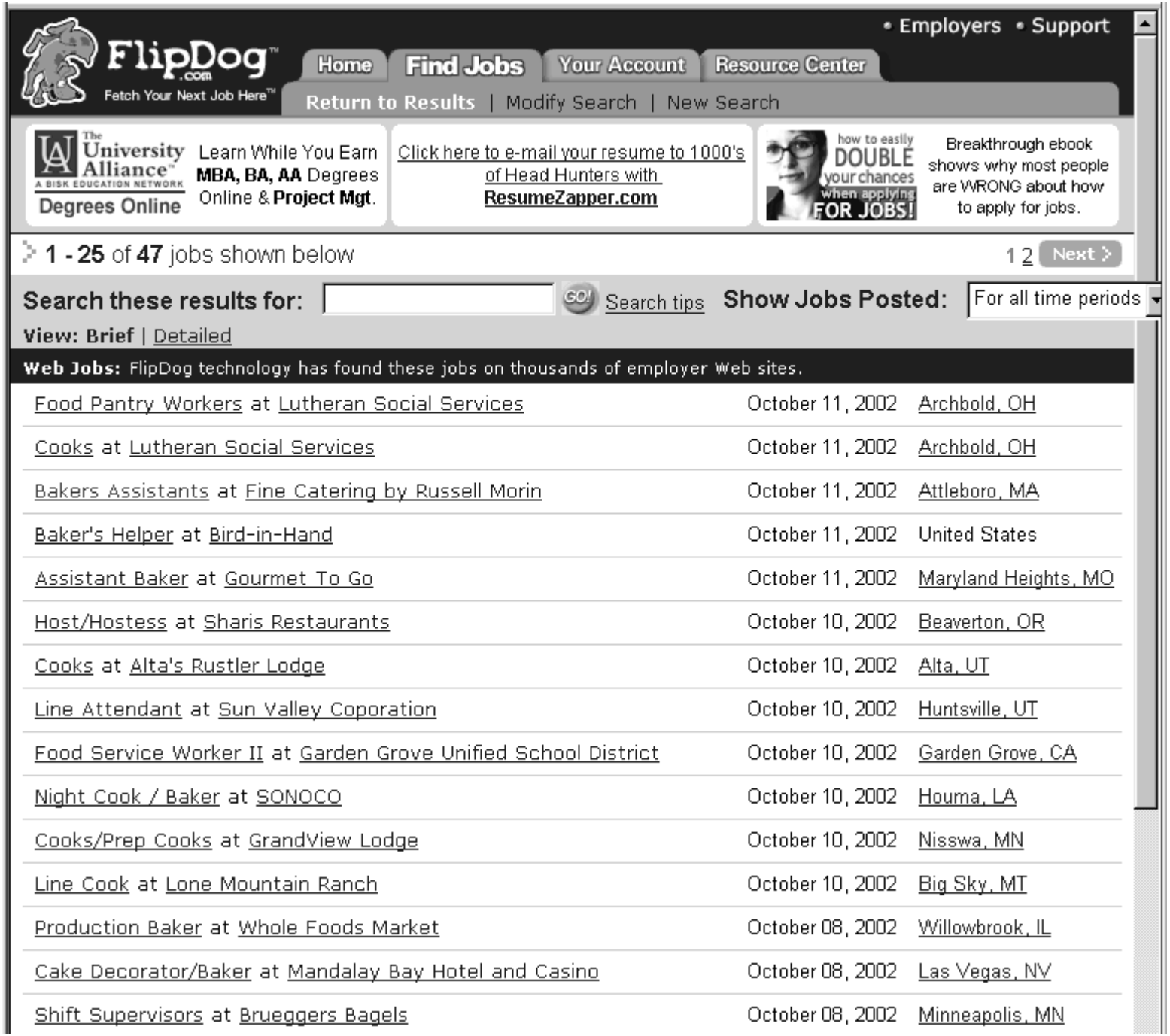
Employers job posting form

Example: A Solution

Extracting Job Openings from the Web



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



FlipDog.com — Fetch Your Next Job Here™

• Employers • Support

Home | Find Jobs | Your Account | Resource Center

Return to Results | Modify Search | New Search

The University Alliance™ — A BISK EDUCATION NETWORK — Degrees Online
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**

how to easily DOUBLE your chances when applying FOR JOBS!
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: GO! 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.

Food Pantry Workers at Lutheran Social Services	October 11, 2002	Archbold, OH
Cooks at Lutheran Social Services	October 11, 2002	Archbold, OH
Bakers Assistants at Fine Catering by Russell Morin	October 11, 2002	Attleboro, MA
Baker's Helper at Bird-in-Hand	October 11, 2002	United States
Assistant Baker at Gourmet To Go	October 11, 2002	Maryland Heights, MO
Host/Hostess at Sharis Restaurants	October 10, 2002	Beaverton, OR
Cooks at Alta's Rustler Lodge	October 10, 2002	Alta, UT
Line Attendant at Sun Valley Coporation	October 10, 2002	Huntsville, UT
Food Service Worker II at Garden Grove Unified School District	October 10, 2002	Garden Grove, CA
Night Cook / Baker at SONOCO	October 10, 2002	Houma, LA
Cooks/Prep Cooks at GrandView Lodge	October 10, 2002	Nisswa, MN
Line Cook at Lone Mountain Ranch	October 10, 2002	Big Sky, MT
Production Baker at Whole Foods Market	October 08, 2002	Willowbrook, IL
Cake Decorator/Baker at Mandalay Bay Hotel and Casino	October 08, 2002	Las Vegas, NV
Shift Supervisors at Brueggers Bagels	October 08, 2002	Minneapolis, MN

Data Mining the Extracted Job Information

What is “Information Extraction”

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

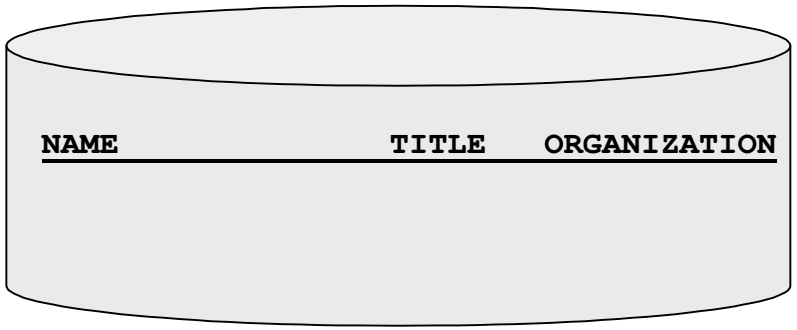
October 14, 2002, 4:00 a.m. PT

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.**

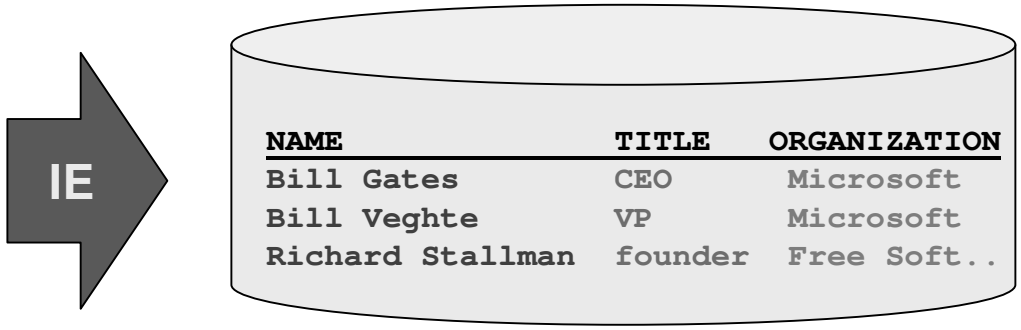
October 14, 2002, 4:00 a.m. PT

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…



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

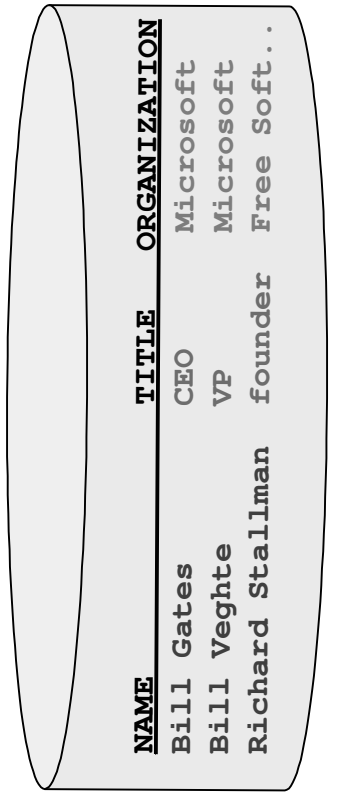
* Microsoft
Gates

* Microsoft

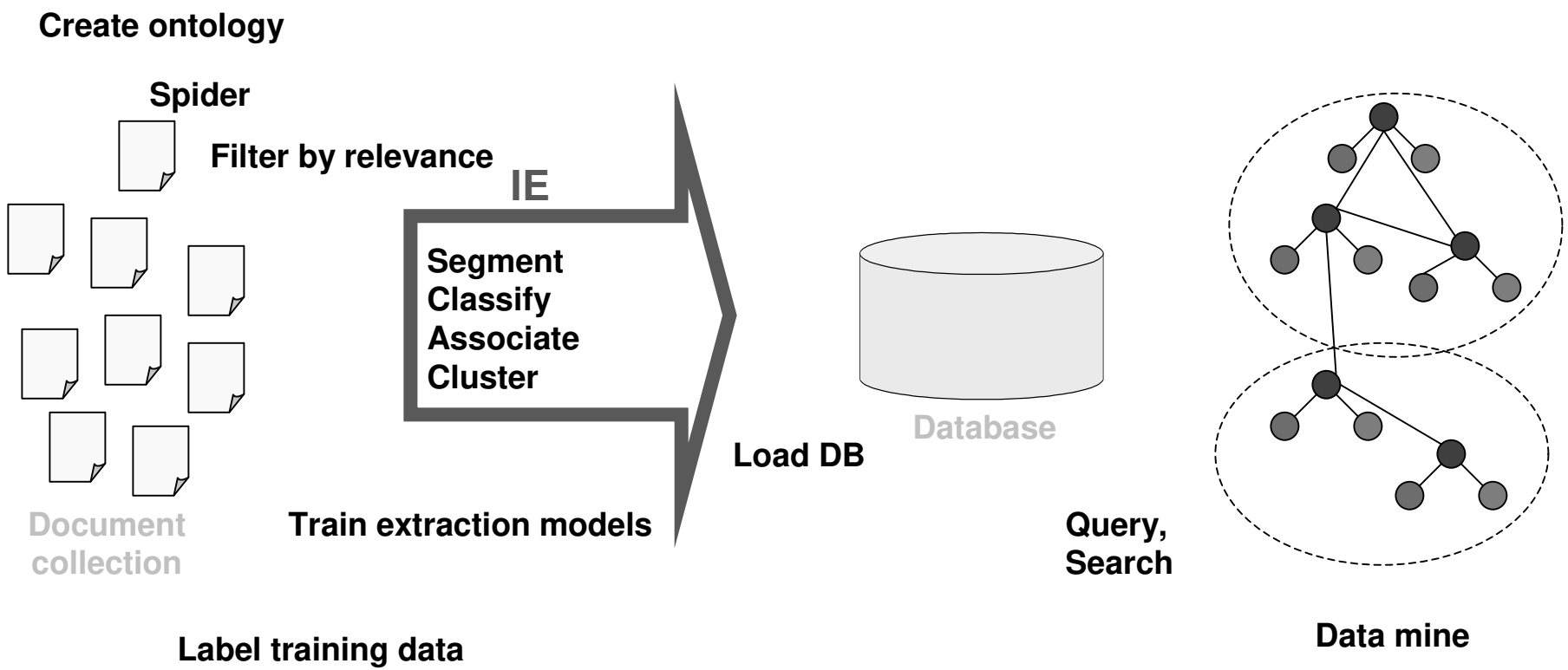
Bill Veghte

* Microsoft
VP

Richard Stallman
founder
Free Software Foundation



IE in Context



Why IE from the Web?

- Science
  - Grand old dream of AI: Build large KB* and reason with it. IE from the Web enables the creation of this KB.
  - IE from the Web 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"

Tutorial Outline

- IE History
- Landscape of problems and solutions
- Parade of models for segmenting/classifying:
  - Sliding window
  - Boundary finding
  - Finite state machines
  - Trees
- Overview of related problems and solutions
- Where to go from here

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

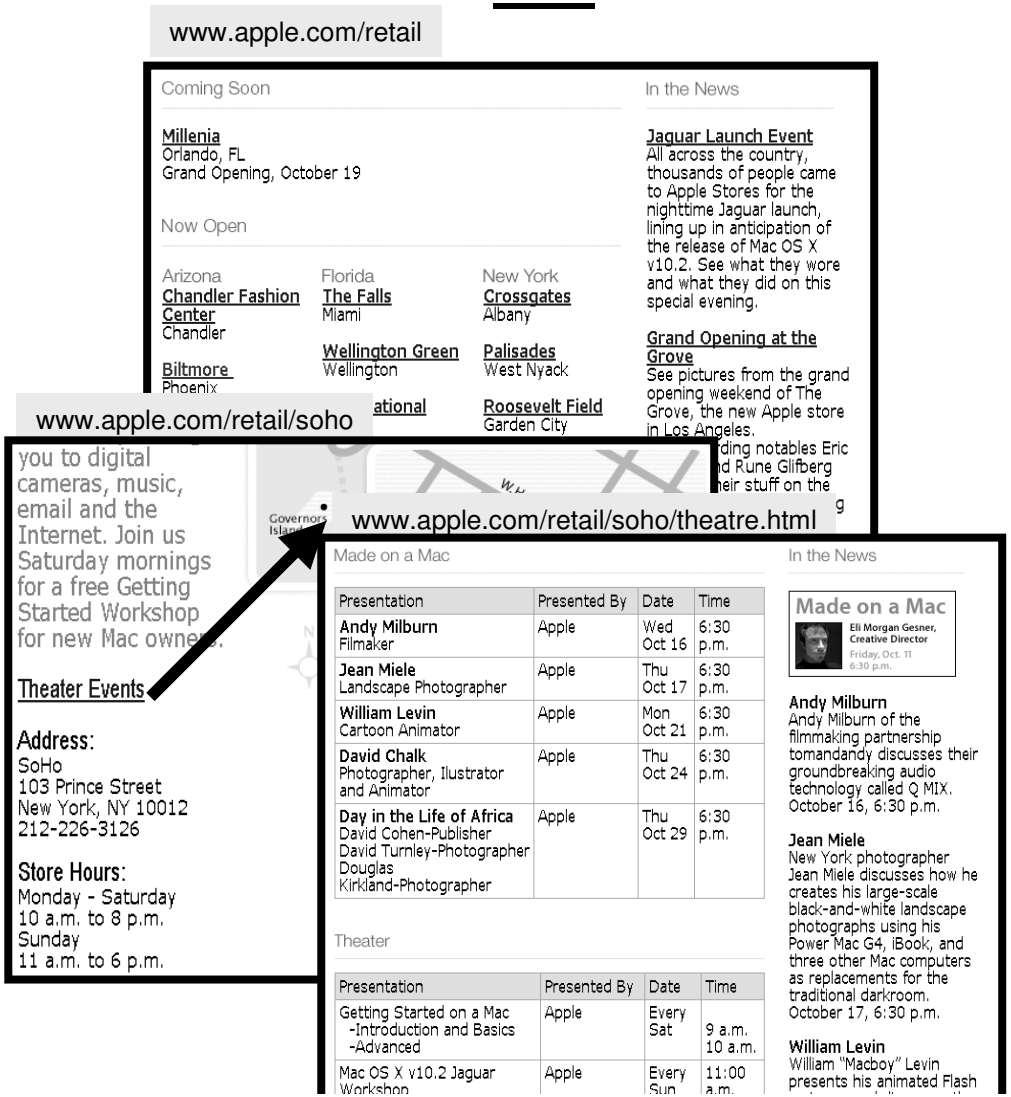
Apple to Open Its First Retail Store in New York City

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



Landscape of IE Tasks (1/4): Pattern Feature Domain

Text paragraphs without formatting

Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.

Grammatical sentences and some formatting & links

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

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

- Press

Contact

- General information
- Directions maps

Non-grammatical snippets, rich formatting & links

Barto, Andrew G.	(413) 545-2109	barto@cs.umass.edu	CS276
Professor. Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.			
Berger, Emery D.	(413) 577-4211	emery@cs.umass.edu	CS344
Assistant Professor.			
Brock, Oliver	(413) 577-0334	oli@cs.umass.edu	CS246
Assistant Professor.			
Clarke, Lori A.	(413) 545-1328	clarke@cs.umass.edu	CS304
Professor. Software verification, testing, and analysis; software architecture and design.			
Cohen, Paul R.	(413) 545-3638	cohen@cs.umass.edu	CS278
Professor. Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.			

Tables

8:30 - 9:30 AM	**Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty** *Joseph Y. Halpern, Cornell University*
9:30 - 10:00 AM	Coffee Break
10:00 - 11:30 AM	Technical Paper Sessions:

Cognitive Robotics	**Logic Programming**	**Natural Language Generation**	**Complexity Analysis**	**Neural Networks**	**Games**
739: A Logical Account of Causal and Topological Maps *Emilio Remolina and Benjamin Kuipers*	116: A-System: Problem Solving through Abduction *Marc Denecker, Antonis Kakas, and Bert Van Nuffelen*	758: Title Generation for Machine-Translated Documents *Rong Jin and Alexander G. Hauptmann*	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories *Marco Cadoli, Thomas Eiter, and Georg Gottlob*	179: Knowledge Extraction and Comparison from Local Function Networks *Kenneth McGarry, Stefan Wermter, and John MacIntyre*	71: Iterative Widening *Tristan Cazenave*
549: Online-Execution of ccGolog Plans *Henrik Grosskreutz*	131: A Comparative Study of Logic Programs with	246: Dealing with Dependencies between Content Planning and	470: A Perspective on Knowledge Compilation	258: Violation-Guided Learning for Constrained	353: Temporal Difference Learning Applied to a

Landscape of IE Tasks (2/4): Pattern Scope

Web site specific

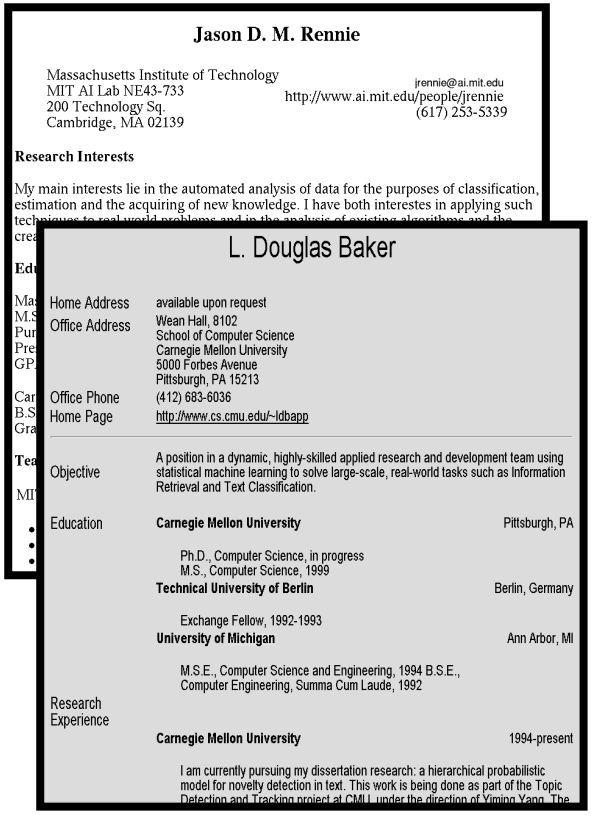
Formatting

Amazon.com Book Pages

Genre specific

Layout

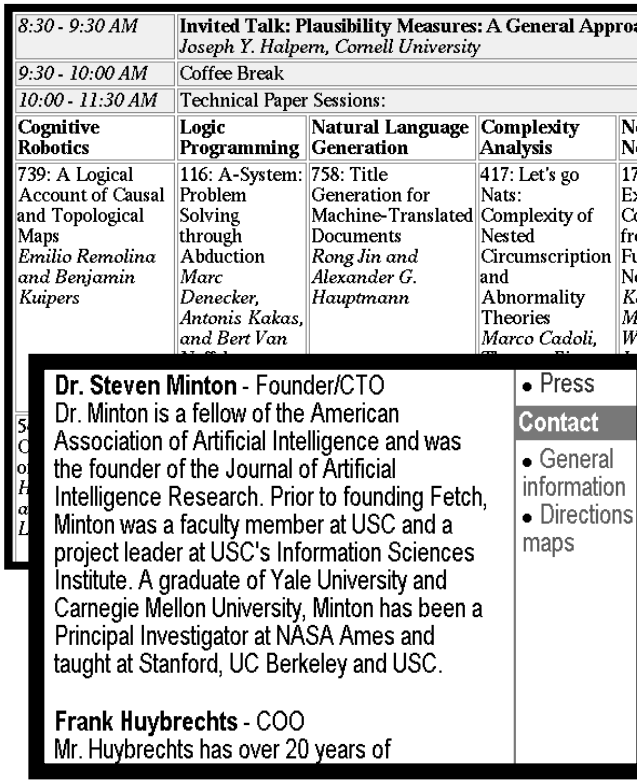
Resumes



Wide, non-specific

Language

University Names



Landscape of IE Tasks (3/4): Pattern Complexity

E.g. word patterns:

Closed set

U.S. states

He was born in Alabama…

The big Wyoming sky…

Regular set

U.S. phone numbers

Phone: (413) 545-1323

The CALD main office can be reached at 412-268-1299

Complex pattern

U.S. postal addresses

University of Arkansas
P.O. Box 140
Hope, AR 71802

Headquarters:
1128 Main Street, 4th Floor
Cincinnati, Ohio 45210

Ambiguous patterns, needing context and many sources of evidence

Person names

…was among the six houses sold by Hope Feldman that year.

Pawel Opalinski, Software Engineer at WhizBang Labs.

Landscape of IE Tasks (4/4): Pattern Combinations

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

Single entity

Person: **Jack Welch**

Person: **Jeffrey Immelt**

Location: **Connecticut**

Binary relationship

Relation: **Person-Title**
Person: **Jack Welch**
Title: **CEO**

Relation: **Company-Location**
Company: **General Electric**
Location: **Connecticut**

N-ary record

Relation: **Succession**
Company: **General Electric**
Title: **CEO**
Out: **Jack Welsh**
In: **Jeffrey Immelt**

"Named entity" extraction

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.

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

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

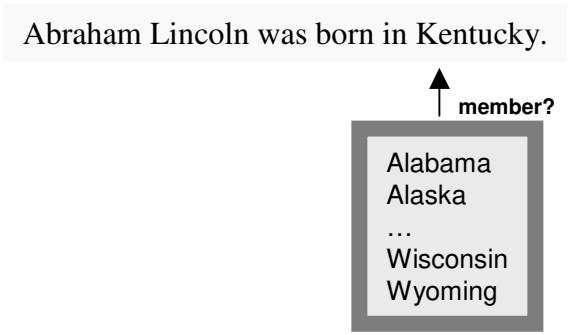
$$\text{F1} = \text{Harmonic mean of Precision \& Recall} = \frac{1}{((1/P) + (1/R)) / 2}$$

State of the Art Performance

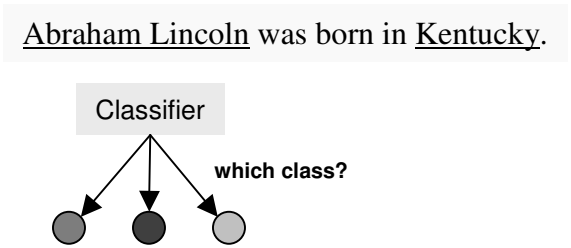
- Named entity recognition
  - Person, Location, Organization, …
  - F1 in high 80's or low- to mid-90's
- Binary relation extraction
  - Contained-in (Location1, Location2)
    Member-of (Person1, Organization1)
  - F1 in 60's or 70's or 80's
- Wrapper induction
  - Extremely accurate performance obtainable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

Landscape of IE Techniques (1/1): Models

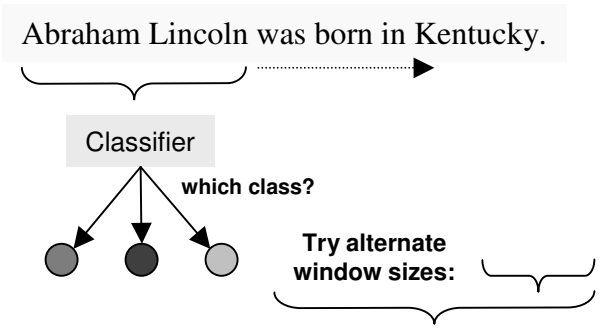
Lexicons



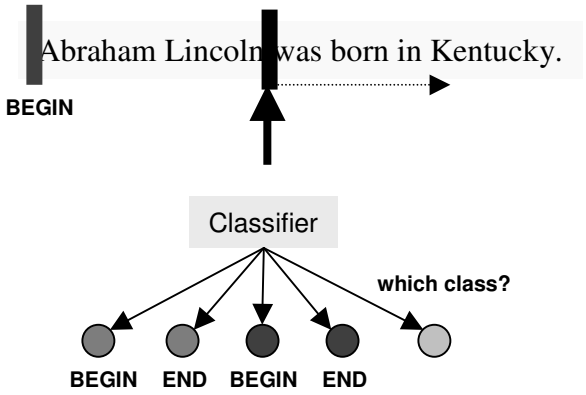
Classify Pre-segmented Candidates



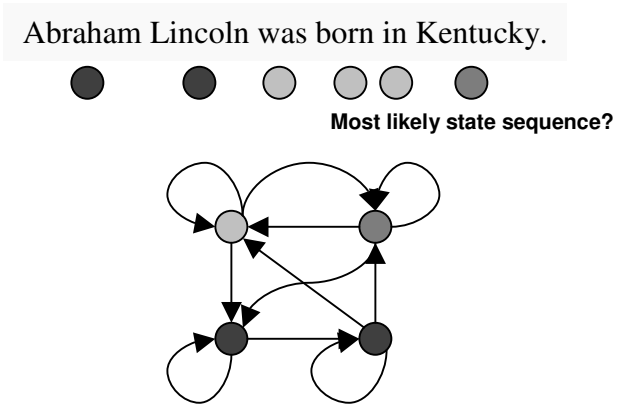
Sliding Window



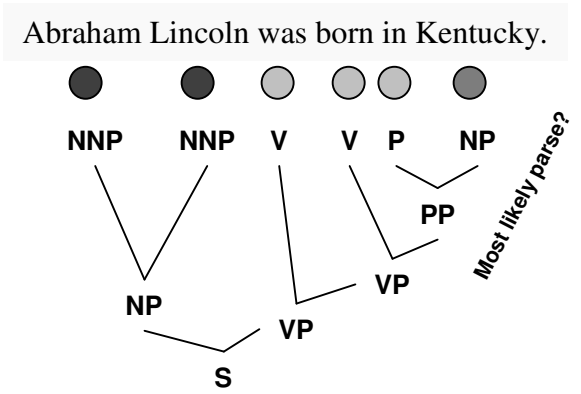
Boundary Models



Finite State Machines



Context Free Grammars



…and beyond

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

Landscape:
Focus of this Tutorial

Pattern complexity	closed set · regular · complex · ambiguous
Pattern feature domain	words · words + formatting · formatting
Pattern scope	site-specific · genre-specific · general
Pattern combinations	entity · binary · n-ary
Models	lexicon · regex · window · boundary · FSM · CFG

Sliding Windows

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

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

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

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

A "Naïve Bayes" Sliding Window Model

[Freitag 1997]

… **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**

P("Wean Hall Rm 5409" = LOCATION) =

$$P(\text{bin}(t) \mid \theta_{\text{start}}) P(n \mid \theta_{\text{length}}) \prod_{i=t-m}^{t-1} P(w_i \mid \theta_{\text{prefix},i-t}) \prod_{i=t}^{t+n} P(w_i \mid \theta_{\text{contents}}) \prod_{i=t+n+1}^{t+n+m} P(w_i \mid \theta_{\text{suffix},i-t-n})$$

Prior probability of start position | **Prior probability of length** | **Probability prefix words** | **Probability contents words** | **Probability suffix words**

Try all start positions and reasonable lengths

Estimate these probabilities by (smoothed) counts from labeled training data.

If P("Wean Hall Rm 5409" = LOCATION) **is above some threshold, extract it.**

Other examples of sliding window: ***[Baluja et al 2000]***
(decision tree over individual words & their context)

“Naïve Bayes” Sliding Window Results

Domain: CMU UseNet Seminar Announcements

```
    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.
```

Field	F1
Person Name:	**30%**
Location:	**61%**
Start Time:	**98%**

SRV: a realistic sliding-window-classifier IE system

[Frietag AAAI '98]

- What windows to consider?
  - *all* windows containing **as many** tokens as the shortest example, but **no more** tokens than the longest example
- How to represent a classifier? It might:
  - Restrict the **length** of window;
  - Restrict the **vocabulary** or formatting used before/after/inside window;
  - Restrict the **relative order** of tokens;
  - Etc…

<title>Course Information for CS213</title>

<h1>CS 213 C++ Programming</h1>

"A token followed by a 3-char numeric token just after the title"

SRV: a rule-learner for sliding-window classification

- Top-down rule learning:

let RULES = ∅;
while (*there are uncovered positive examples*) {
  // construct a rule R to add to RULES
  let R be a rule covering all examples;
  while (R *covers too many negative examples)* {
    let C = argmax_C VALUE(R, R ∧ C, *uncoveredExamples*)
      over some set of candidate conditions C
    let R = R ∧ C;
  }
  let RULES = RULES ∪ {R};
}

SRV: a rule-learner for sliding-window classification

Search metric: SRV algorithm greedily adds conditions to maximize "information gain" of *R*

$$\text{VALUE}(R,R',\text{Data}) = |\text{Data}|*p\,(\,p \log p - p' \log p')$$

where *p (p')* is fraction of data covered by *R (R')*

To prevent overfitting:

rules are built on 2/3 of data, then their false positive rate is estimated with a Dirichlet on the 1/3 holdout set.

Candidate conditions: …

Learning “first-order” rules

- A sample “zero-th” order rule set:

  (tok1InTitle ∧ ¬tok1StartsPara ∧ tok2triple)
  ∨ (prevtok2EqCourse ∧ prevtok1EqNumber) ∨ **...**

- First-order “rules” can be learned the same way—with additional search to find best “condition”

  phrase(X) ← firstToken(X,A), ¬startPara(A),
  nextToken(A,B), triple(B)
  phrase(X) ← firstToken(X,A), prevToken(A,C), eq(C,’number’),
  prevToken(C,D), eq(D,’course’)

- Semantics:

  “p(X) ← q(X),r(X,Y),s(Y)” = “{X : ∃ Y : q(X) ∧ r(X,Y) ∧ s(Y)}”

SRV: a rule-learner for sliding-window classification

- Primitive predicates used by SRV:
  - *token(X,W), allLowerCase(W), numerical(W), …*
  - *nextToken(W,U), previousToken(W,V)*

- HTML-specific predicates:
  - *inTitleTag(W), inH1Tag(W), inEmTag(W),…*
  - *emphasized(W) = "inEmTag(W) or inBTag(W) or …"*
  - *tableNextCol(W,U) = "U is some token in the column after the column W is in"*
  - *tablePreviousCol(W,V), tableRowHeader(W,T),…*

SRV: a rule-learner for sliding-window classification

- Non-primitive "conditions" used by SRV:
  - *every(+X, f, c) = $\forall$ W$\in$X : f(W)=c*
    - variables tagged "+" must be used in earlier conditions
    - underlined values will be replaced by constants, e.g., "every(X, isCapitalized, true)"
  - *some(+X, W, <f_1,…,f_k>, g, c)= $\exists$ W: g(f_k(…(f_1(W)…))=c*
    - *e.g., some(X, W, [prevTok,prevTok],inTitle,false)*
    - *set of "paths" <f_1,…,f_k> considered grows over time.*
  - *tokenLength(+X, relop, c):*
  - *position(+W,direction,relop, c):*
    - *e.g., tokenLength(X,>,4), position(W,fromEnd,<,2)*

Utility of non-primitive conditions in greedy rule search

Greedy search for first-order rules is hard because useful conditions can give **no** immediate benefit:

phrase(X) ← token(X,A), prevToken(A,B),inTitle(B),
nextToken(A,C), tripleton(C)

"A token followed by a 3-char numeric token just after the title"

<title>Course Information for CS213</title>

<h1>CS 213 C++ Programming</h1>

courseNumber(X) ←
tokenLength(X,=,2),
every(X, inTitle, false),
some(X, A, <previousToken>, inTitle, true),
some(X, B, <>. tripleton, true)

Non-primitive conditions make greedy search easier

Rapier: an alternative approach

[Califf & Mooney, AAAI '99]

A bottom-up rule learner:

initialize RULES to be one rule per example;
repeat {
  randomly pick *N* pairs of rules (R_i,R_j);
  let {$G_{1...}$,G_N} be the consistent pairwise generalizations;
  let $G^* = argmin_G\ COST(G,RULES)$;
  let RULES = RULES ∪ {G*} – {R': G* ⊇ R'}
}
where *COST(G,RULES) = size of RULES*- {R': G ⊇ R'} and
  "G⊇ R" means every *example* matching G matches R

<title>Course Information for CS213</title>

<h1>CS 213 C++ Programming</h1> …

Differences dropped

courseNum(window1) ← token(window1,'CS'), doubleton('CS'),
prevToken('CS','CS213'), inTitle('CS213'), nextTok('CS','213'),
numeric('213'), tripleton('213'), nextTok('213','C++'),
tripleton('C++'), ….

<title>Syllabus and meeting times for Eng 214</title>

<h1>Eng 214 Software Engineering for Non-programmers </h1>…

courseNum(window2) ← token(window2,'Eng'), tripleton('Eng'),
prevToken('Eng','214'), inTitle('214'), nextTok('Eng','214'),
numeric('214'), tripleton('214'), nextTok('214','Software'), …

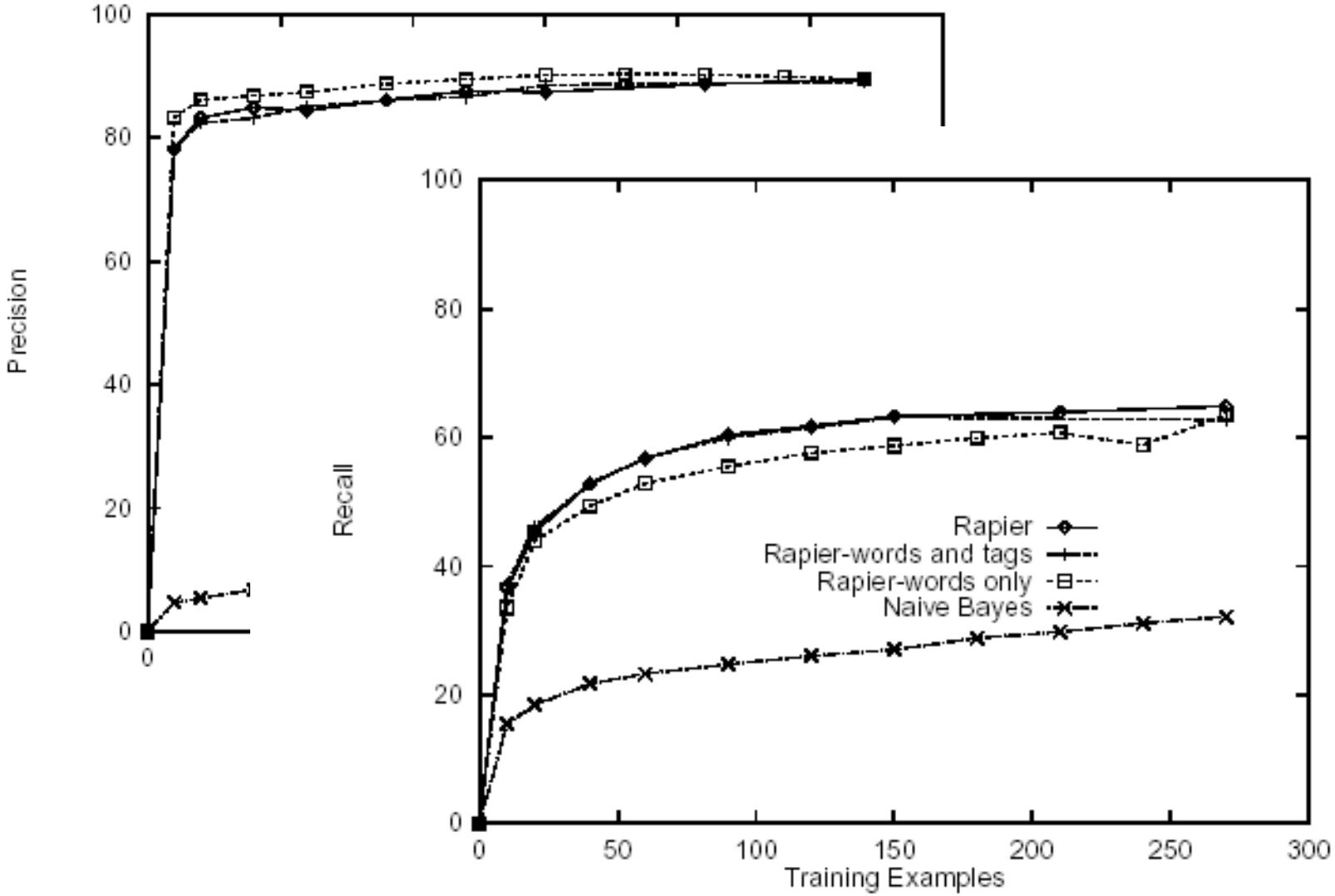
courseNum(X) ← token(X,A),
prevToken(A, B), inTitle(B), nextTok(A,C)),
numeric(C), tripleton(C), nextTok(C,D), …

Common conditions carried over to generalization

Rapier: an alternative approach

- Combines top-down and bottom-up learning
  - Bottom-up to find common restrictions on **content**
  - Top-down greedy addition of restrictions on **context**
- Use of part-of-speech and semantic features (from WORDNET).
- Special “pattern-language” based on sequences of tokens, each of which satisfies one of a set of given constraints
  - **< <tok∈{‘ate’,’hit’},POS∈{‘vb’}>, <tok∈{‘the’}>, <POS∈{‘nn’>>**

Rapier: results – precision/recall



Rapier – results *vs.* SRV

System	stime		etime		loc		speaker	
	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
RAPIER	93.9	92.9	95.8	94.6	91.0	60.5	80.9	39.4
RAP-WT	96.5	95.3	94.9	94.4	91.0	61.5	79.0	40.0
RAP-W	96.5	95.9	96.8	96.6	90.0	54.8	76.9	29.1
NAIBAY	98.2	98.2	49.5	95.7	57.3	58.8	34.5	25.6
SRV	98.6	98.4	67.3	92.6	74.5	70.1	54.4	58.4
WHISK	86.2	100.0	85.0	87.2	83.6	55.4	52.6	11.1
WH-PR	96.2	100.0	89.5	87.2	93.8	36.1	0.0	0.0

Rule-learning approaches to sliding-window classification: Summary

- SRV, Rapier, and WHISK *[Soderland KDD '97]*
  - Representations for **classifiers** allow restriction of the **relationships** between tokens, etc
  - Representations are carefully chosen **subsets** of **even more powerful** representations based on logic programming (ILP and Prolog)
  - Use of these "heavyweight" representations is **complicated**, but seems to pay off in results
- Can simpler representations for classifiers work?

BWI: Learning to detect boundaries

[Freitag & Kushmerick, AAAI 2000]

- Another formulation: learn **three** probabilistic classifiers:
  - *START(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

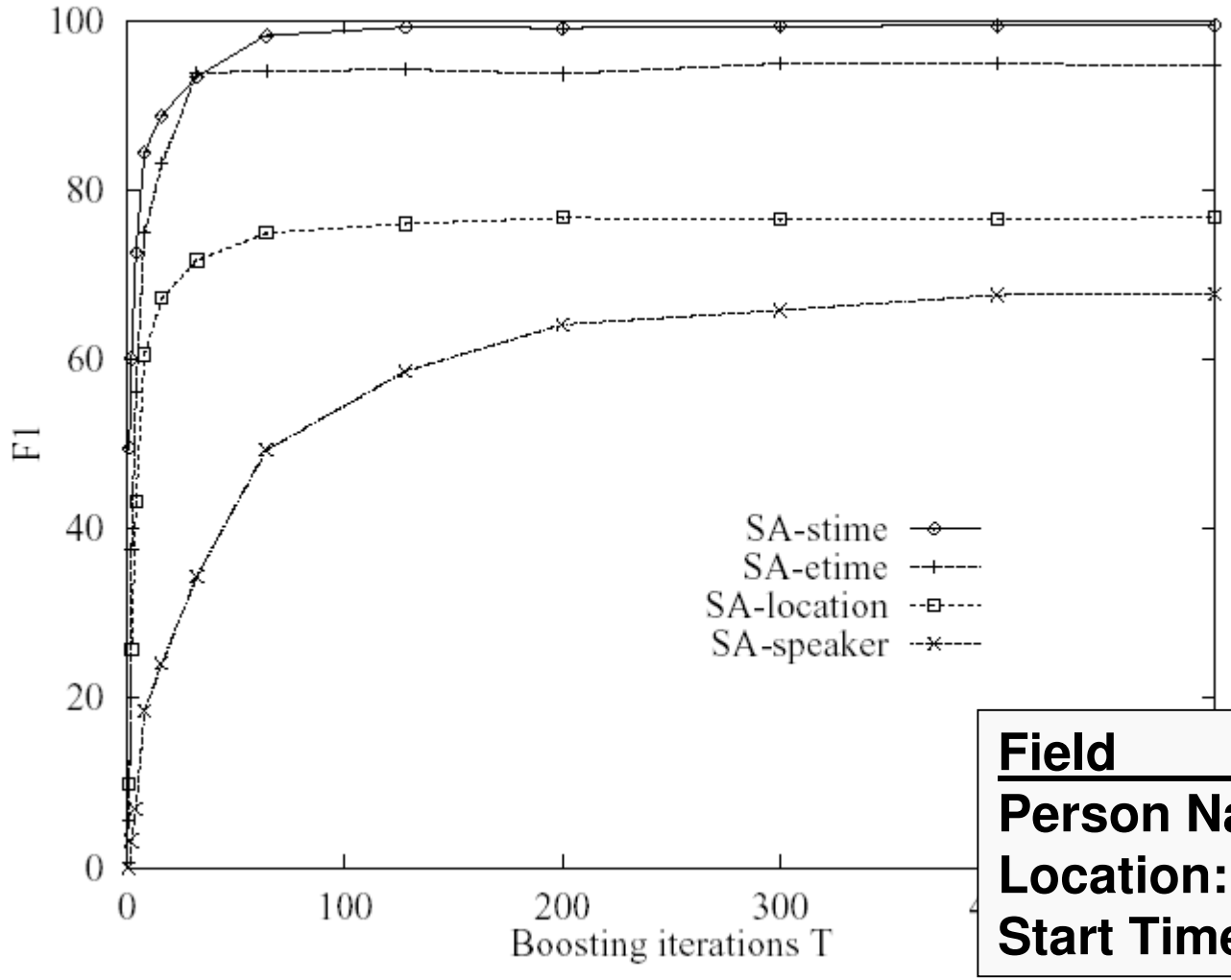
  *START(i) * END(j) * LEN(j-i)*

- *LEN(k)* is estimated from a histogram

BWI: Learning to detect boundaries

- BWI uses **boosting** to find "detectors" for *START* and *END*
- Each weak detector has a *BEFORE* and *AFTER* pattern (on tokens before/after position *i).*
- Each "pattern" is a sequence of tokens and/or wildcards like: anyAlphabeticToken, anyToken, anyUpperCaseLetter, anyNumber, …
- Weak learner for "patterns" uses greedy search (+ lookahead) to repeatedly extend a pair of empty *BEFORE,AFTER* patterns

BWI: Learning to detect boundaries



Field	F1
Person Name:	30%
Location:	61%
Start Time:	98%

Problems with Sliding Windows and Boundary Finders

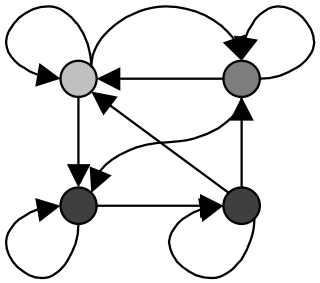
- Decisions in neighboring parts of the input are made independently from each other.
  - Naïve Bayes Sliding Window may predict a "seminar end time" before the "seminar start time".
  - It is possible for two *overlapping* windows to both be above threshold.
  - In a Boundary-Finding system, left boundaries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

Finite State Machines

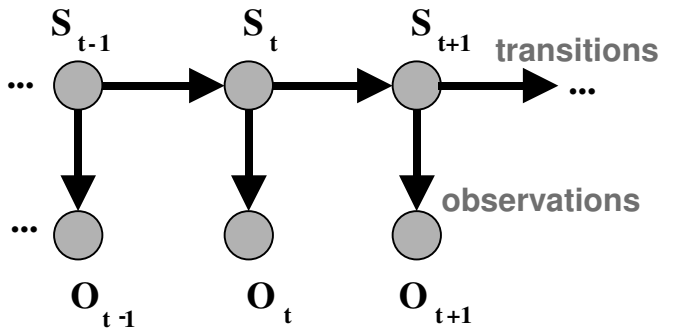
Hidden Markov Models

HMMs are the standard sequence modeling tool in genomics, music, speech, NLP, …

Finite state model



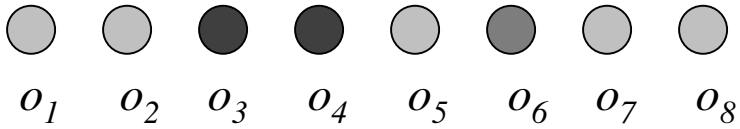
Graphical model



Generates:

State sequence

Observation sequence



$$P(\vec{s},\vec{o}) \propto \prod_{t=1}^{|\vec{o}|} P(s_t \mid s_{t-1}) P(o_t \mid s_t)$$

Parameters: for all states $S=\{s_1,s_2,\ldots\}$

Start state probabilities: $P(s_t)$

Transition probabilities: $P(s_t|s_{t-1})$

Observation (emission) probabilities: $P(o_t|s_t)$ Usually a multinomial over atomic, fixed alphabet

Training:

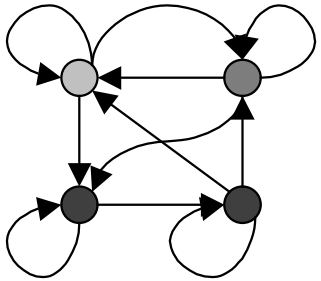
Maximize probability of training observations (w/ prior)

IE with Hidden Markov Models

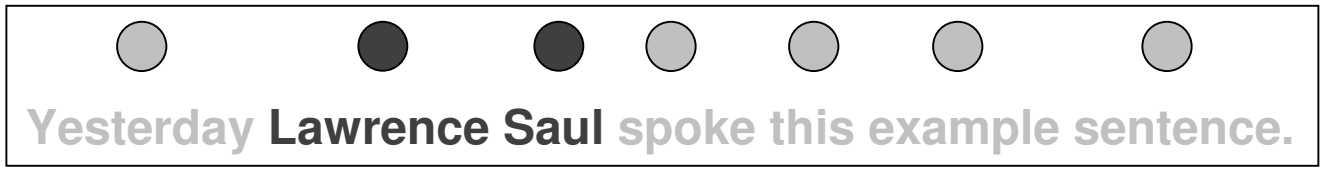
Given a sequence of observations:

Yesterday Lawrence Saul spoke this example sentence.

and a trained HMM:



Find the most likely state sequence: (Viterbi) $\arg\max_{\vec{s}} P(\vec{s}, \vec{o})$



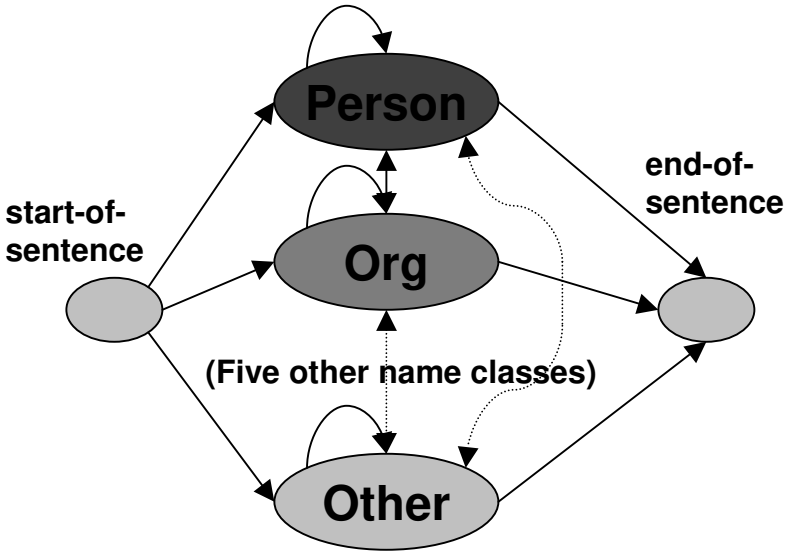
Any words said to be generated by the designated "person name" state extract as a person name:

Person name: Lawrence Saul

HMM Example: "Nymble"

[Bikel, et al 1998], [BBN "IdentiFinder"]

Task: Named Entity Extraction



Transition probabilities

$$P(s_t \mid s_{t-1}, o_{t-1})$$

Back-off to:

$$P(s_t \mid s_{t-1})$$

$$P(s_t)$$

Observation probabilities

$$P(o_t \mid s_t, s_{t-1})$$

or $$P(o_t \mid s_t, o_{t-1})$$

Back-off to:

$$P(o_t \mid s_t)$$

$$P(o_t)$$

Train on 450k words of news wire text.

Results:

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

Other examples of shrinkage for HMMs in IE: *[Freitag and McCallum '99]*

Regrets from Atomic View of Tokens

Would like richer representation of text:
multiple overlapping features, whole chunks of text.

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"

line, sentence, or paragraph features:

- length
- is centered in page
- percent of non-alphabetics
- white-space aligns with next line
- containing sentence has two verbs
- grammatically contains a question
- contains links to "authoritative" pages
- *emissions that are uncountable*
- *features at multiple levels of granularity*

Problems with Richer Representation and a Generative Model

- These arbitrary features are not independent:
  - Overlapping and long-distance dependences
  - Multiple levels of granularity (words, characters)
  - Multiple modalities (words, formatting, layout)
  - Observations from past and future
- HMMs are *generative* models of the text: $P(\vec{s}, \vec{o})$
- Generative models do not easily handle these non-independent features. Two choices:
  - **Model the dependencies**. Each state would have its own Bayes Net. But we are already starved for training data!
  - **Ignore the dependencies**. This causes "over-counting" of evidence (ala naïve Bayes). Big problem when combining evidence, as in Viterbi!

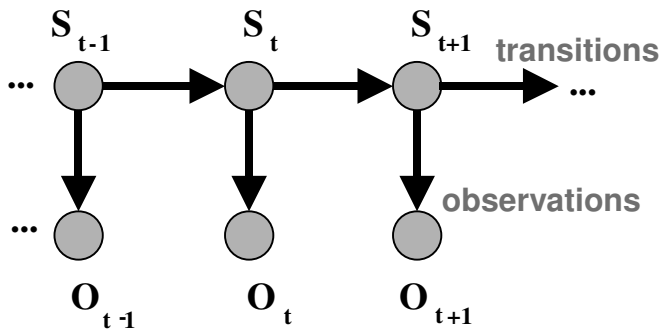
Conditional Sequence Models

- We would prefer a *conditional* model: $P(\overline{s}|\overline{o})$ instead of $P(\overline{s},\overline{o})$:
  - Can examine features, but not responsible for generating them.
  - Don't have to explicitly model their dependencies.
  - Don't "waste modeling effort" trying to generate what we are given at test time anyway.
- If successful, this answers the challenge of integrating the ability to handle many arbitrary features with the full power of finite state automata.

Locally Normalized Conditional Sequence Model

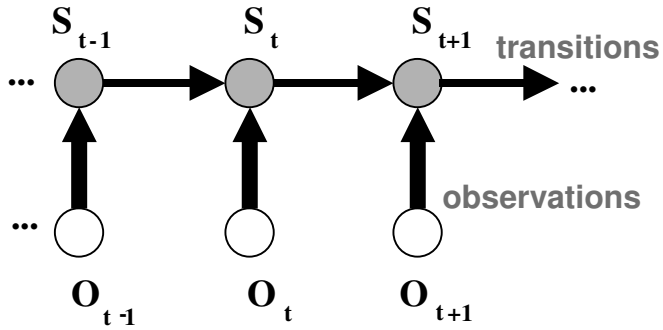
Maximum Entropy Markov Models ***[McCallum, Freitag & Pereira, 2000]***
MaxEnt POS Tagger ***[Ratnaparkhi, 1996]***
SNoW-based Markov Model ***[Punyakanok & Roth, 2000]***

Generative (traditional HMM)



$$P(\vec{s},\vec{o}) \propto \prod_{t=1}^{|\vec{o}|} P(s_t \mid s_{t-1}) P(o_t \mid s_t)$$

Conditional



$$P(\vec{s} \mid \vec{o}) \propto \prod_{t=1}^{|\vec{o}|} P(s_t \mid s_{t-1}, o_t)$$

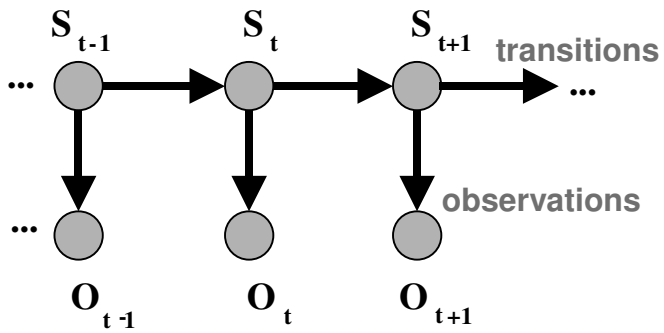
Standard belief propagation: forward-backward procedure.
Viterbi and Baum-Welch follow naturally.

Locally Normalized Conditional Sequence Model

Maximum Entropy Markov Models ***[McCallum, Freitag & Pereira, 2000]***
MaxEnt POS Tagger ***[Ratnaparkhi, 1996]***
SNoW-based Markov Model ***[Punyakanok & Roth, 2000]***

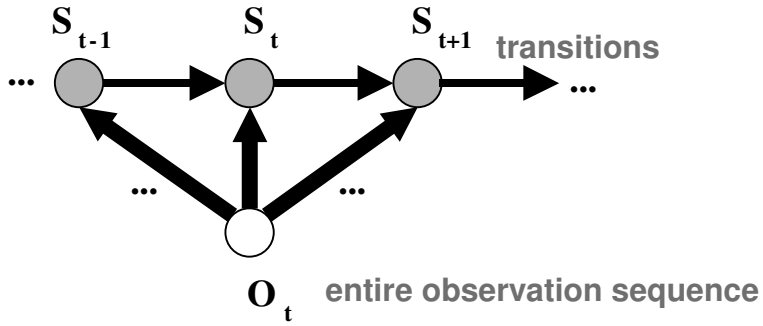
Or, more generally:

Generative (traditional HMM)



$$P(\vec{s},\vec{o}) \propto \prod_{t=1}^{|\vec{o}|} P(s_t \mid s_{t-1}) P(o_t \mid s_t)$$

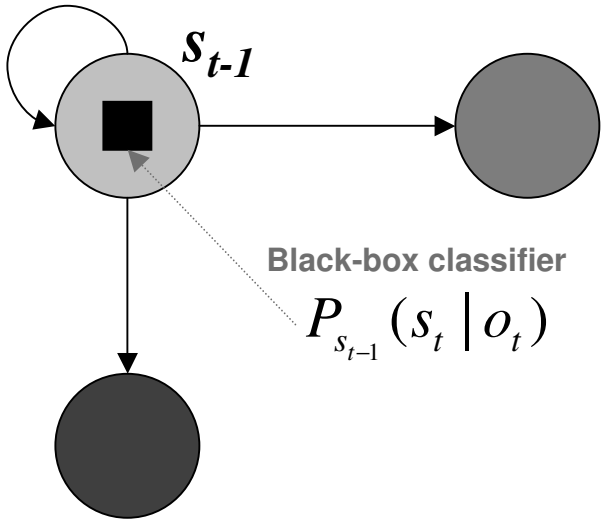
Conditional



$$P(\vec{s} \mid \vec{o}) \propto \prod_{t=1}^{|\vec{o}|} P(s_t \mid s_{t-1}, \vec{o}, t)$$

Standard belief propagation: forward-backward procedure.
Viterbi and Baum-Welch follow naturally.

Exponential Form for "Next State" Function



$$P(s_t \mid s_{t-1}, \vec{o}, t) \Rightarrow$$

$$P_{s_{t-1}}(s_t \mid \vec{o}, t) = \frac{1}{Z(s_{t-1}, \vec{o}, t)} \exp\left(\sum_k \underbrace{\lambda_k}_{\text{weight}} \underbrace{f_k(s_t, s_{t-1}, \vec{o}, t)}_{\text{feature}} \right)$$

Overall Recipe:

- Labeled data is assigned to transitions.
- Train each state's exponential model by maximum likelihood (iterative scaling or conjugate gradient).

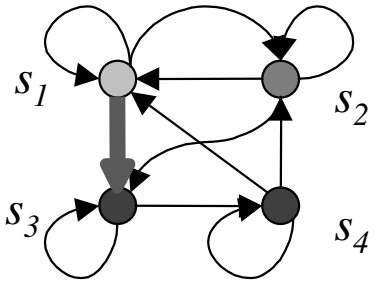
Feature Functions

Example $f_k(\vec{o}, t, s_t, s_{t-1})$:

$$f_{<\text{Capitalized},s_i,s_j>}(s_t, s_{t-1}, \vec{o}, t) = \begin{cases} 1 & \text{if Capitalized}(o_t) \wedge \text{s}_\text{i} = s_{t-1} \wedge \text{s}_\text{j} = s_t \\ 0 & \text{otherwise} \end{cases}$$

$\overline{o}$ = **Yesterday Lawrence Saul spoke this example sentence.**

o_1 o_2 o_3 o_4 o_5 o_6 o_7



$$f_{<Capitalized,s_1,s_3>}(s_1, s_2, \vec{o}, 2) = 1$$

Experimental Data

38 files belonging to 7 UseNet FAQs

Example:

```
<head>          X-NNTP-Poster: NewsHound v1.33
<head>          Archive-name: acorn/faq/part2
<head>          Frequency: monthly
<head>
<question>      2.6) What configuration of serial cable should I use?
<answer>
<answer>            Here follows a diagram of the necessary connection
<answer>        programs to work properly.  They are as far as I know
<answer>        agreed upon by commercial comms software developers fo
<answer>
<answer>           Pins 1, 4, and 8 must be connected together inside
<answer>        is to avoid the well known serial port chip bugs.  The
```

Procedure: For each FAQ, train on one file, test on other; average.

Features in Experiments

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
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

Models Tested

- **ME-Stateless**: A single maximum entropy classifier applied to each line independently.
- **TokenHMM**: A fully-connected HMM with four states, one for each of the line categories, each of which generates individual tokens (groups of alphanumeric characters and individual punctuation characters).
- **FeatureHMM**: Identical to TokenHMM, only the lines in a document are first converted to sequences of features.
- **MEMM**: The Maximum Entropy Markov Model described in this talk.

Results

Learner	*Segmentation precision*	*Segmentation recall*
ME-Stateless	0.038	0.362
TokenHMM	0.276	0.140
FeatureHMM	0.413	0.529
MEMM	0.867	0.681

Conditional Random Fields (CRFs)

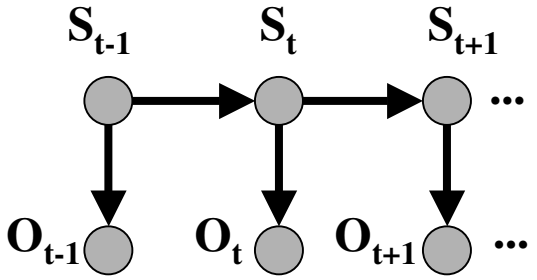
[Lafferty, McCallum, Pereira '2001]

From HMMs to MEMMs to CRFs

$\vec{s} = s_1, s_2, \ldots s_n \qquad \vec{o} = o_1, o_2, \ldots o_n$

HMM

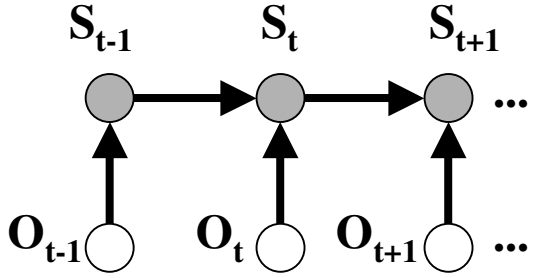
$$P(\vec{s}, \vec{o}) \propto \prod_{t=1}^{|\vec{o}|} P(s_t \mid s_{t-1}) P(o_t \mid s_t)$$



MEMM

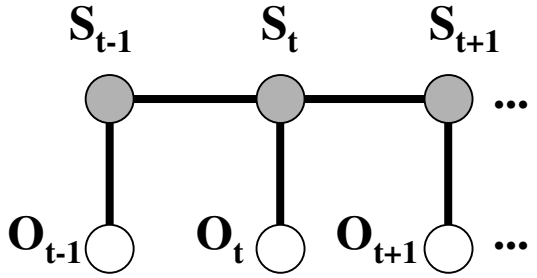
$$P(\vec{s} \mid \vec{o}) \propto \prod_{t=1}^{|\vec{o}|} P(s_t \mid s_{t-1}, o_t)$$

$$\propto \prod_{t=1}^{|\vec{o}|} \frac{1}{Z_{s_{t-1}, o_t}} \exp\left(\sum_j \lambda_j f_j(s_t, s_{t-1}) + \sum_k \mu_k g_k(s_t, o_t) \right)$$



CRF

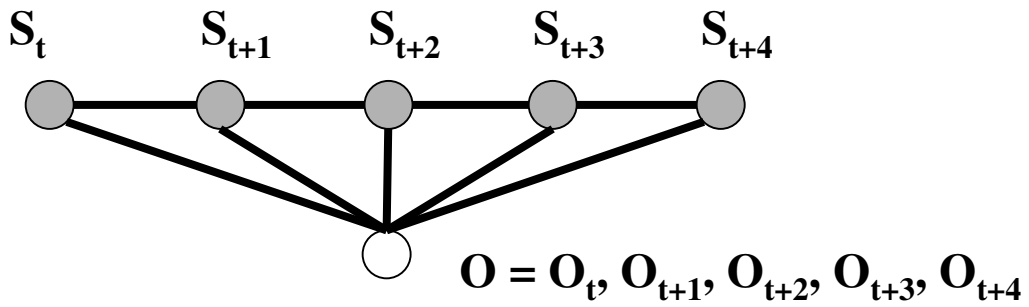
$$P(\vec{s} \mid \vec{o}) \propto \frac{1}{Z_{\vec{o}}} \prod_{t=1}^{|\vec{o}|} \exp\left(\sum_j \lambda_j f_j(s_t, s_{t-1}) + \sum_k \mu_k g_k(s_t, o_t) \right)$$



(A special case of MEMMs and CRFs.)

Conditional Random Fields (CRFs)

[Lafferty, McCallum, Pereira '2001]



Markov on *s*, conditional dependency on *o*.

$$P(\vec{s} \mid \vec{o}) \propto \frac{1}{Z_{\vec{o}}} \prod_{t=1}^{|\vec{o}|} \exp\left(\sum_k \lambda_k f_k(s_t, s_{t-1}, \vec{o}, t) \right)$$

Hammersley-Clifford-Besag theorem stipulates that the CRF has this form—an exponential function of the cliques in the graph.

Assuming that the dependency structure of the states is tree-shaped (linear chain is a trivial tree), inference can be done by dynamic programming in time $O(|\overline{o}|\ |S|^2)$—just like HMMs.

General CRFs vs. HMMs

- More general and expressive modeling technique
- Comparable computational efficiency
- Features may be arbitrary functions of *any* or *all* observations
- Parameters need not fully specify generation of observations; require less training data
- Easy to incorporate domain knowledge
- State means only “state of process”, vs “state of process” and “observational history I’m keeping”

Training CRFs

Maximize log-likelihood of parameters given training data :

$$L(\{\lambda_k\} \mid \{\langle \vec{o}, \vec{s} \rangle^{(i)}\})$$

Log-likelihood gradient :

feature count using correct labels - feature count using labels assigned by current parameters - smoothing penalty

$$\frac{\partial L}{\partial \lambda_k} = \sum_i C_k(\vec{s}^{(i)}, \vec{o}^{(i)}) - \sum_i \sum_{\vec{s}} P_{\{\lambda_k\}}(\vec{s} \mid \vec{o}^{(i)})\, C_k(\vec{s}, \vec{o}^{(i)}) - {\lambda_k}^2$$

$$C_k(\vec{s}, \vec{o}) = \sum_t f_k(\vec{o}, t, s_{t-1}, s_t)$$

Methods:

- **iterative scaling** (quite slow)
- **conjugate gradient** (much faster)
- **conjugate gradient with preconditioning** (super fast)
- **limited-memory quasi-Newton methods** (also super fast)

[Sha & Pereira 2002] & [Malouf 2002]

Complexity comparable to standard Baum-Welch

Voted Perceptron Sequence Models

[Collins 2002]

Like CRFs with stochastic gradient ascent and a Viterbi approximation.

Given training data : $\{\langle \vec{o}, \vec{s} \rangle^{(i)}\}$

Initialize parameters to zero : $\forall \mathrm{k} \;\; \lambda_{\mathrm{k}} = 0$

Iterate to convergence :

for all training instances, i

$$\vec{s}_{Viterbi} = \arg\max_{\vec{s}} \prod_t \exp\left(\sum_k \lambda_k f_k(s_t, s_{t-1}, \vec{o}, t) \right)$$

$$\forall \mathrm{k} : \quad \lambda_k += C_k(\vec{s}^{(i)}, \vec{o}^{(i)}) - C_k(\vec{s}_{Viterbi}, \vec{o}^{(i)})$$

Analogous to the gradient for this one training instance

$$\left(\text{where } C_k(\vec{s}, \vec{o}) = \sum_t f_k(s_{t-1}, s_t, \vec{o}, t) \text{ as before} \right)$$

Avoids calculating the partition function (normalizer), Z_o, but gradient ascent, not 2nd-order or conjugate gradient method.

MEMM & CRF Related Work

- Maximum entropy for language tasks:
  - Language modeling *[Rosenfeld '94, Chen & Rosenfeld '99]*
  - Part-of-speech tagging *[Ratnaparkhi '98]*
  - Segmentation *[Beeferman, Berger & Lafferty '99]*
  - Named entity recognition "MENE" *[Borthwick, Grishman,…'98]*
- HMMs for similar language tasks
  - Part of speech tagging *[Kupiec '92]*
  - Named entity recognition *[Bikel et al '99]*
  - Other Information Extraction *[Leek '97], [Freitag & McCallum '99]*
- Serial Generative/Discriminative Approaches
  - Speech recognition *[Schwartz & Austin '93]*
  - Reranking Parses *[Collins, '00]*
- Other conditional Markov models
  - Non-probabilistic local decision models *[Brill '95], [Roth '98]*
  - Gradient-descent on state path *[LeCun et al '98]*
  - Markov Processes on Curves (MPCs) *[Saul & Rahim '99]*
  - Voted Perceptron-trained FSMs *[Collins '02]*

Part-of-speech Tagging

[Pereira 2001 personal comm.]

45 tags, 1M words training data, Penn Treebank

DT NN NN , NN , VBZ RB JJ IN
The asbestos fiber , crocidolite, is unusually resilient once

PRP VBZ DT NNS , IN RB JJ NNS TO PRP VBG
it enters the lungs , with even brief exposures to it causing

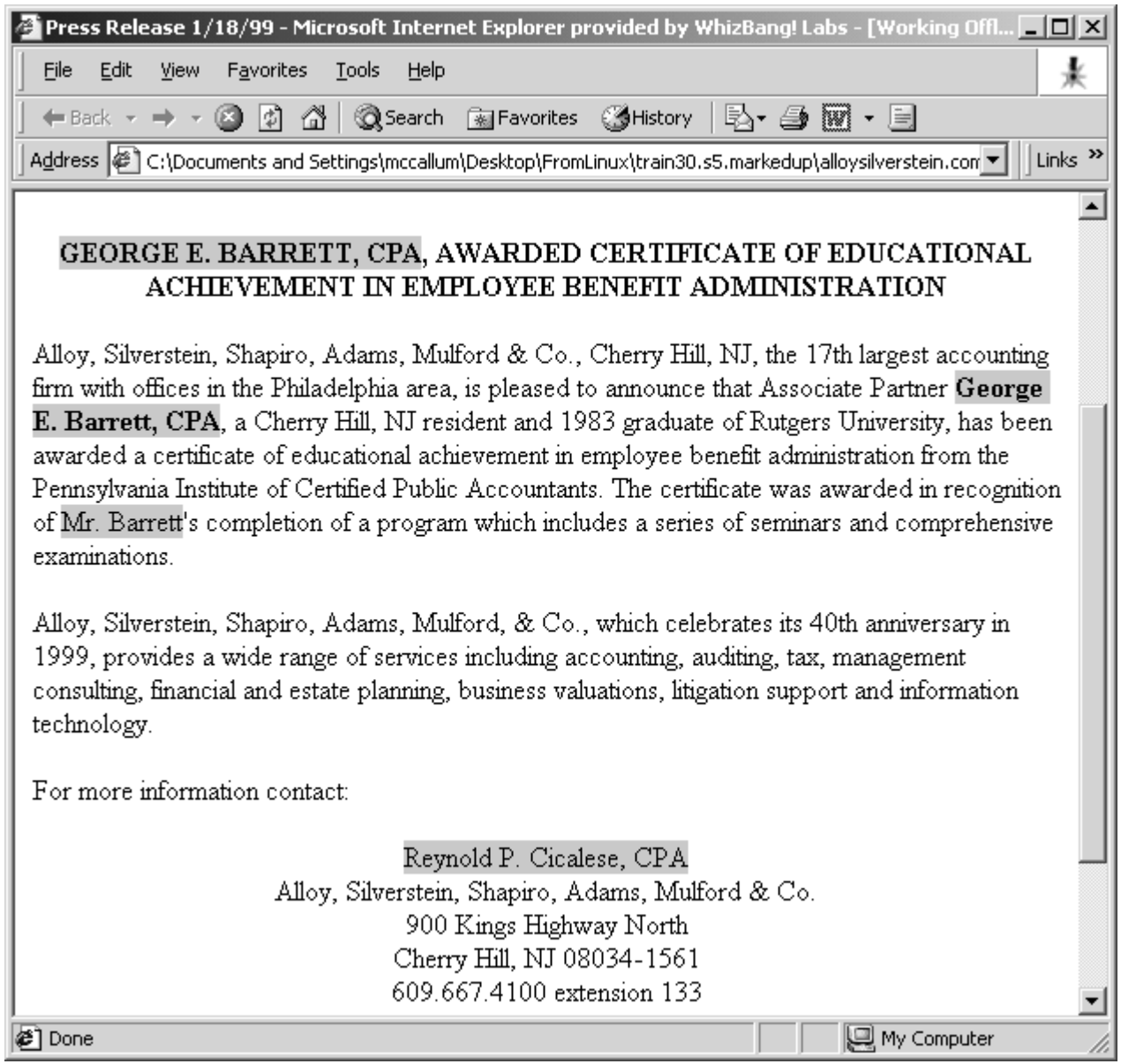
NNS WDT VBP RP NNS JJ , NNS VBD .
symptoms that show up decades later , researchers said .

			Using spelling features*			
	Error	**oov error**	**error**	**Δ err**	**oov error**	**Δ err**
HMM	5.69%	45.99%				
CRF	5.55%	48.05%	4.27%	-24%	23.76%	-50%

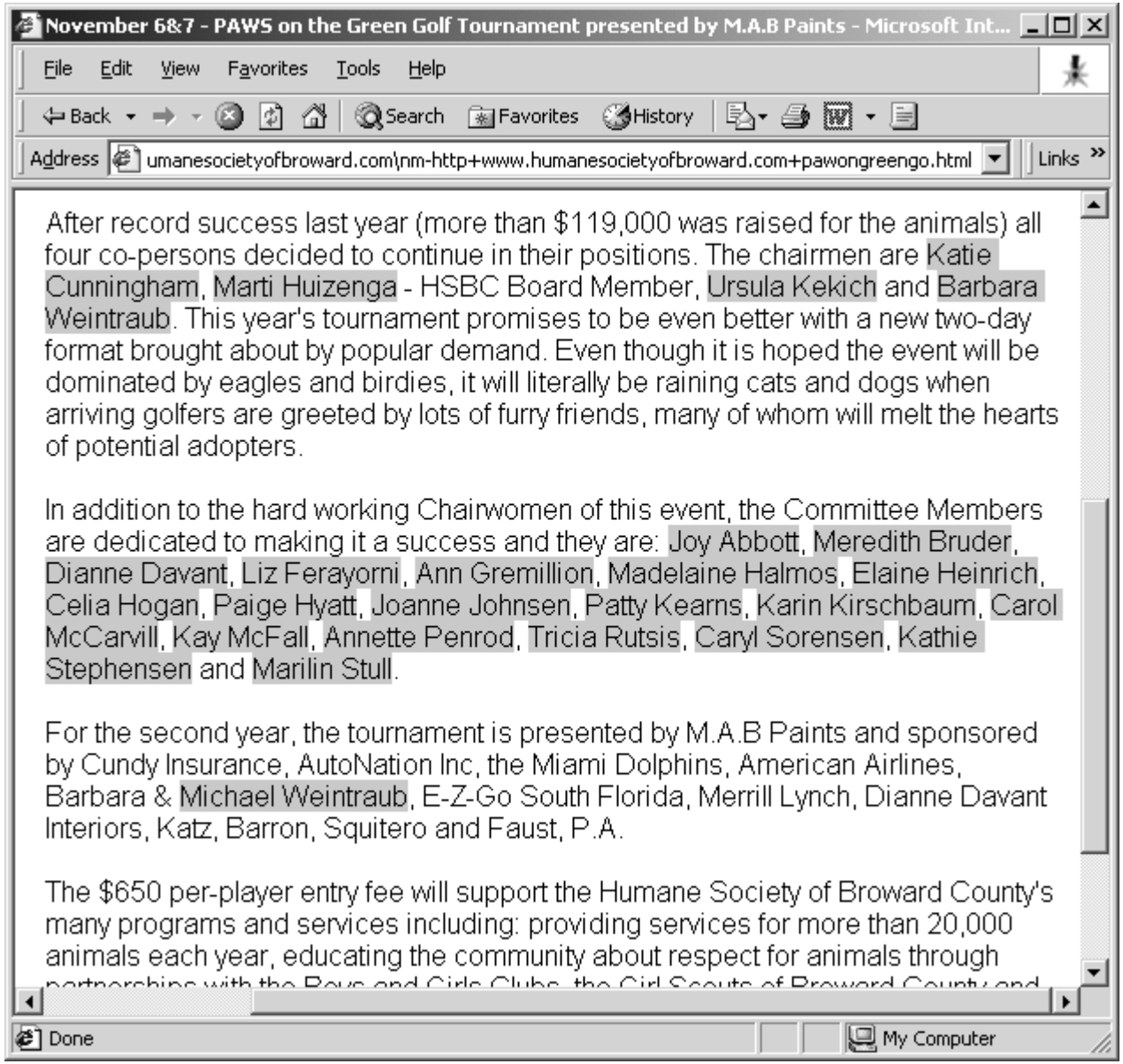
* use words, *plus* overlapping features: capitalized, begins with #, contains hyphen, ends in -ing, -ogy, -ed, -s, -ly, -ion, -tion, -ity, -ies.

Person name Extraction

[McCallum 2001, unpublished]



Person name Extraction



Features in Experiment

Capitalized	Xxxxx
Mixed Caps	XxXxxx
All Caps	XXXXX
Initial Cap	X….
Contains Digit	xxx5
All lowercase	xxxx
Initial	X
Punctuation	.,:;!(), etc
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)

Hand-built FSM person-name extractor says yes, (prec/recall ~ 30/95)

Conjunctions of all previous feature pairs, evaluated at the current time step.

Conjunctions of all previous feature pairs, evaluated at current step and one step ahead.

All previous features, evaluated two steps ahead.

All previous features, evaluated one step behind.

Total number of features = ~200k

Training and Testing

- Trained on 65469 words from 85 pages, 30 different companies' web sites.
- Training takes 4 hours on a 1 GHz Pentium.
- Training precision/recall is 96% / 96%.

- Tested on different set of web pages with similar size characteristics.
- Testing precision is 92 – 95%,
  recall is 89 – 91%.

Chinese Word Segmentation

[McCallum & Feng, to appear]

- Trained on 800 segmented sentences from UPenn Chinese Treebank.
- Training time: ~2 hours with L-BFGS.

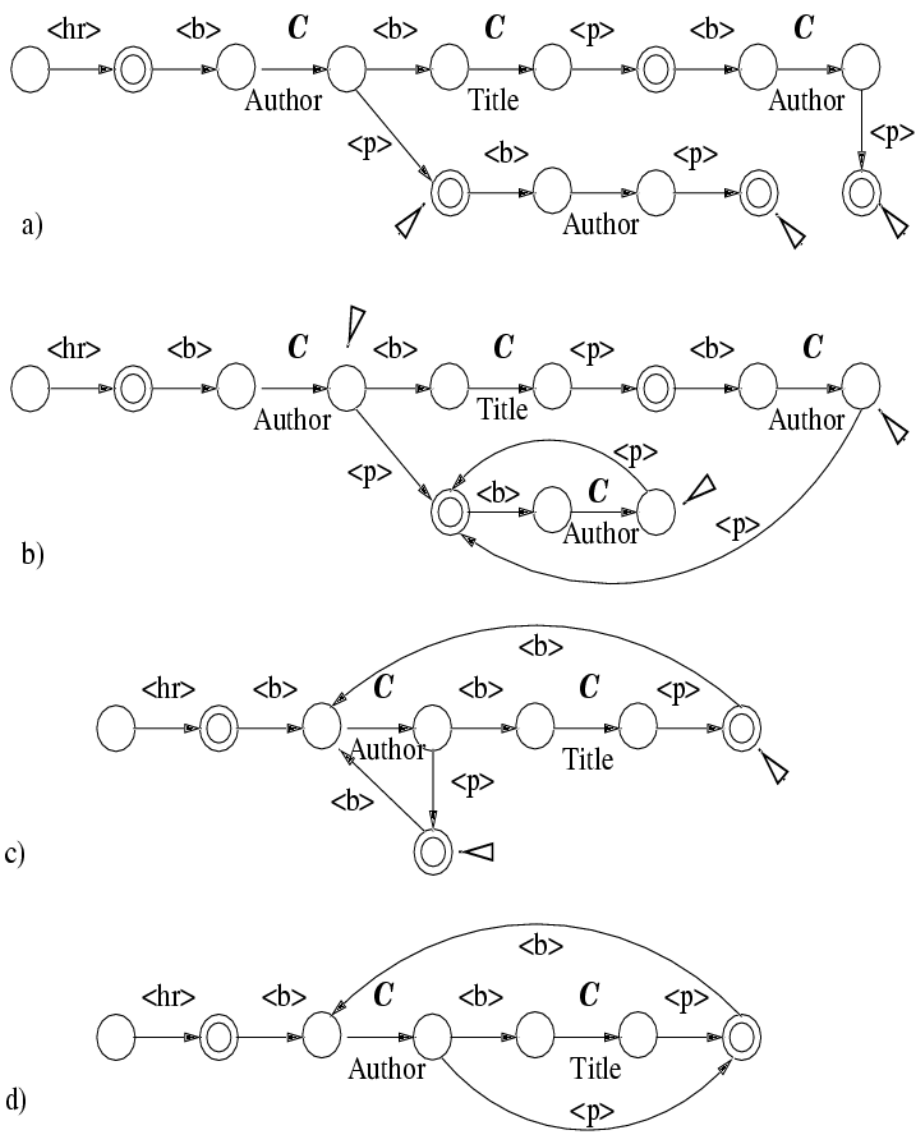
- Training F1: 99.4%
- Testing F1: 99.3%

- Previous top contendors' F1: ~85-95%

Inducing State-Transition Structure

[Chidlovskii, 2000]

K-reversible grammars

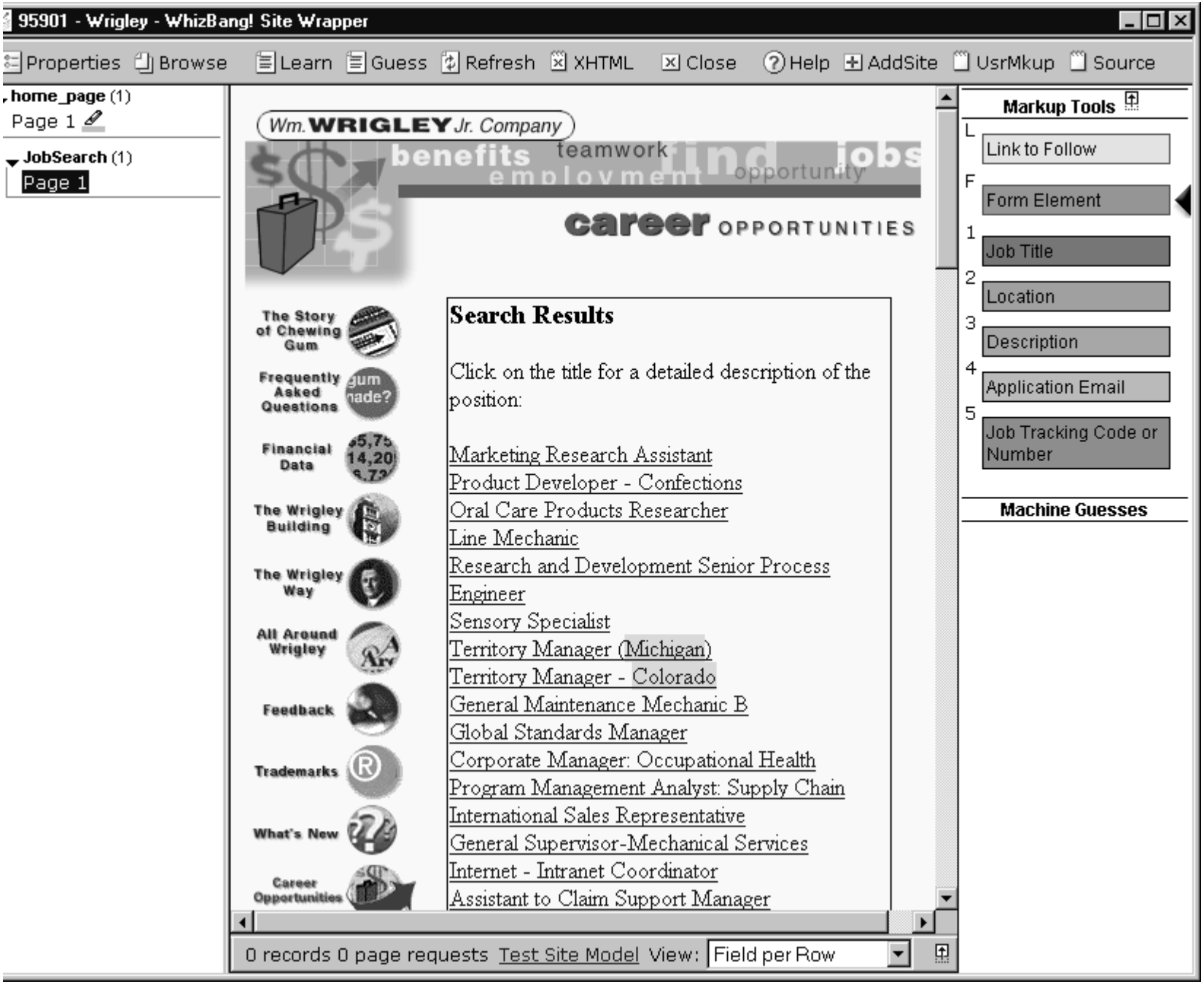


Limitations of HMM/CRF models

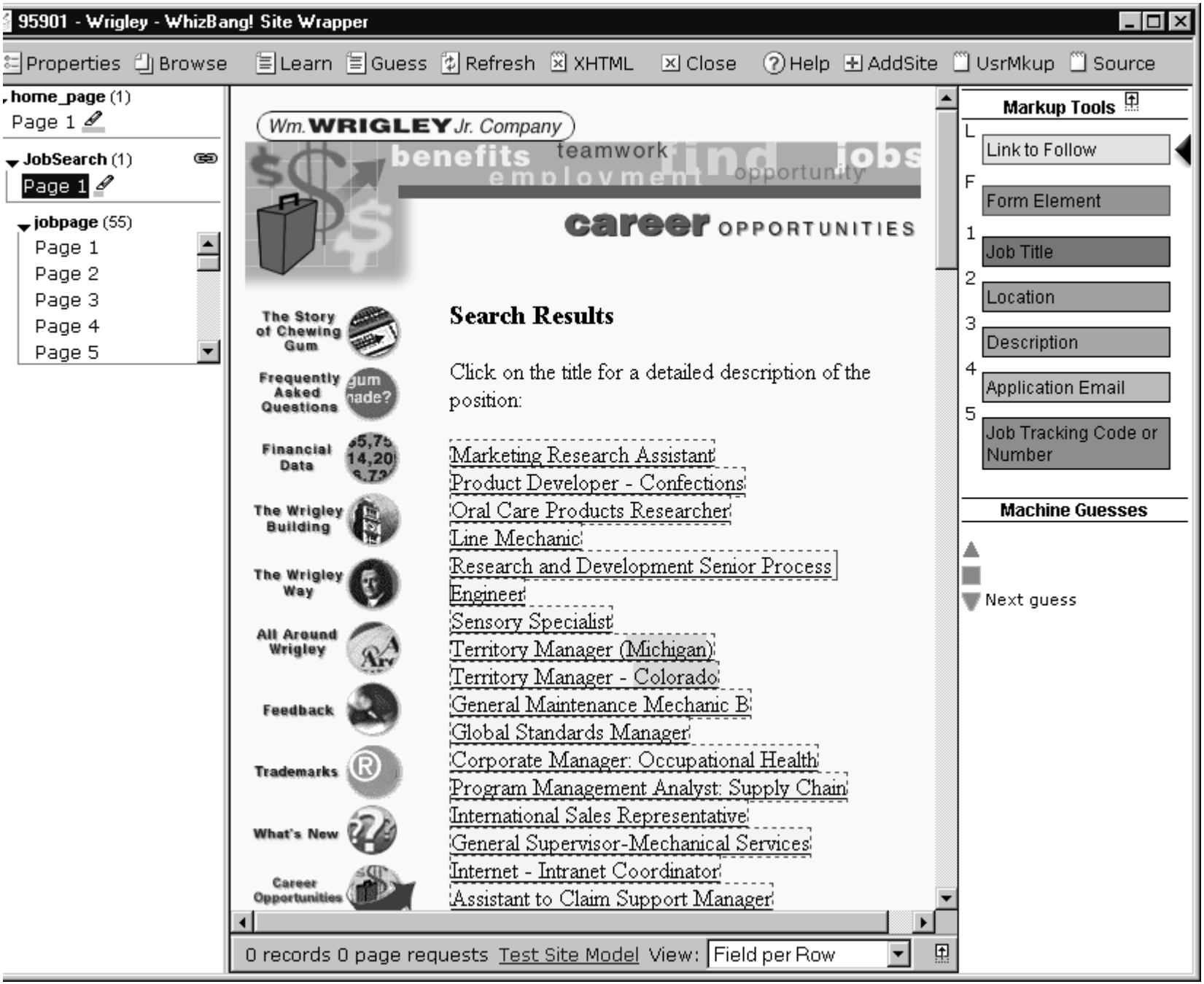
- HMM/CRF models have a **linear** structure
- Web documents have a **hierarchical** structure
  - Are we suffering by not modeling this structure more explicitly?
- How can one learn a **hierarchical** extraction model?
  - Coming up: STALKER, a hierarchical **wrapper-learner**
  - But first: how do we train wrapper-learners?

Tree-based Models

- Extracting from **one** web site
  - Use *site-specific* formatting information: e.g., “the JobTitle is a bold-faced paragraph in column 2”
  - For large well-structured sites, like parsing a **formal language**
- Extracting from **many** web sites:
  - Need general solutions to entity extraction, grouping into records, etc.
  - Primarily use *content* information
  - Must deal with a *wide range* of ways that users present data.
  - Analogous to parsing **natural language**
- Problems are **complementary:**
  - Site-dependent learning can collect training data for a site-independent learner
  - Site-dependent learning can boost accuracy of a site-independent learner on selected key sites


95901 - Wrigley - WhizBang! Site Wrapper
Properties
Browse
Learn
Guess
Refresh
XHTML
Close
Help
AddSite
UsrMkup
Source
home_page (1)
Page 1
JobSearch (1)
Page 1
Wm. WRIGLEY Jr. Company
benefits
teamwork
employment
opportunity
career OPPORTUNITIES
The Story of Chewing Gum
Frequently Asked Questions
Financial Data
The Wrigley Building
The Wrigley Way
All Around Wrigley
Feedback
Trademarks
What's New
Career Opportunities
Search Results
Click on the title for a detailed description of the position:
Marketing Research Assistant
Product Developer - Confections
Oral Care Products Researcher
Line Mechanic
Research and Development Senior Process Engineer
Sensory Specialist
Territory Manager (Michigan)
Territory Manager - Colorado
General Maintenance Mechanic B
Global Standards Manager
Corporate Manager: Occupational Health
Program Management Analyst: Supply Chain
International Sales Representative
General Supervisor-Mechanical Services
Internet - Intranet Coordinator
Assistant to Claim Support Manager
0 records 0 page requests
Test Site Model
View:
Field per Row
Markup Tools
L
Link to Follow
F
Form Element
1
Job Title
2
Location
3
Description
4
Application Email
5
Job Tracking Code or Number
Machine Guesses

95901 - Wrigley - WhizBang! Site Wrapper
Properties
Browse
Learn
Guess
Refresh
XHTML
Close
Help
AddSite
UsrMkup
Source
home_page (1)
Page 1
JobSearch (1)
Page 1
Wm. WRIGLEY Jr. Company
benefits
teamwork
employment
opportunity
jobs
career OPPORTUNITIES
The Story of Chewing Gum
Frequently Asked Questions
Financial Data
The Wrigley Building
The Wrigley Way
All Around Wrigley
Feedback
Trademarks
What's New
Career Opportunities
User gives first K positive—and thus many implicit negative examples
Marketing Research Assistant
Product Developer - Confections
Oral Care Products Researcher
Line Mechanic
Research and Development Senior Process Engineer
Sensory Specialist
Territory Manager (Michigan)
Territory Manager - Colorado
General Maintenance Mechanic B
Global Standards Manager
Corporate Manager: Occupational Health
Program Management Analyst: Supply Chain
International Sales Representative
General Supervisor-Mechanical Services
Internet - Intranet Coordinator
Assistant to Claim Support Manager
0 records 0 page requests
Test Site Model
View:
Field per Row
Markup Tools
L
Link to Follow
F
Form Element
1
Job Title
2
Location
3
Description
4
Application Email
5
Job Tracking Code or Number
Machine Guesses
Learner


95901 - Wrigley - WhizBang! Site Wrapper
Properties
Browse
Learn
Guess
Refresh
XHTML
Close
Help
AddSite
UsrMkup
Source
home_page (1)
Page 1
JobSearch (1)
Page 1
jobpage (55)
Page 1
Page 2
Page 3
Page 4
Page 5
Wm. WRIGLEY Jr. Company
benefits
teamwork
employment
opportunity
career OPPORTUNITIES
The Story of Chewing Gum
Frequently Asked Questions
Financial Data
The Wrigley Building
The Wrigley Way
All Around Wrigley
Feedback
Trademarks
What's New
Career Opportunities
Search Results
Click on the title for a detailed description of the position:
Marketing Research Assistant
Product Developer - Confections
Oral Care Products Researcher
Line Mechanic
Research and Development Senior Process Engineer
Sensory Specialist
Territory Manager (Michigan)
Territory Manager - Colorado
General Maintenance Mechanic B
Global Standards Manager
Corporate Manager: Occupational Health
Program Management Analyst: Supply Chain
International Sales Representative
General Supervisor-Mechanical Services
Internet - Intranet Coordinator
Assistant to Claim Support Manager
0 records 0 page requests
Test Site Model
View:
Field per Row
Markup Tools
L
Link to Follow
F
Form Element
1
Job Title
2
Location
3
Description
4
Application Email
5
Job Tracking Code or Number
Machine Guesses
Next guess

STALKER: Hierarchical boundary finding

[Muslea,Minton & Knoblock 99]

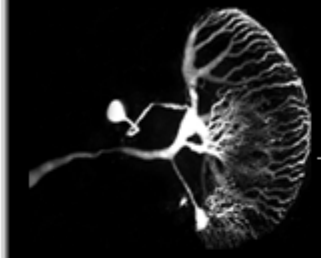
- Main idea:
  - To train a hierarchical extractor, pose a **series** of learning problems, one for each node in the hierarchy
  - At each stage, extraction is simplified by knowing about the "**context**."

Neural Information Processing Systems 2002 (NIPS*2002) Main Page - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media Address Links

Google nips-2002 Search Web Search Site Page Info Up Highlight nips 2002



NIPS 2002 www.nips.cc

Vancouver, Canada, December 9-14

- Main
- Program ■ Tutorials ■ Invited Talks
- Registration and Financial/Travel Support
- Invitation Letters ■ Volunteering
- Hotels and Local Transportation
- For Authors and Presenters

NOTE REGISTER NOW online. (More info).

NOTE Apply for a travel grant before **October 18**. (More info).

NOTE If you are attending the workshops, please help our planning by completing a brief workshop preferences survey.

Neural Information Processing Systems: Natural and Synthetic

You are invited to participate in the Sixteenth Annual Conference on Neural Information Processing Systems, which is the premier scientific meeting on neural computation. The Conference is extraordinarily interdisciplinary, with contributions from many intellectual communities united by an interest in the study of artificial and natural neural information processing. Interdisciplinary interactions are encouraged by having a single track of presentations and high-quality poster sessions. Presentation topics include learning algorithms and architectures, learning theory, neuroscience, cognitive science, vision, speech and signal processing, reinforcement learning and control, implementations,and diverse applications. All papers have been rigorously reviewed (with a 30% acceptance rate). The Conference is preceeded by one day of tutorials and is followed by two days of workshop sessions in nearby Whistler, British Columbia, Canada.

(Click here for a larged Vancouver Image)

Tutorials and Conference

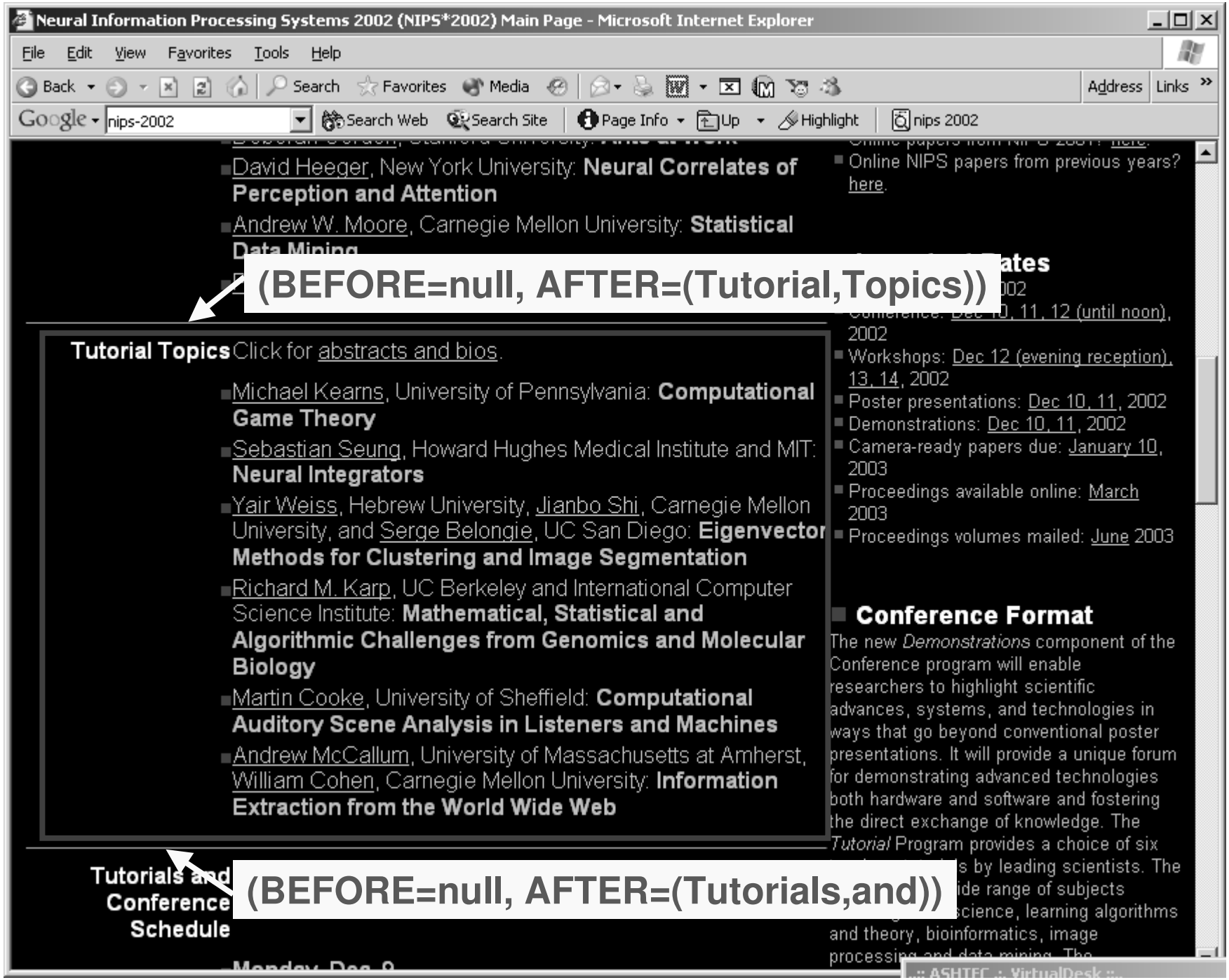
Conference Invited Speakers Click for abstracts and bios.

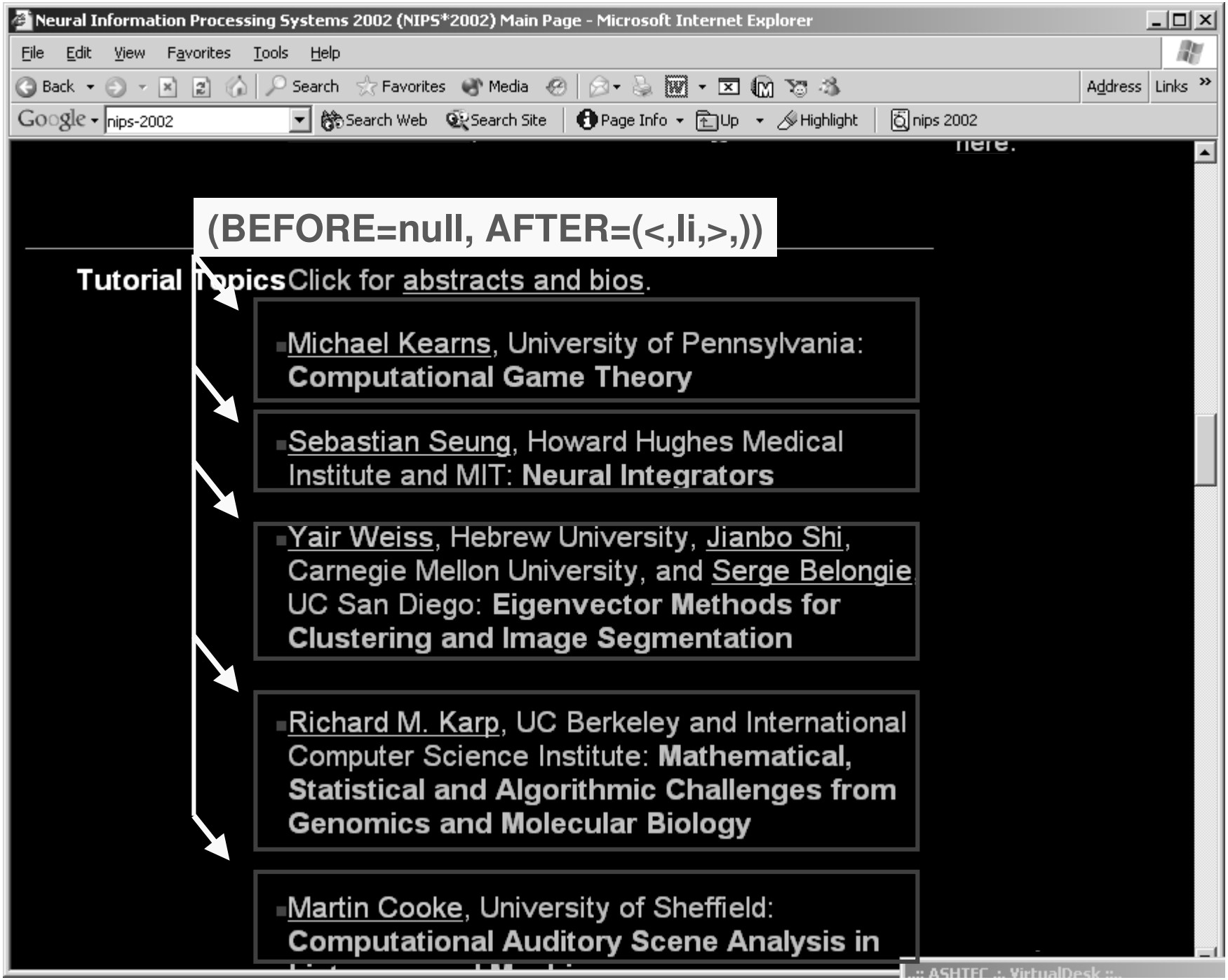
- Hugh Durrant-Whyte, University of Sydney: **Information Flow in Sensor Networks**
- Paul Glimcher, New York University: **Decisions, Uncertainty and the Brain: Neuroeconomics**

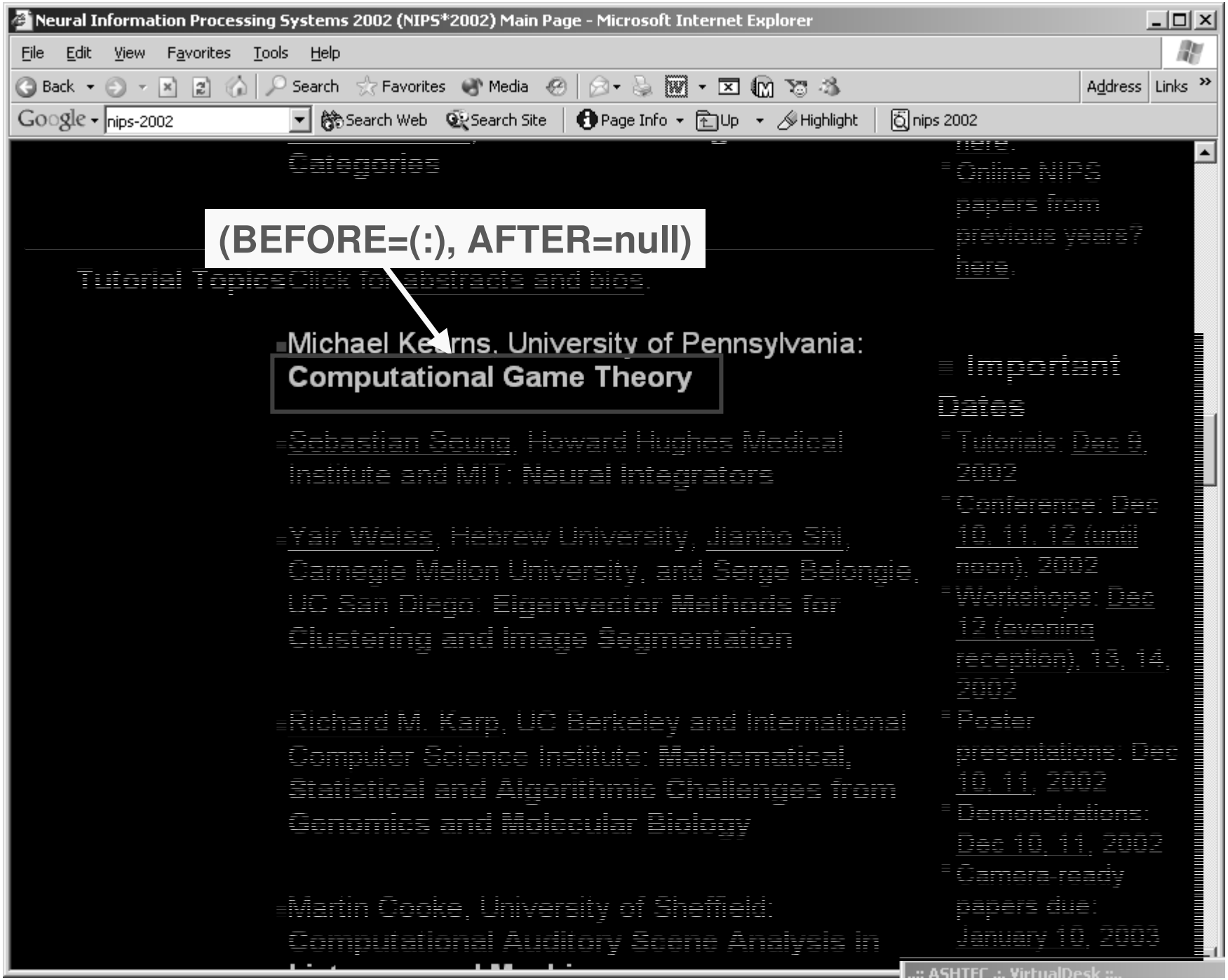
Where is Information on...

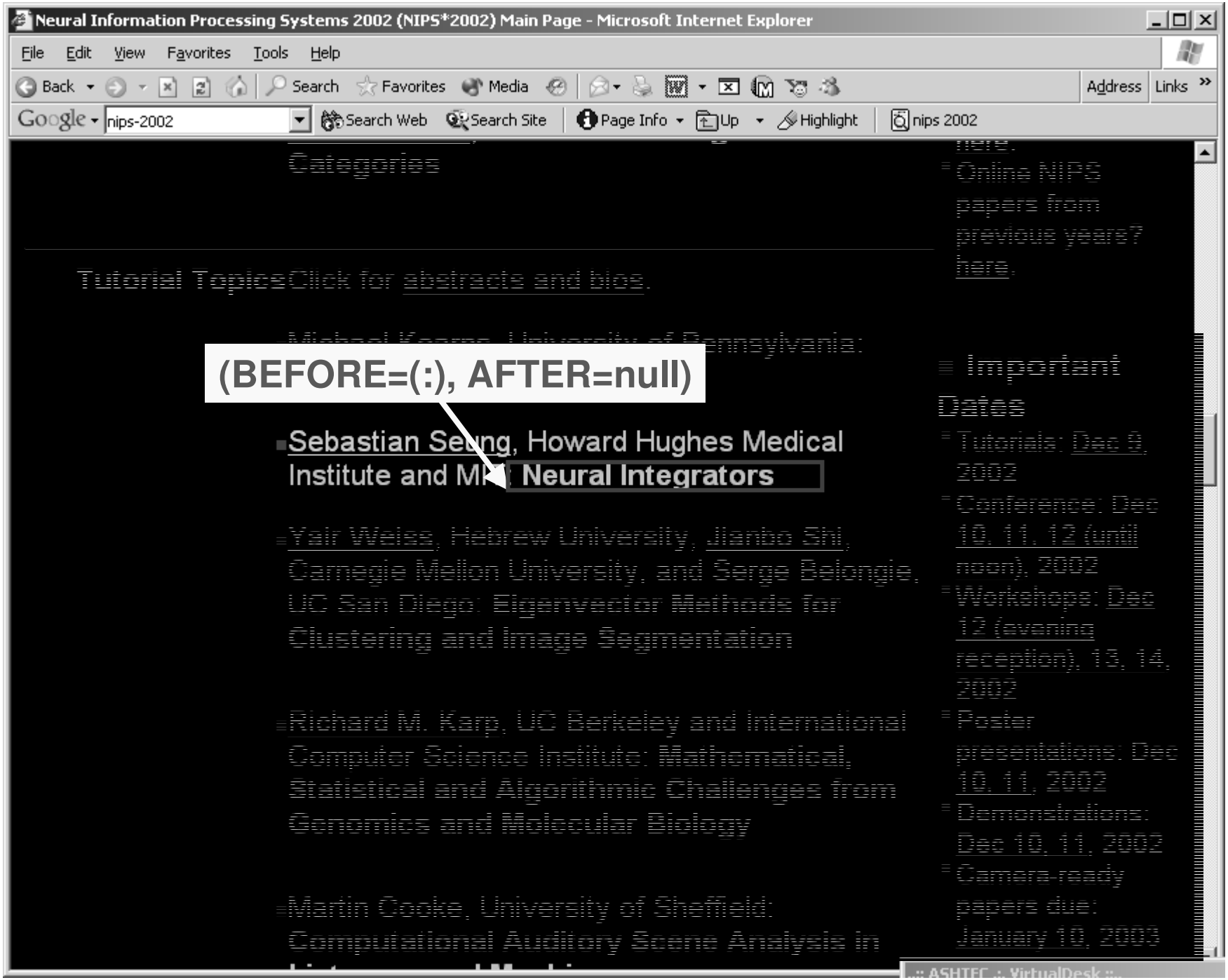
- Online registration? here.
- Conference proceedings? here.
- Proceedings on CD-ROM? here.
- Poster formats? here.
- Skiing discounts? here.
- Driving directions? here.
- Organizers of NIPS? here.
- Call for papers/proposals (now past)? here.
- NIPS*2001 conference? here.
- Online

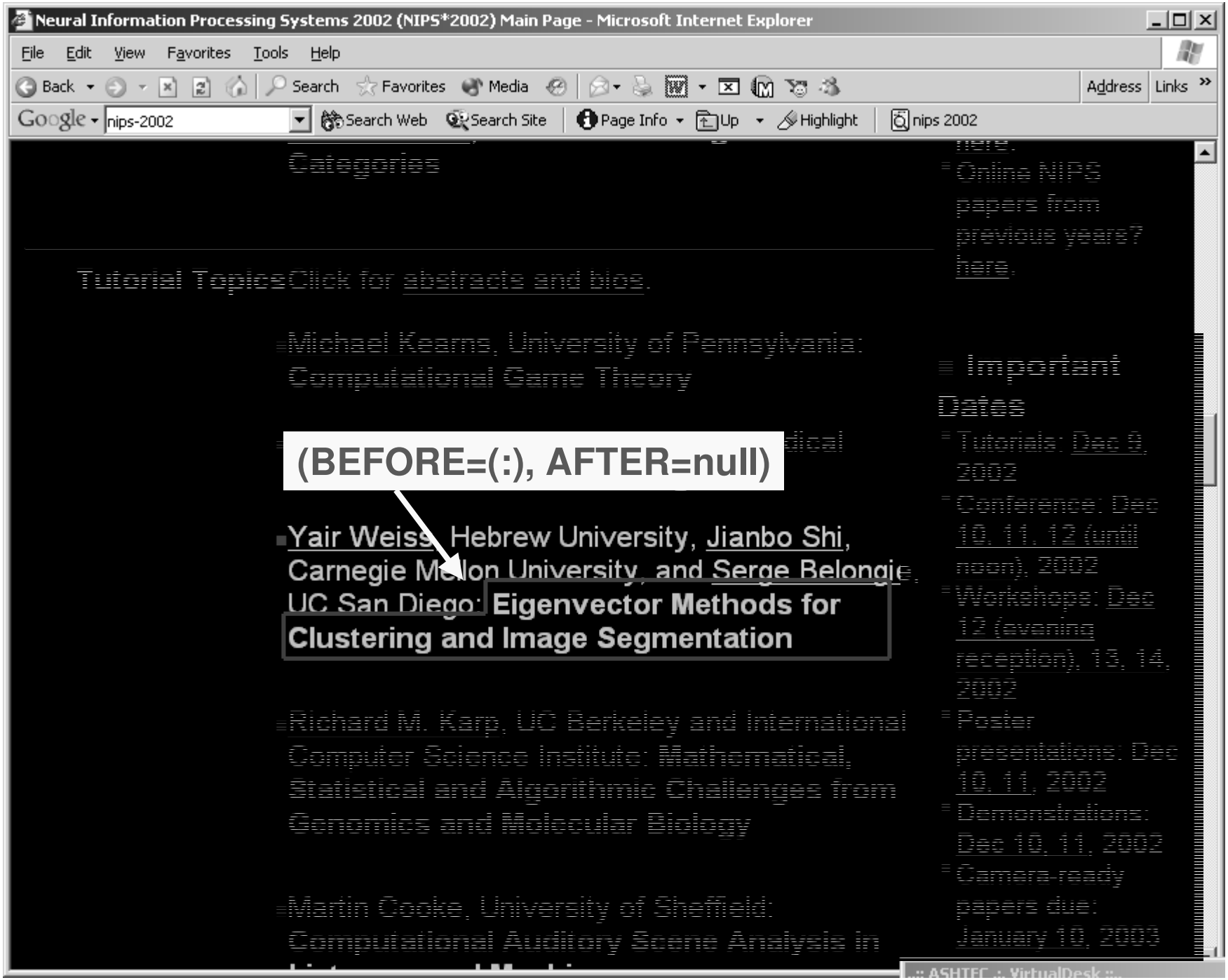
.:: ASHTEC :: VirtualDesk ::.


Neural Information Processing Systems 2002 (NIPS*2002) Main Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media Address Links
Google nips-2002 Search Web Search Site Page Info Up Highlight nips 2002
David Heeger, New York University: Neural Correlates of Perception and Attention
Andrew W. Moore, Carnegie Mellon University: Statistical Data Mining
(BEFORE=null, AFTER=(Tutorial,Topics))
Tutorial Topics Click for abstracts and bios.
Michael Kearns, University of Pennsylvania: Computational Game Theory
Sebastian Seung, Howard Hughes Medical Institute and MIT: Neural Integrators
Yair Weiss, Hebrew University, Jianbo Shi, Carnegie Mellon University, and Serge Belongie, UC San Diego: Eigenvector Methods for Clustering and Image Segmentation
Richard M. Karp, UC Berkeley and International Computer Science Institute: Mathematical, Statistical and Algorithmic Challenges from Genomics and Molecular Biology
Martin Cooke, University of Sheffield: Computational Auditory Scene Analysis in Listeners and Machines
Andrew McCallum, University of Massachusetts at Amherst, William Cohen, Carnegie Mellon University: Information Extraction from the World Wide Web
Tutorials and Conference Schedule
(BEFORE=null, AFTER=(Tutorials,and))
Online NIPS papers from previous years? here.
Workshops: Dec 12 (evening reception), 13, 14, 2002
Poster presentations: Dec 10, 11, 2002
Demonstrations: Dec 10, 11, 2002
Camera-ready papers due: January 10, 2003
Proceedings available online: March 2003
Proceedings volumes mailed: June 2003
Conference Format
The new Demonstrations component of the Conference program will enable researchers to highlight scientific advances, systems, and technologies in ways that go beyond conventional poster presentations. It will provide a unique forum for demonstrating advanced technologies both hardware and software and fostering the direct exchange of knowledge. The Tutorial Program provides a choice of six
by leading scientists. The
range of subjects
science, learning algorithms and theory, bioinformatics, image

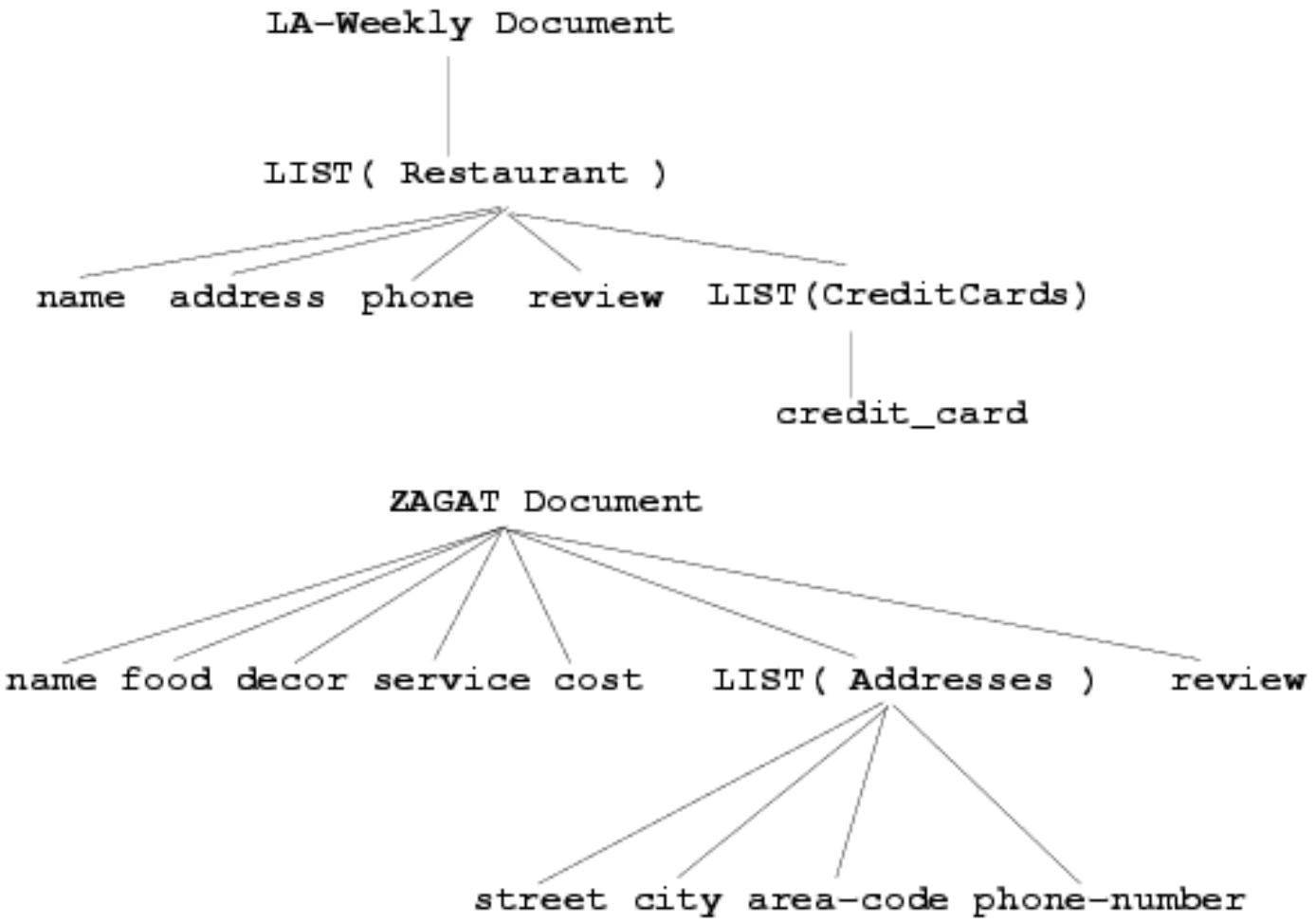

Neural Information Processing Systems 2002 (NIPS*2002) Main Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media Address Links
Google nips-2002 Search Web Search Site Page Info Up Highlight nips 2002
here.
(BEFORE=null, AFTER=(<,li,>,))
Tutorial Topics Click for abstracts and bios.
Michael Kearns, University of Pennsylvania: Computational Game Theory
Sebastian Seung, Howard Hughes Medical Institute and MIT: Neural Integrators
Yair Weiss, Hebrew University, Jianbo Shi, Carnegie Mellon University, and Serge Belongie, UC San Diego: Eigenvector Methods for Clustering and Image Segmentation
Richard M. Karp, UC Berkeley and International Computer Science Institute: Mathematical, Statistical and Algorithmic Challenges from Genomics and Molecular Biology
Martin Cooke, University of Sheffield: Computational Auditory Scene Analysis in


Neural Information Processing Systems 2002 (NIPS*2002) Main Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media Address Links
Google nips-2002 Search Web Search Site Page Info Up Highlight nips 2002
Categories
(BEFORE=(:), AFTER=null)
Tutorial TopicsClick for abstracts and bios.
Michael Kearns, University of Pennsylvania: Computational Game Theory
Sebastian Seung, Howard Hughes Medical Institute and MIT: Neural Integrators
Yair Weiss, Hebrew University, Jianbo Shi, Carnegie Mellon University, and Serge Belongie, UC San Diego: Eigenvector Methods for Clustering and Image Segmentation
Richard M. Karp, UC Berkeley and International Computer Science Institute: Mathematical, Statistical and Algorithmic Challenges from Genomics and Molecular Biology
Martin Cooke, University of Sheffield: Computational Auditory Scene Analysis in
here.
Online NIPS papers from previous years? here.
Important Dates
Tutorials: Dec 9, 2002
Conference: Dec 10, 11, 12 (until noon), 2002
Workshops: Dec 12 (evening reception), 13, 14, 2002
Poster presentations: Dec 10, 11, 2002
Demonstrations: Dec 10, 11, 2002
Camera-ready papers due: January 10, 2003


Neural Information Processing Systems 2002 (NIPS*2002) Main Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media Address Links
Google nips-2002 Search Web Search Site Page Info Up Highlight nips 2002
Categories
Tutorial TopicsClick for abstracts and bios.
(BEFORE=(:), AFTER=null)
Sebastian Seung, Howard Hughes Medical Institute and MI
Neural Integrators
Yair Weiss, Hebrew University, Jianbo Shi, Carnegie Mellon University, and Serge Belongie, UC San Diego: Eigenvector Methods for Clustering and Image Segmentation
Richard M. Karp, UC Berkeley and International Computer Science Institute: Mathematical, Statistical and Algorithmic Challenges from Genomics and Molecular Biology
Martin Cooke, University of Sheffield: Computational Auditory Scene Analysis in
here.
Online NIPS papers from previous years? here.
Important Dates
Tutorials: Dec 9, 2002
Conference: Dec 10, 11, 12 (until noon), 2002
Workshops: Dec 12 (evening reception), 13, 14, 2002
Poster presentations: Dec 10, 11, 2002
Demonstrations: Dec 10, 11, 2002
Camera-ready papers due: January 10, 2003


Neural Information Processing Systems 2002 (NIPS*2002) Main Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media Address Links
Google nips-2002 Search Web Search Site Page Info Up Highlight nips 2002
Categories
Tutorial TopicsClick for abstracts and bios.
Michael Kearns, University of Pennsylvania: Computational Game Theory
(BEFORE=(:), AFTER=null)
Yair Weiss, Hebrew University, Jianbo Shi, Carnegie Mellon University, and Serge Belongie, UC San Diego: Eigenvector Methods for Clustering and Image Segmentation
Richard M. Karp, UC Berkeley and International Computer Science Institute: Mathematical, Statistical and Algorithmic Challenges from Genomics and Molecular Biology
Martin Cooke, University of Sheffield: Computational Auditory Scene Analysis in
here.
Online NIPS papers from previous years? here.
Important Dates
Tutorials: Dec 9, 2002
Conference: Dec 10, 11, 12 (until noon), 2002
Workshops: Dec 12 (evening reception), 13, 14, 2002
Poster presentations: Dec 10, 11, 2002
Demonstrations: Dec 10, 11, 2002
Camera-ready papers due: January 10, 2003

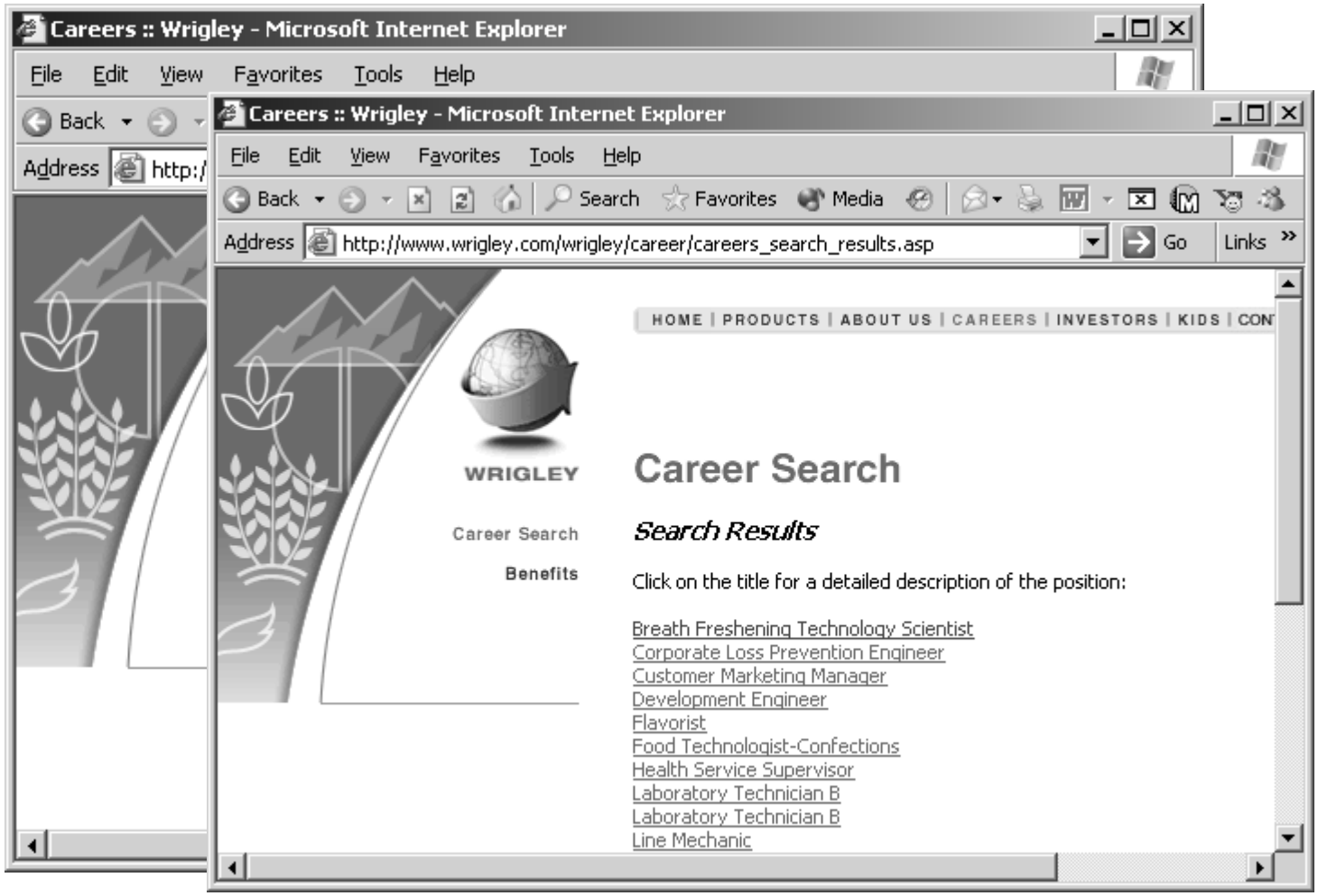
Stalker: hierarchical decomposition of two web sites



Stalker: summary and results

- Rule format:
  - "landmark automata" format for rules which extended BWI's format
    - E.g.: <a>W. Cohen CMU: Web IE
    - BWI: BEFORE=(<, /, a,>, ANY, :)
    - STALKER: BEGIN = SkipTo(<, /, a, >), SkipTo(:)
- Top-down rule learning algorithm
  - Carefully chosen ordering between types of rule specializations
- Very fast learning: e.g. 8 examples *vs. 274*
- **A lesson:** we often control the IE training data!

Why low sample complexity is important in "wrapper learning"



"Wrapster": a hybrid approach to representing wrappers

[Cohen,Jensen&Hurst WWW02]

- Common representations for web pages include:
  - a rendered **image**
  - a **DOM tree** (tree of HTML markup & text)
    - gives some of the power of hierarchical decomposition
  - a sequence of **tokens**
  - a **bag of words**, a sequence of **characters**, a node in a directed graph, . . .
- Questions:
  - How can we engineer a system to generalize quickly?
  - How can we explore representational choices easily?

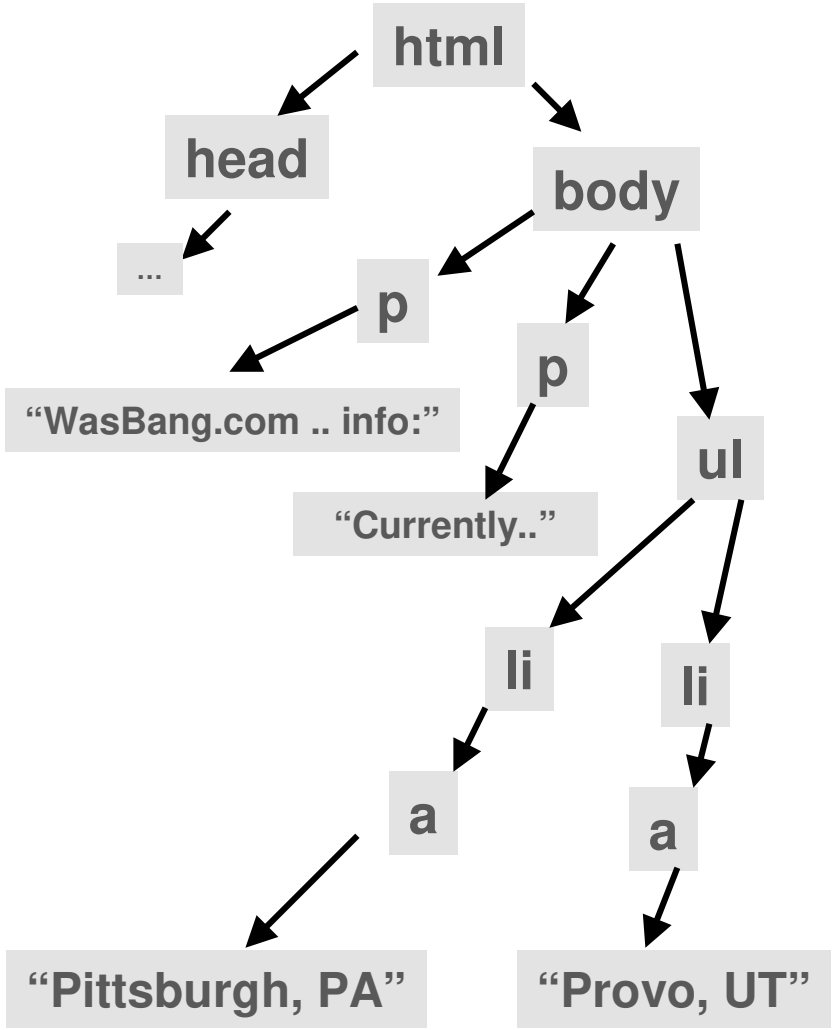
Wrapster architecture

- Bias is an ordered set of “builders”.
- Builders are simple “micro-learners”.
- A single master algorithm co-ordinates learning.
  - Hybrid top-down/bottom-up rule learning
- Terminology:
  - *Span:* substring of page, created by a *predicate*
  - *Predicate:* subset of span×span, created by a *builder*
  - *Builder:* a “micro-learner”, created by hand

Wrapster predicates

- A predicate is a binary relation on spans:
  - $p(s,t)$ means that t is extracted from s.
- Membership in a predicate can be tested:
  - Given *(s,t),* is *p(s,t)* true?
- Predicates can be executed:
  - EXECUTE(*s,t*) = { *t : p(s,t)* }

Example Wrapster predicate



http://wasBang.org/aboutus.html

WasBang.com contact info:

Currently we have offices in two locations:

- Pittsburgh, PA
- Provo, UT

Example Wrapster predicate

Example:

p(s_1,s_2) *iff* s_2 are the tokens below an li node inside a ul node inside s_1.

EXECUTE(*p*,s_1) extracts

– "Pittsburgh, PA"

– "Provo, UT"

http://wasBang.org/aboutus.html

WasBang.com contact info:

Currently we have offices in two locations:

- Pittsburgh, PA
- Provo, UT

Wrapster builders

- Builders are based on *simple, restricted languages,* for example:
  - $L_{tagpath}$: *p* is defined by $tag_1,\ldots,tag_k$ and $p_{tag1,\ldots,tagk}(s_1,s_2)$ is true iff s_1 and s_2 correspond to DOM nodes and s_2 is reached from s_1 by following a path ending in $tag_1,\ldots,tag_k$
    - EXECUTE($p_{ul,li}$,s_1) = {“Pittsburgh,PA”, “Provo, UT”}
  - $L_{bracket}$: *p* is defined by a pair of strings *(l,r),* and $p_{l,r}(s_1,s_2)$ is true iff s_2 is preceded by *l* and followed by *r*.
    - EXECUTE($p_{in,locations}$,s_1) = {“two”}

Wrapster builders

For each language L there is a builder B which implements:

- LGG(positive examples of $p(s_1,s_2)$): least general $p \in L$ that covers all the positive examples (like pairwise generalization)
  - For $L_{bracket}$, longest common prefix and suffix of the examples.
- REFINE(p, examples): a set of p's that cover some but not all of the examples.
  - For $L_{tagpath}$, extend the path with one additional tag that appears in the examples.
- Builders/languages can be combined:
  - E.g. to construct a builder for $(L_1 \wedge L_{2)}$ or $(L_1 \circ L_2)$

Wrapster builders - examples

- Compose `tagpaths' and `brackets'
  - E.g., "extract strings between '(' and ')' inside a list item inside an unordered list"
- Compose `tagpaths' and language-based extractors
  - E.g., "extract city names inside the first paragraph"
- Extract items based on position inside a *rendered* table, or properties of the *rendered* text
  - E.g., "extract items inside any column headed by text containing the words 'Job' and 'Title'"
  - E.g. "extract items in boldfaced italics"

Composing builders

- Composing builders for $L_{tagpath}$ and $L_{bracket}$.
- LGG of the locations would be $p_{tags} \circ p_{L,R}$ where
  - *tags = ul,li*
  - *L* = "("
  - *R* = ")"

- Jobs at WasBang.com:
  Call (888)-555-1212 now to apply!
- Webmaster (New York). Perl, servlets essential.
- Librarian (Pittsburgh). MLS required.
- Ski Instructor (Vancouver). Snowboarding skills also useful.

Composing builders – structural/global

- Composing builders for $L_{tagpath}$ and L_{city}
- $L_{city} = \{p_{city}\}$ where $p_{city}(s_1,s_2)$ iff s_2 is a city name inside of s_2.
- LGG of the locations would be $p_{tags} \circ p_{city}$

- Jobs at WasBang.com:

  Call Alberta Hill at 1-888-555-1212 now to apply!

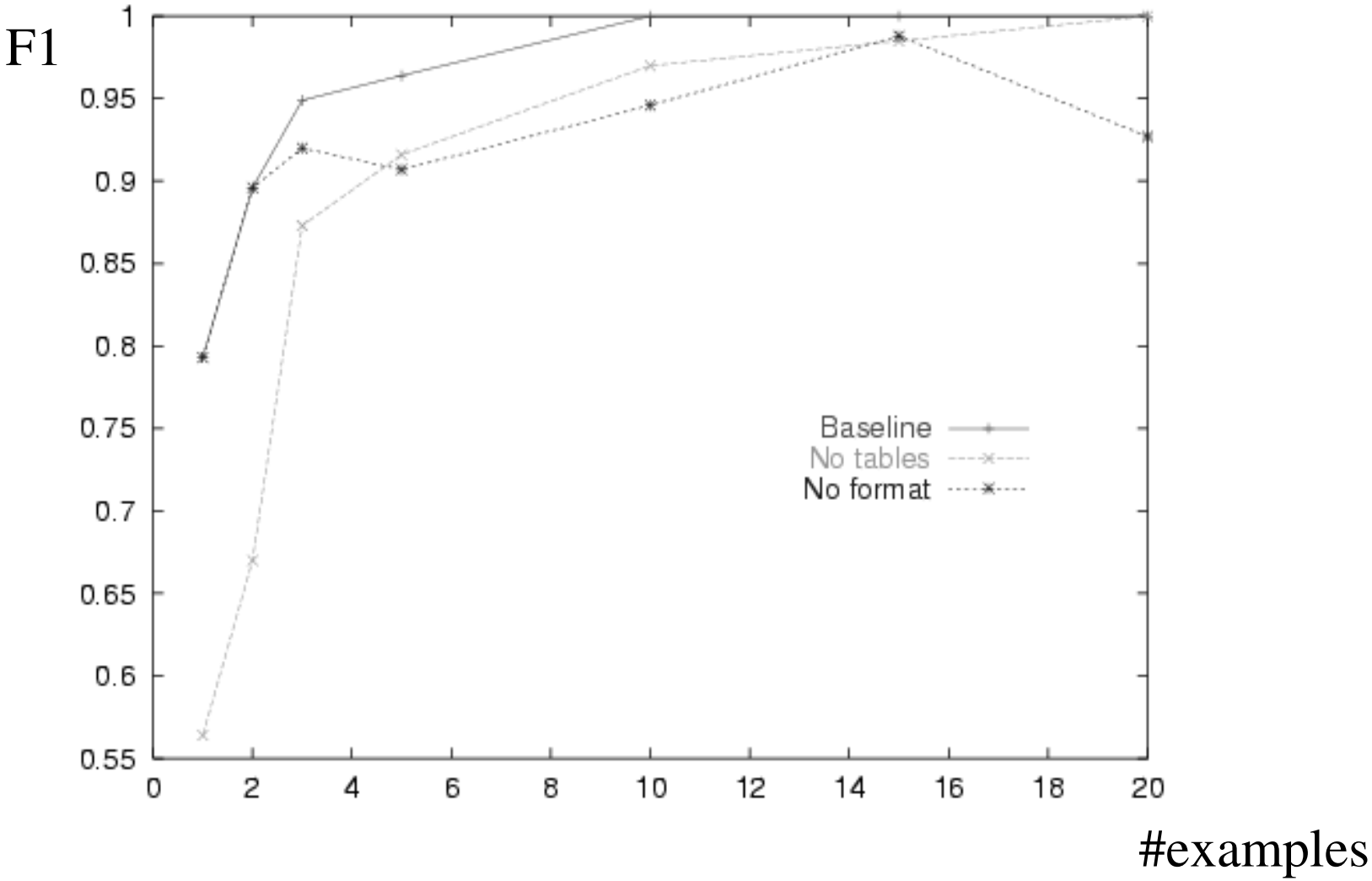
- Webmaster (New York). Perl, servlets essential.
- Librarian (Pittsburgh). MLS required.
- Ski Instructor (Vancouver). Snowboarding skills also useful.

Table-based builders

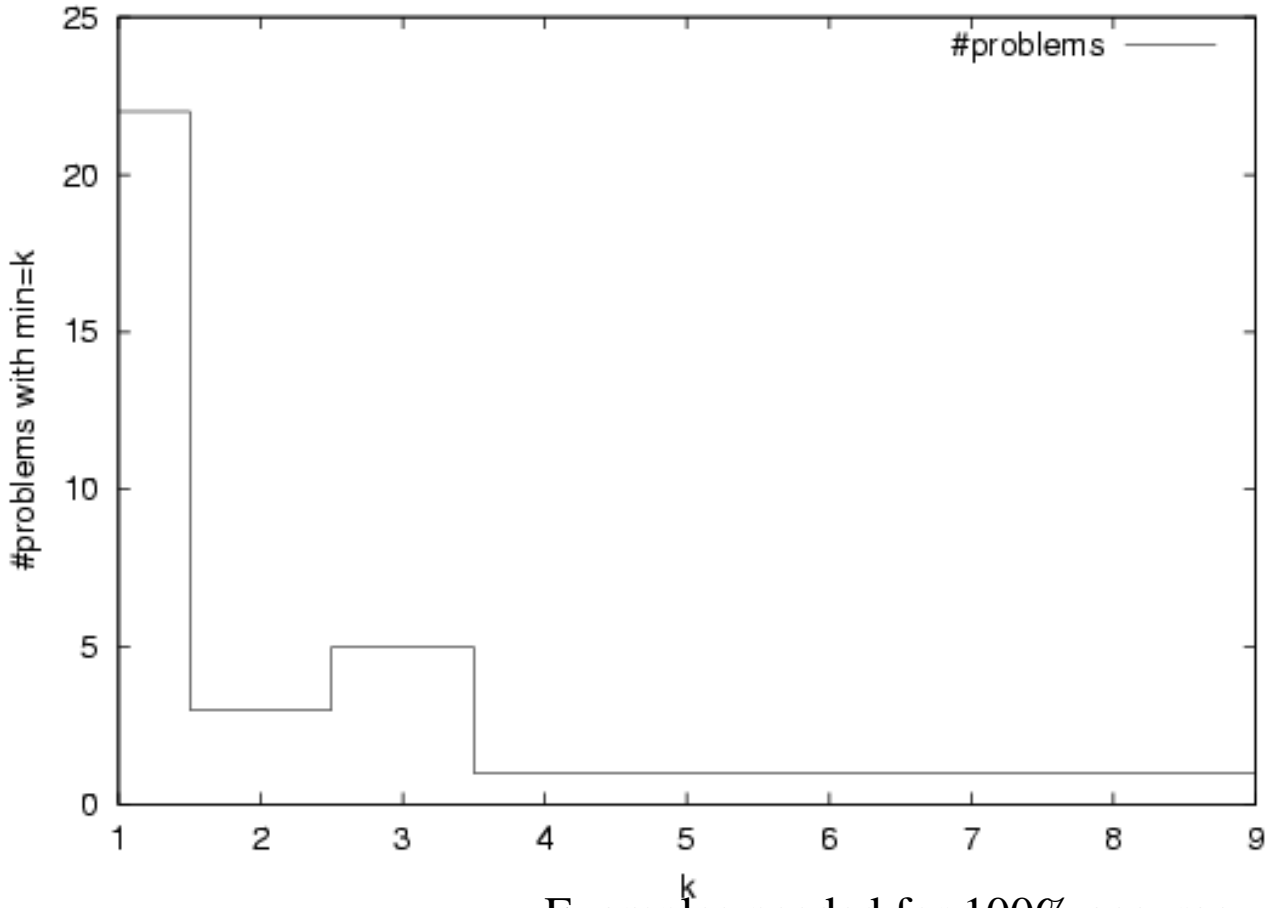
"Actresses"			
Lucy	Lawless	images	links
Angelina	Jolie	images	links
...	...	...	...
"Singers"			
Madonna		images	links
Brittany	Spears	images	links
...	...	...	...

How to represent "links to pages about singers"?
Builders can be based on a **geometric** view of a page.

Wrapster results



Wrapster results



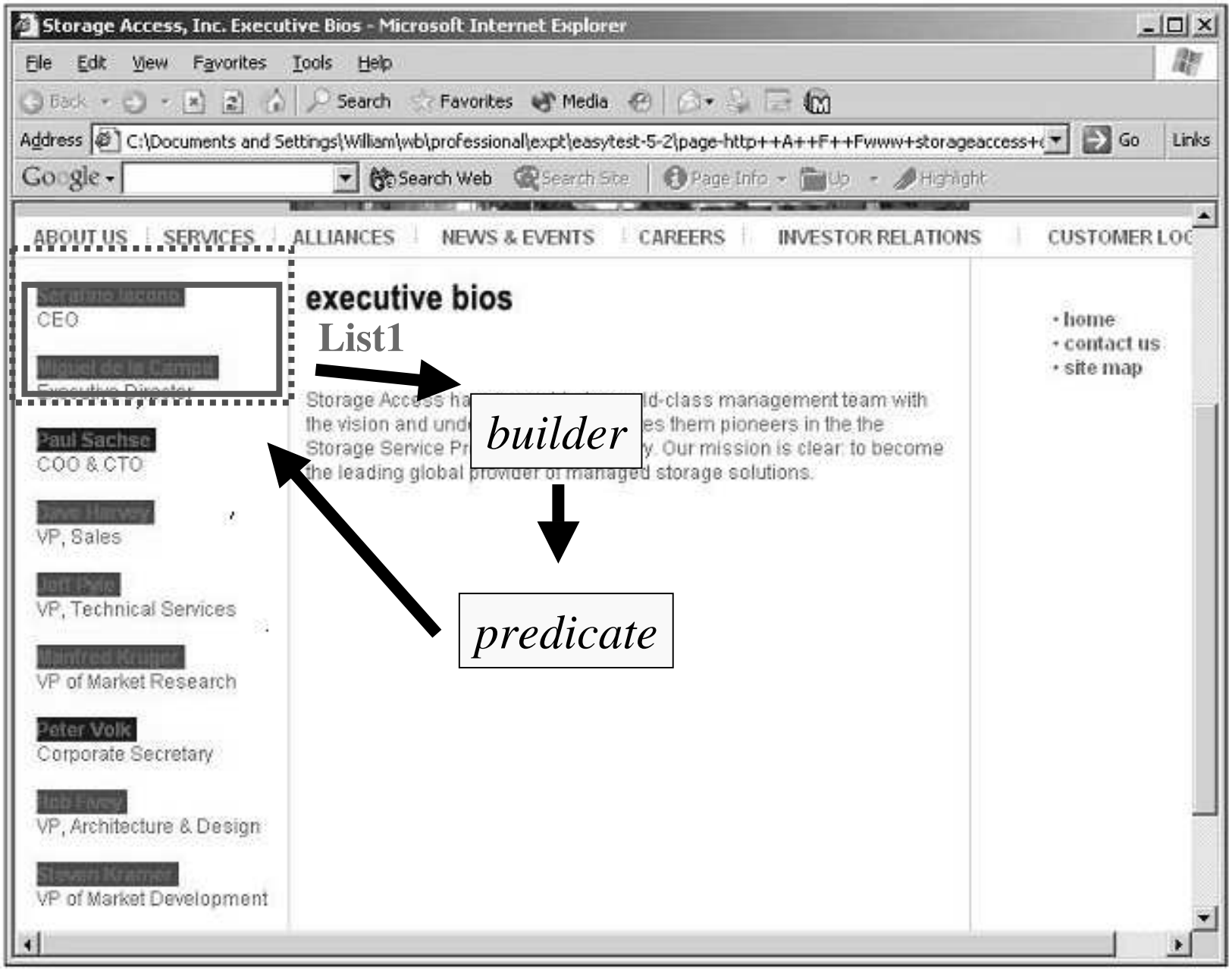
Examples needed for 100% accuracy

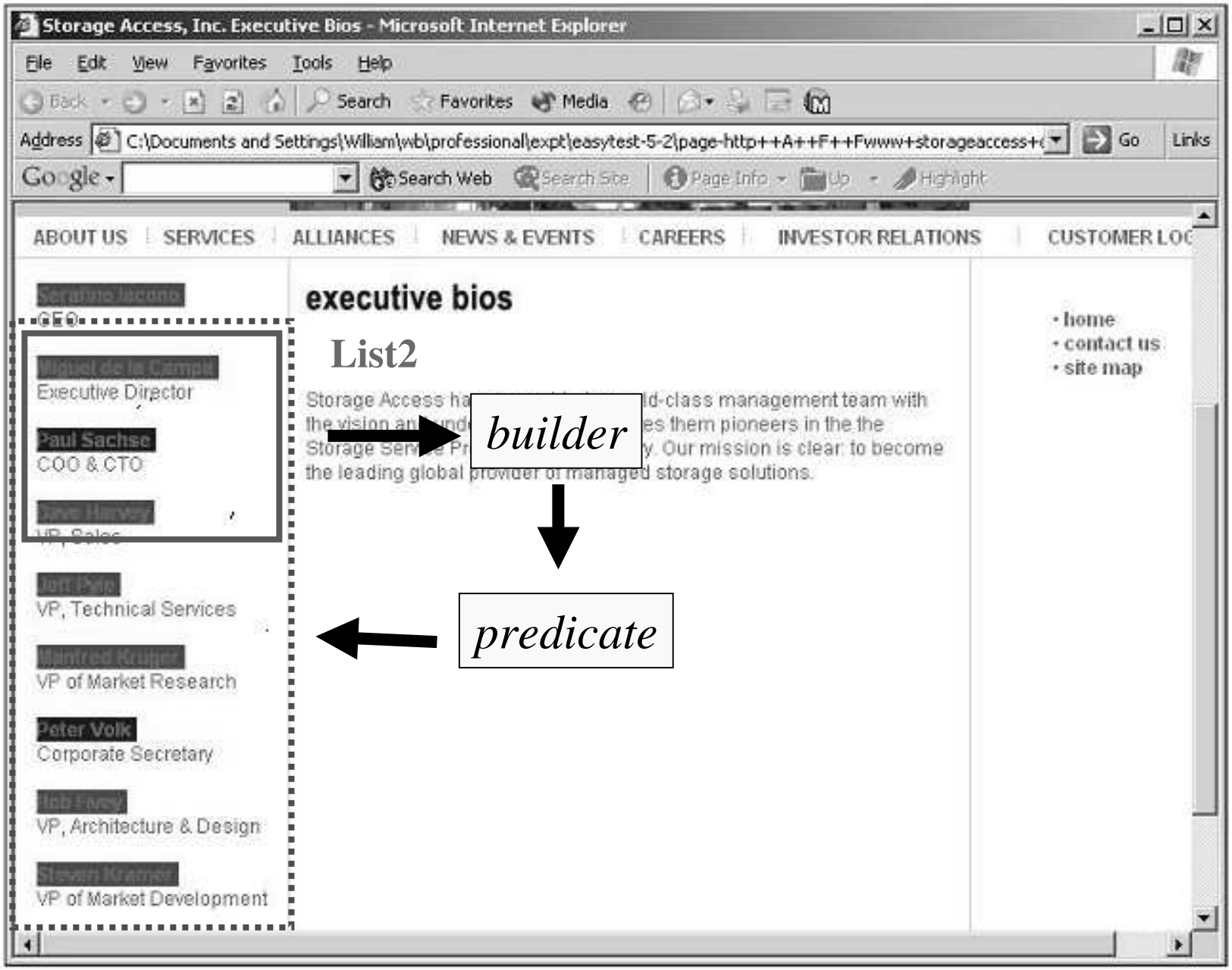
Site-dependent *vs.* site-independent IE

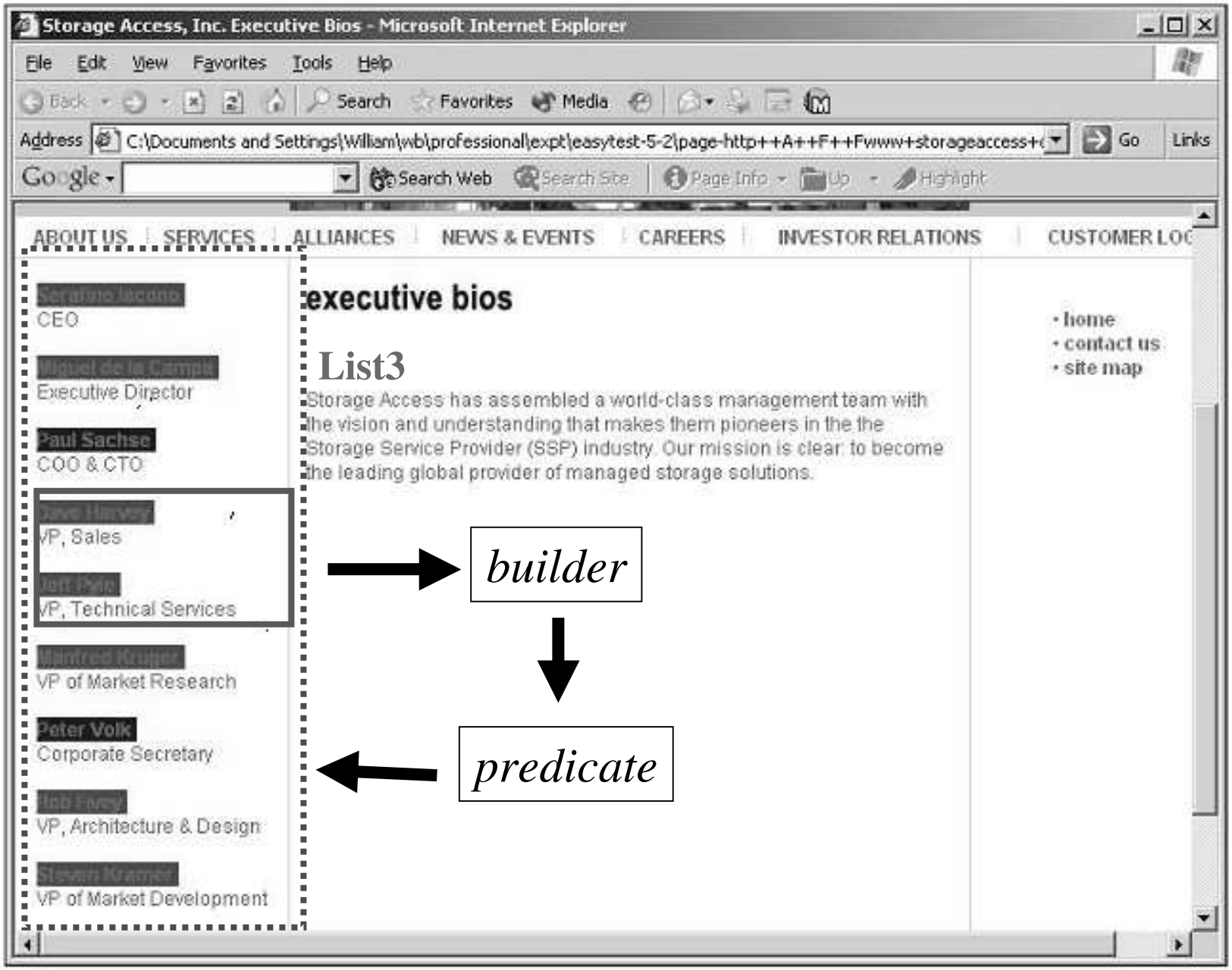
- When is *formatting* information useful?
  - On a **single site,** format is extremely consistent.
  - Across **many sites,** format can vary widely.
- Can we **improve** a site-independent classifier using site-dependent format features? For instance:
  - "Smooth" predictions toward ones that are **locally consistent** with formatting.
  - Learn a "wrapper" from "noisy" labels given by a site-independent IE system.
- First step: obtaining features from the builders

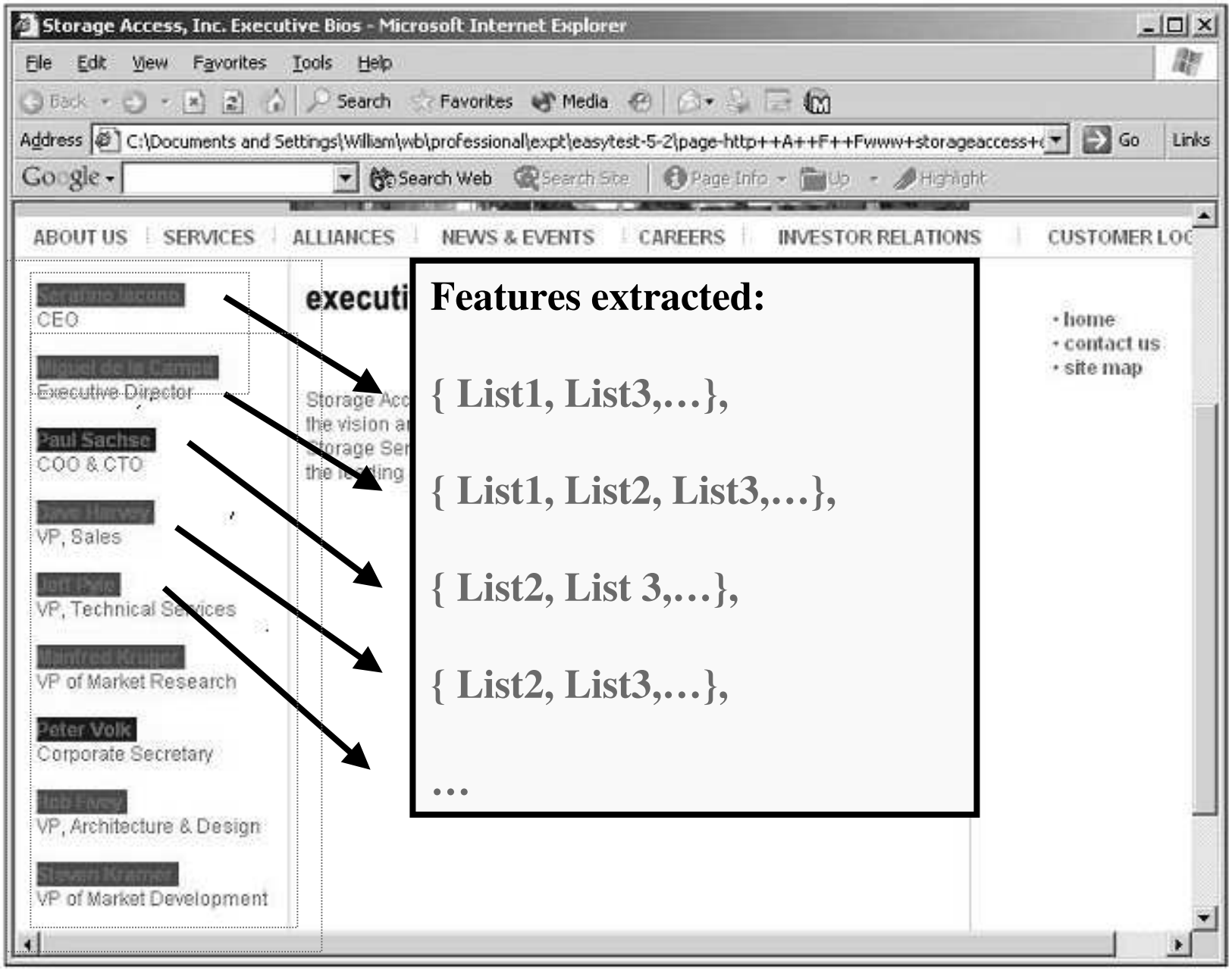
Feature construction using builders

- Let D be the set of all positive examples. Generate many small training sets D_i from D, by sliding small windows over D.
- Let P be the set of all predicates found by any builder from any subset D_i.
- For each predicate p, add a new feature f_p that is true for exactly those $x \in D$ that are extracted from their containing page by p.


Storage Access, Inc. Executive Bios - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address C:\Documents and Settings\William\wb\professional\expt\easytest-5-2\page-http++A++F++Fwww+storageaccess+
Go Links
Google Search Web Search Site Page Info Up Highlight
ABOUT US SERVICES ALLIANCES NEWS & EVENTS CAREERS INVESTOR RELATIONS CUSTOMER LOC
CEO
Executive Director
Paul Sachse
COO & CTO
VP, Sales
VP, Technical Services
VP of Market Research
Peter Volk
Corporate Secretary
VP, Architecture & Design
VP of Market Development
executive bios
List1
builder
predicate
• home
• contact us
• site map

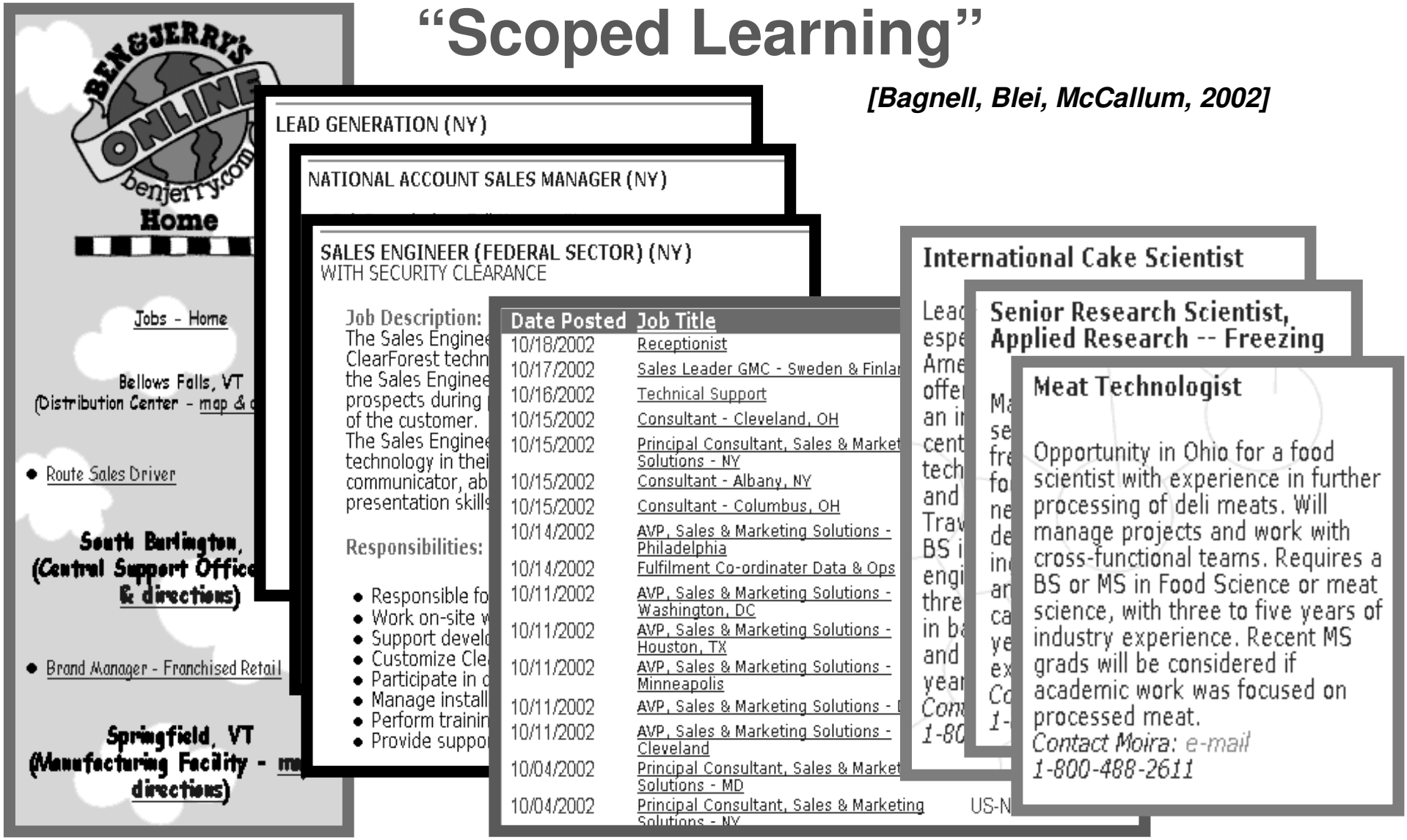

Storage Access, Inc. Executive Bios - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address C:\Documents and Settings\William\wb\professional\expt\easytest-5-2\page-http++A++F++Fwww+storageaccess+
Go Links
Google Search Web Search Site Page Info Up Highlight
ABOUT US SERVICES ALLIANCES NEWS & EVENTS CAREERS INVESTOR RELATIONS CUSTOMER LOC
executive bios
List2
Executive Director
COO & CTO
VP, Technical Services
VP of Market Research
Corporate Secretary
VP, Architecture & Design
VP of Market Development
builder
predicate
• home
• contact us
• site map


Storage Access, Inc. Executive Bios - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address C:\Documents and Settings\William\wb\professional\expt\easytest-5-2\page-http++A++F++Fwww+storageaccess+
Go Links
Google Search Web Search Site Page Info Up Highlight
ABOUT US SERVICES ALLIANCES NEWS & EVENTS CAREERS INVESTOR RELATIONS CUSTOMER LOC
CEO
Executive Director
Paul Sachse
COO & CTO
VP, Sales
VP, Technical Services
VP of Market Research
Peter Volk
Corporate Secretary
VP, Architecture & Design
VP of Market Development
executive bios
List3
Storage Access has assembled a world-class management team with the vision and understanding that makes them pioneers in the the Storage Service Provider (SSP) industry. Our mission is clear: to become the leading global provider of managed storage solutions.
• home
• contact us
• site map
builder
predicate


Storage Access, Inc. Executive Bios - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Search Favorites Media
Address C:\Documents and Settings\William\wb\professional\expt\easytest-5-2\page-http++A++F++Fwww+storageaccess+c
Go Links
Google Search Web Search Site Page Info Up Highlight
ABOUT US SERVICES ALLIANCES NEWS & EVENTS CAREERS INVESTOR RELATIONS CUSTOMER LOC
CEO
Executive Director
Paul Sachse
COO & CTO
VP, Sales
VP, Technical Services
VP of Market Research
Peter Volk
Corporate Secretary
VP, Architecture & Design
VP of Market Development
executi
Storage Acc
the vision a
Storage Ser
the leading
• home
• contact us
• site map
Features extracted:
{ List1, List3,…},
{ List1, List2, List3,…},
{ List2, List 3,…},
{ List2, List3,…},
…

Learning Formatting Patterns "On the Fly": "Scoped Learning"

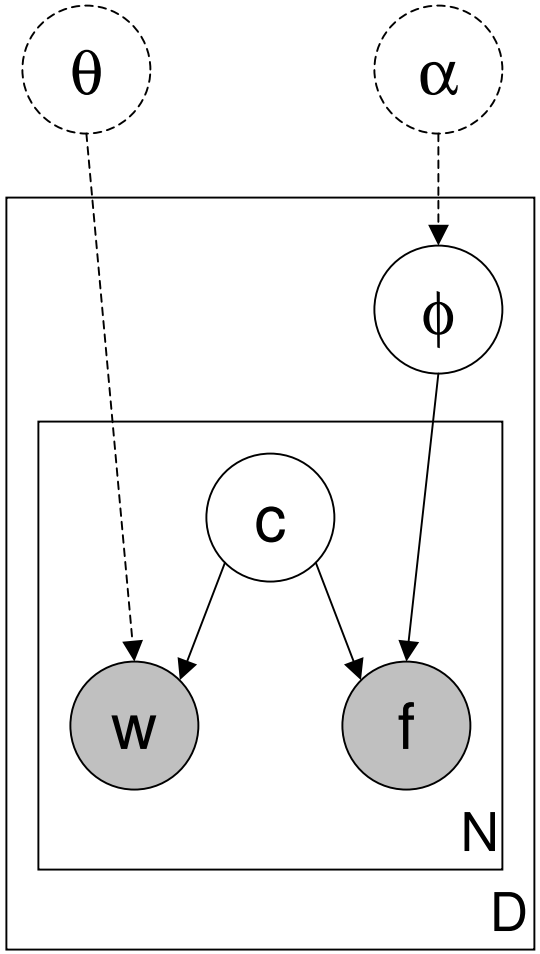
[Bagnell, Blei, McCallum, 2002]



Formatting is regular on each site, but there are too many different sites to wrap.
Can we get the best of both worlds?

Scoped Learning Generative Model

1. For each of the D documents:
   a) Generate the multinomial formatting feature parameters ϕ from $p(\phi|\alpha)$
2. For each of the N words in the document:
   a) Generate the *n*th category c_n from $p(c_n)$.
   b) Generate the *n*th word (global feature) from $p(w_n|c_n, \theta)$
   c) Generate the *n*th formatting feature (local feature) from $p(f_n|c_n, \phi)$



$$p(\phi, \mathbf{c}, \mathbf{w}, \mathbf{f}) = p_\alpha(\phi) \prod_{n=1}^{N} p(c_n) p_\theta(w_n|c_n) p(f_n|c_n, \phi)$$

Inference

Given a new web page, we would like to classify each word resulting in $\mathbf{c} = \{c_1, c_2, \ldots, c_n\}$

$$p(\mathbf{c}|\mathbf{w},\mathbf{f}) = \frac{\int \prod_{n=1}^{N} p(w_n|c_n)p(f_n|c_n,\phi)p(c_n)p(\phi)d\phi}{\int \prod_{n=1}^{N} \sum_{c_n} p(w_n|c_n)p(f_n|c_n,\phi)p(c_n)p(\phi)d\phi}$$

This is not feasible to compute because of the integral and sum in the denominator. We experimented with two approximations:

- MAP point estimate of ϕ
- Variational inference

MAP Point Estimate

If we approximate ϕ with a point estimate, $\hat{\phi}$, then the integral disappears and c decouples. We can then label each word with:

$$\hat{c}_n = \arg\max_{c_n} p(w_n|c_n)p(f_n|c_n, \hat{\phi})p(c_n)$$

A natural point estimate is the posterior mode: a maximum likelihood estimate for the local parameters given the document in question:

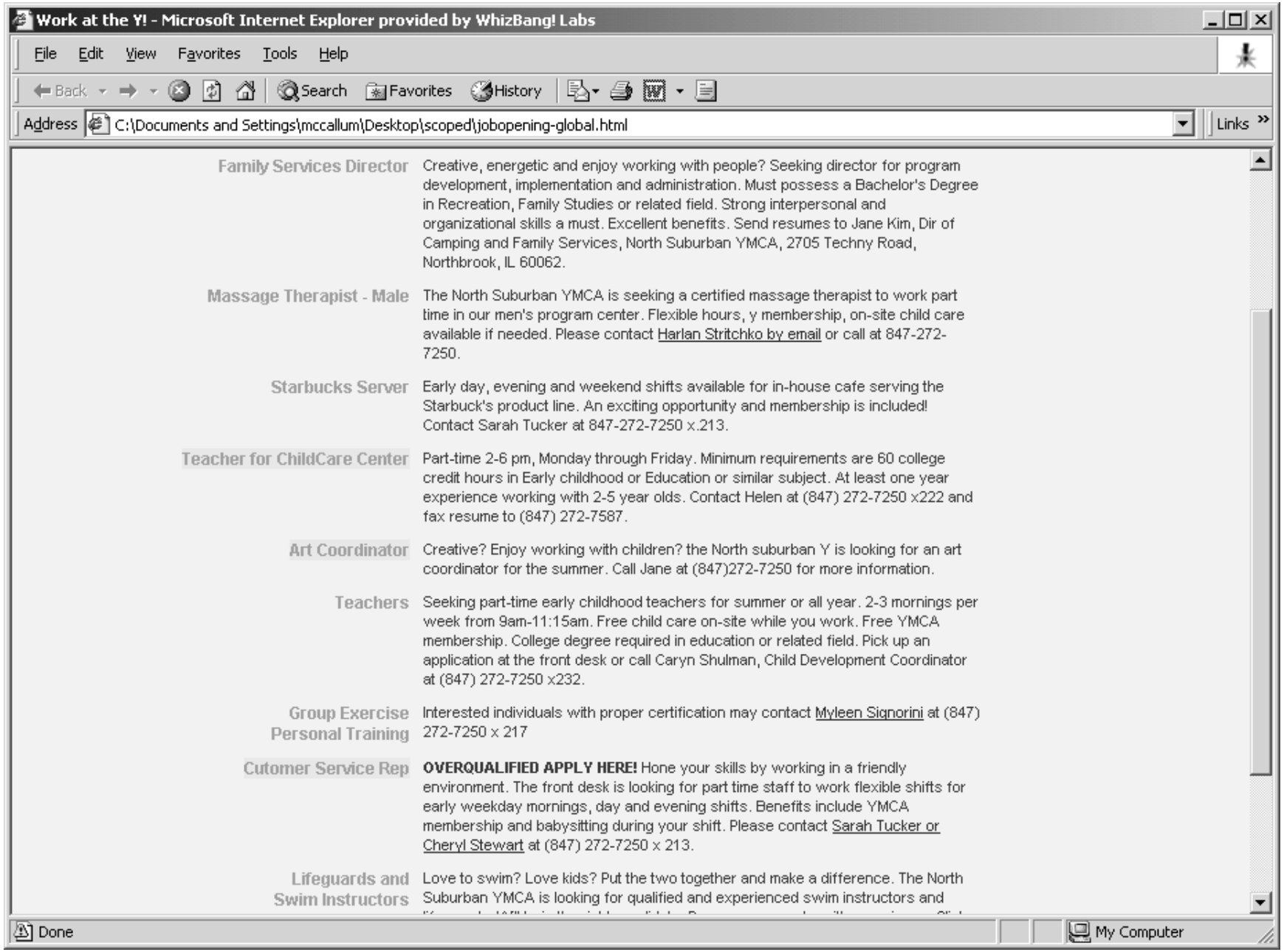
$$\hat{\phi} = \arg\max_{\phi} p(\phi|\mathbf{f}, \mathbf{w})$$

E-step:

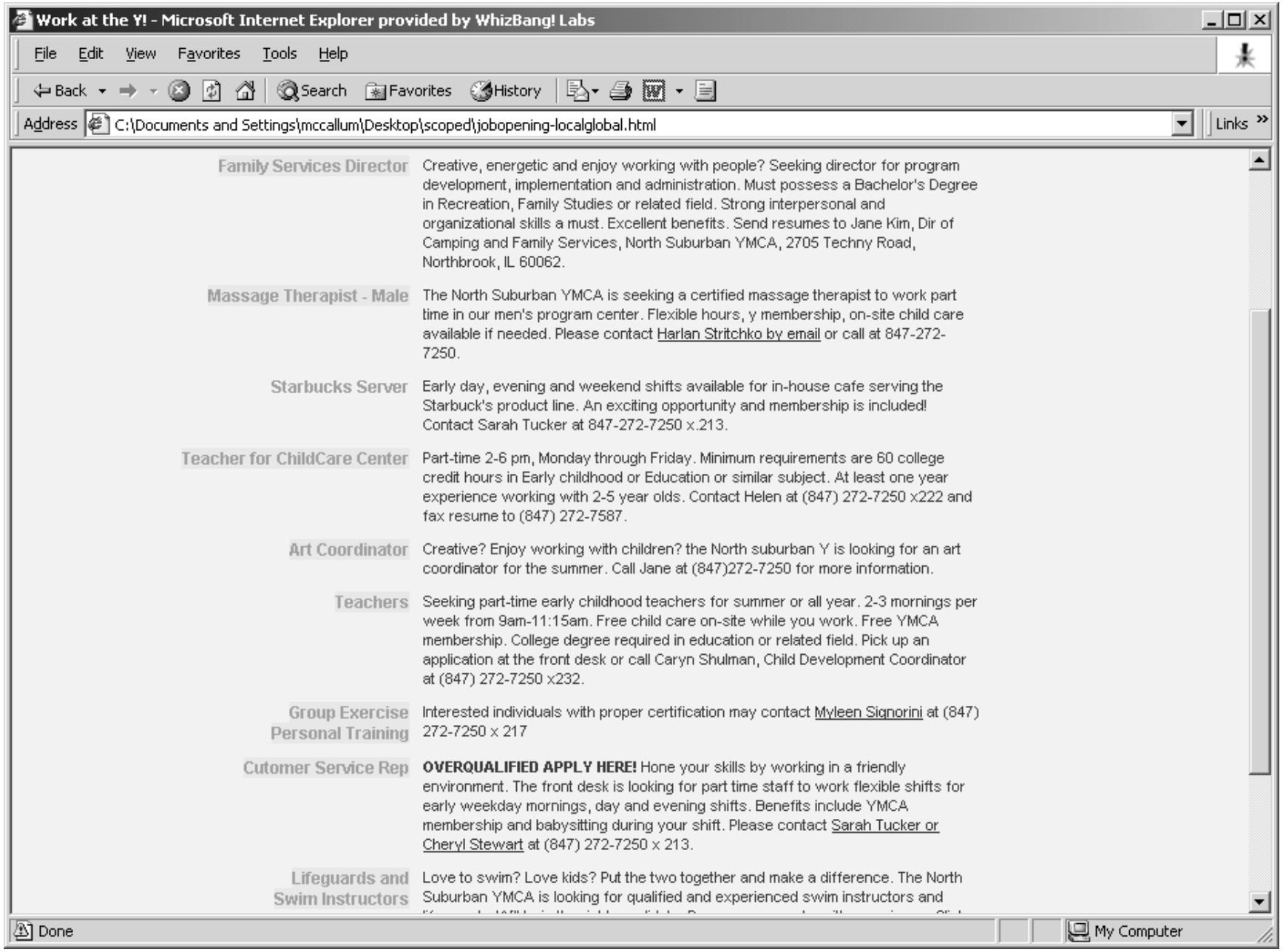
$$p^{(t+1)}(c_n|w_n, f_n; \phi) \propto p^{(t)}(f_n|c_n; \phi)p(w_n|c_n)p(c_n)$$

M-step:

$$\hat{\phi}_{c,f} = p^{(t+1)}(f|c; \phi) \propto \sum_{\{n:c_n=c, f_n=f\}} p^{(t)}(c_n|f_n, w_n)$$



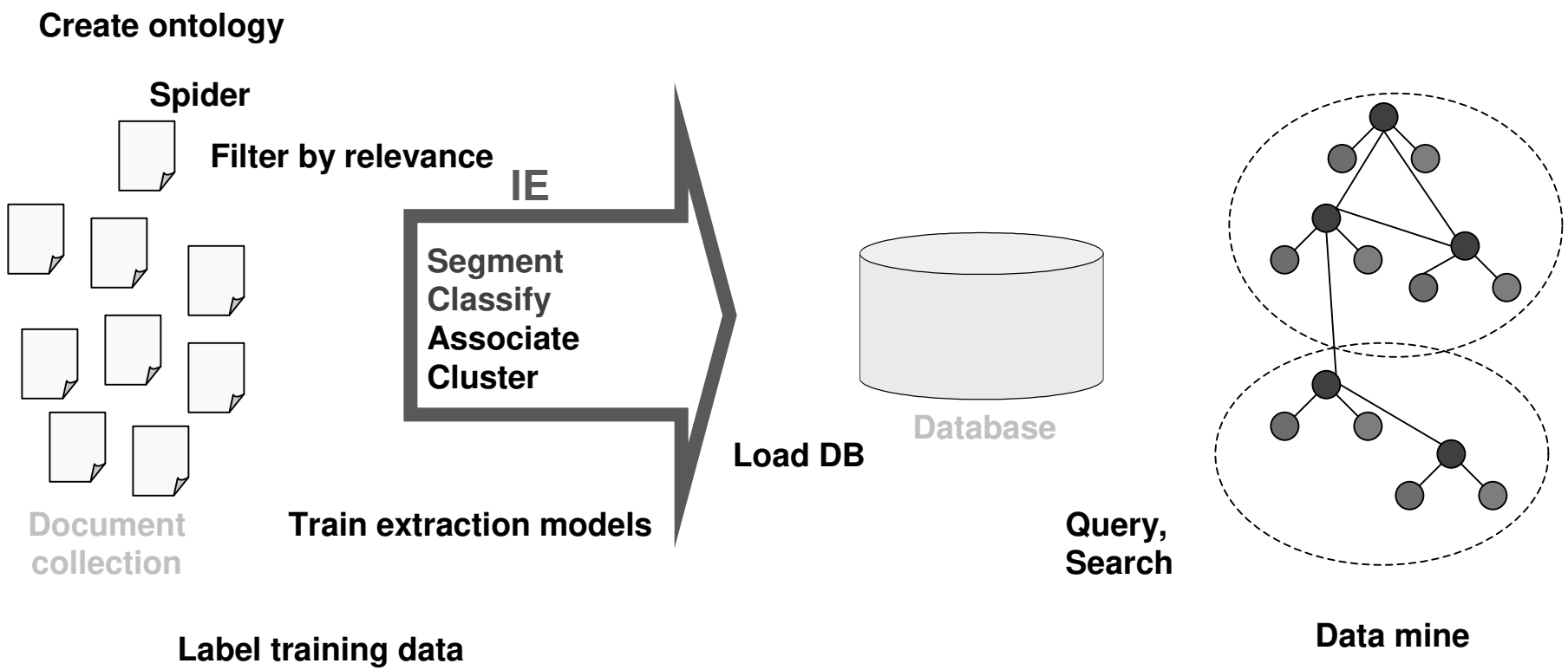
Global Extractor: Precision = 46%, Recall = 75%



Scoped Learning Extractor: Precision = 58%, Recall = 75% ***Δ Error = -22%***

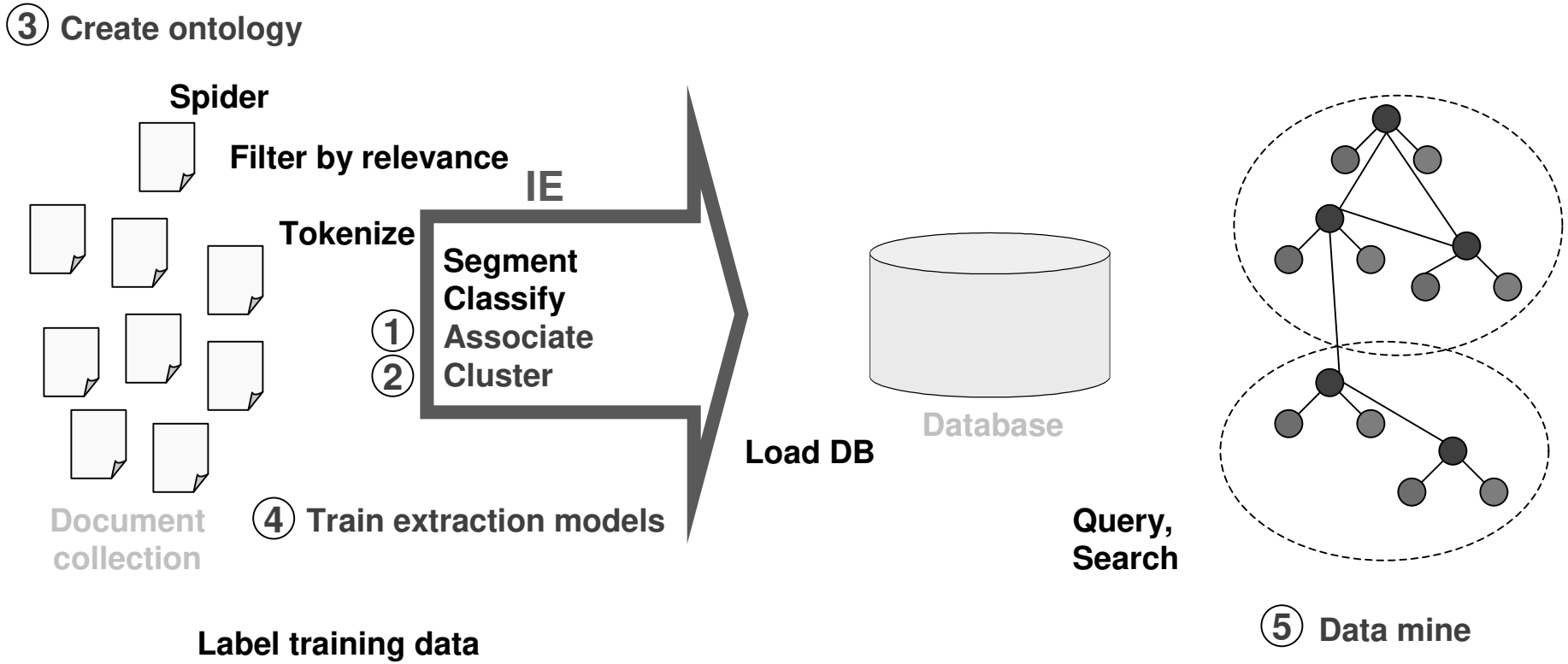
Broader View

Up to now we have been focused on segmentation and classification



Broader View

Now touch on some other issues



(1) Association as Binary Classification

Sebastian Thrun conferred with Sue Becker, the NIPS*2002 General Chair.

Person Person Role

Person-Role (Sebastian Thrun, NIPS*2002 General Chair) → NO

Person-Role (Sue Becker, NIPS*2002 General Chair) → YES

Do this with SVMs and tree kernels over parse trees.

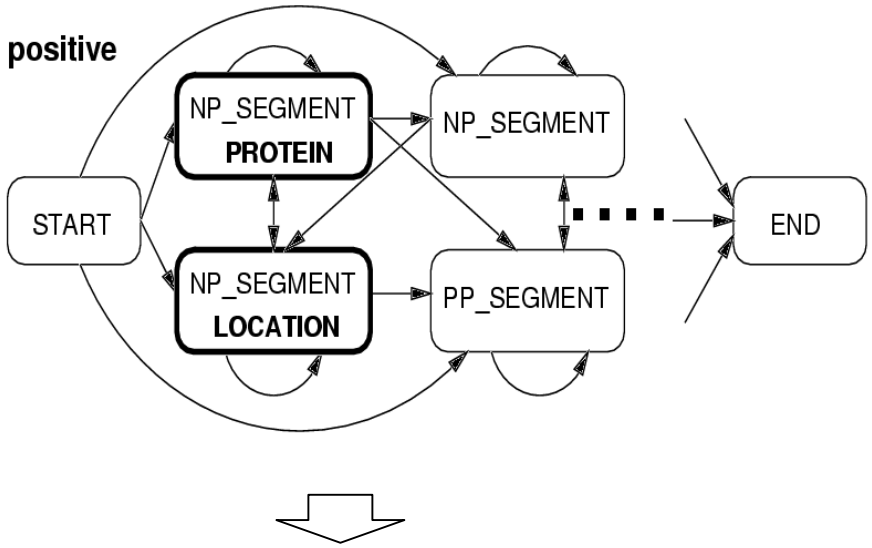
[Zelenko et al, 2002]

(1) Association with Finite State Machines

[Ray & Craven, 2001]

… This enzyme, UBC6, localizes to the endoplasmic reticulum, with the catalytic domain facing the cytosol. …

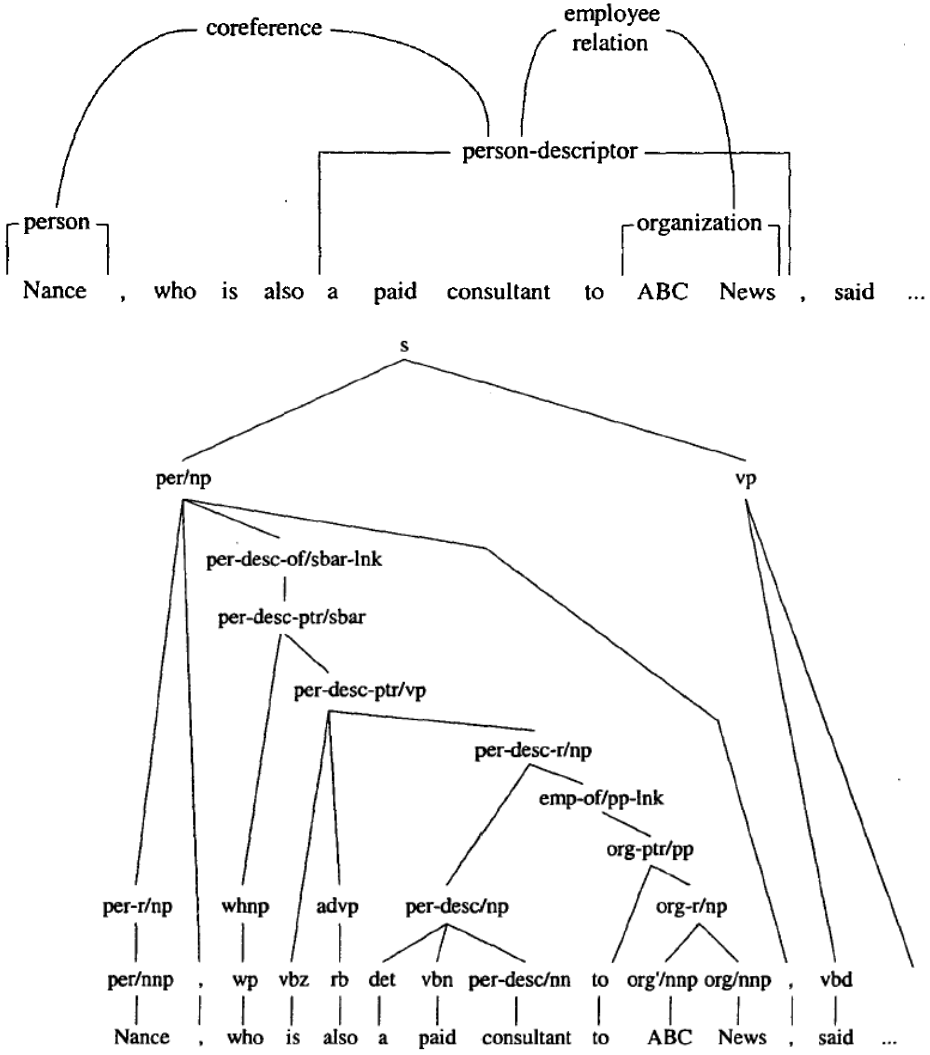
DET	this
N	enzyme
N	ubc6
V	localizes
PREP	to
ART	the
ADJ	endoplasmic
N	reticulum
PREP	with
ART	the
ADJ	catalytic
N	domain
V	facing
ART	the
N	cytosol



Subcellular-localization (UBC6, endoplasmic reticulum)

(1) Association using Parse Tree

Simultaneously POS tag, parse, extract & associate! *[Miller et al 2000]*



Increase space of parse constitutes to include entity and relation tags

Notation	Description
c_h	head constituent category
c_m	modifier constituent category
X_p	X of parent node
t	POS tag
w	word

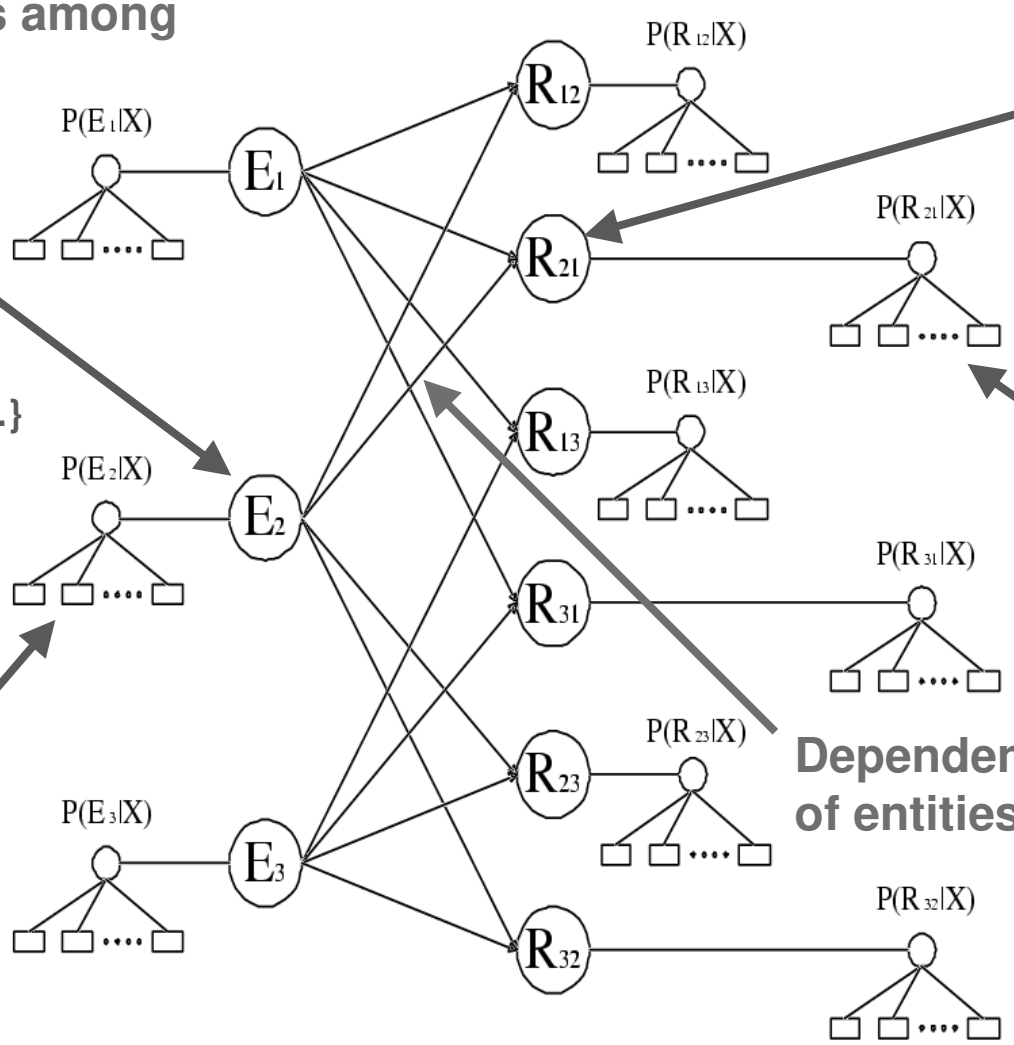
Parameters	e.g.
$P(c_h \mid c_p)$	P(vp|s)
$P(c_m \mid c_p, c_{hp}, c_{m-1}, w_p)$	P(per/np|s,vp,null,said)
$P(t_m \mid c_m, t_h, w_h)$	P(per/nnp|per/np,vbd,said)
$P(w_m \mid c_m, t_m, t_h, w_h)$	P(nance|per/np,per/nnp,vbd,said)

(This is also a great example of extraction using a tree model.)

(1) Association with Graphical Models

[Roth & Yih 2002]

Capture arbitrary-distance dependencies among predictions.



Random variable over the class of entity #1, e.g. over {person, location,…}

Random variable over the class of relation between entity #2 and #1, e.g. over {lives-in, is-boss-of,…}

Local language models contribute evidence to entity classification.

Local language models contribute evidence to relation classification.

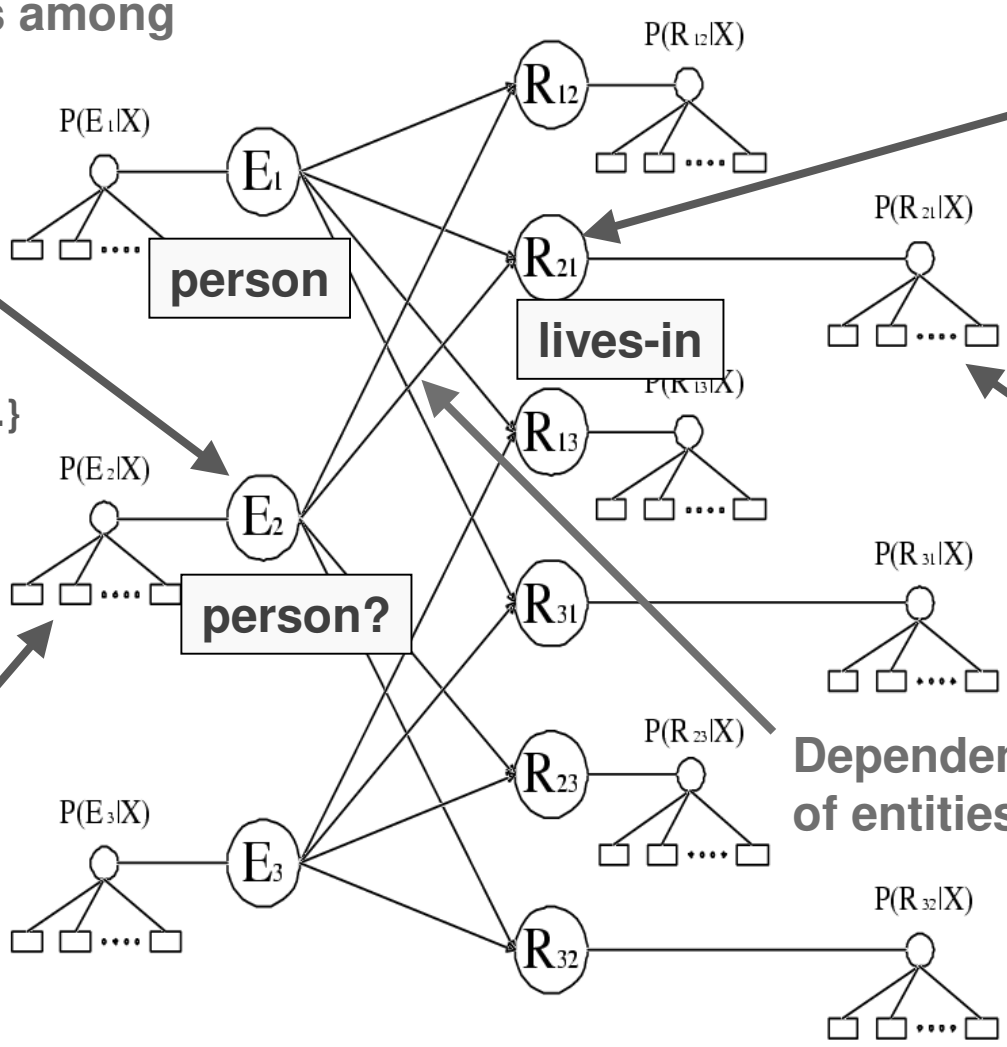
Dependencies between classes of entities and relations!

Inference with loopy belief propagation.

(1) Association with Graphical Models

[Roth & Yih 2002]

Also capture long-distance dependencies among predictions.



Random variable over the class of entity #1, e.g. over {person, location,…}

Random variable over the class of relation between entity #2 and #1, e.g. over {lives-in, is-boss-of,…}

Local language models contribute evidence to entity classification.

Local language models contribute evidence to relation classification.

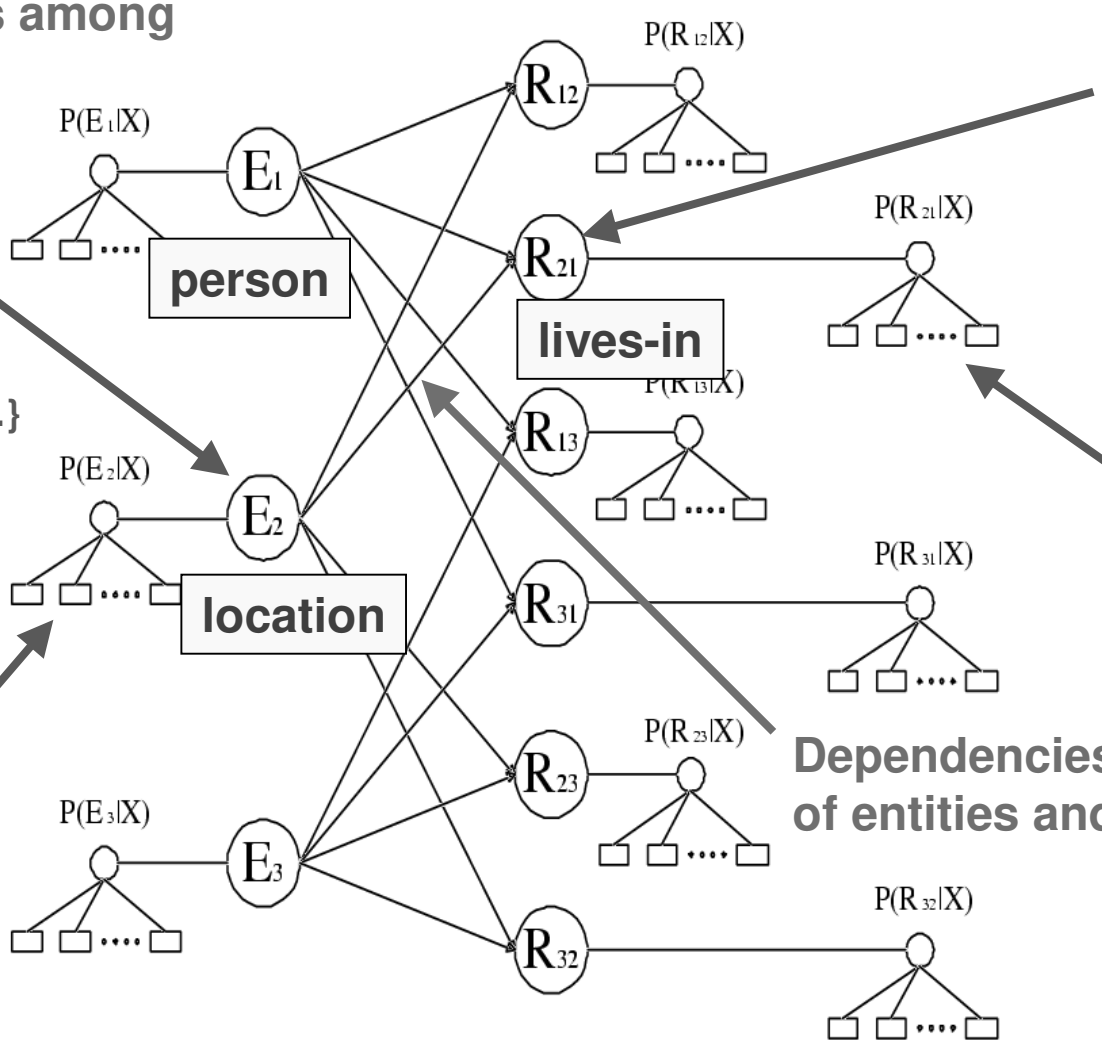
Dependencies between classes of entities and relations!

Inference with loopy belief propagation.

(1) Association with Graphical Models

[Roth & Yih 2002]

Also capture long-distance dependencies among predictions.



Random variable over the class of entity #1, e.g. over {person, location,…}

Random variable over the class of relation between entity #2 and #1, e.g. over {lives-in, is-boss-of,…}

Local language models contribute evidence to entity classification.

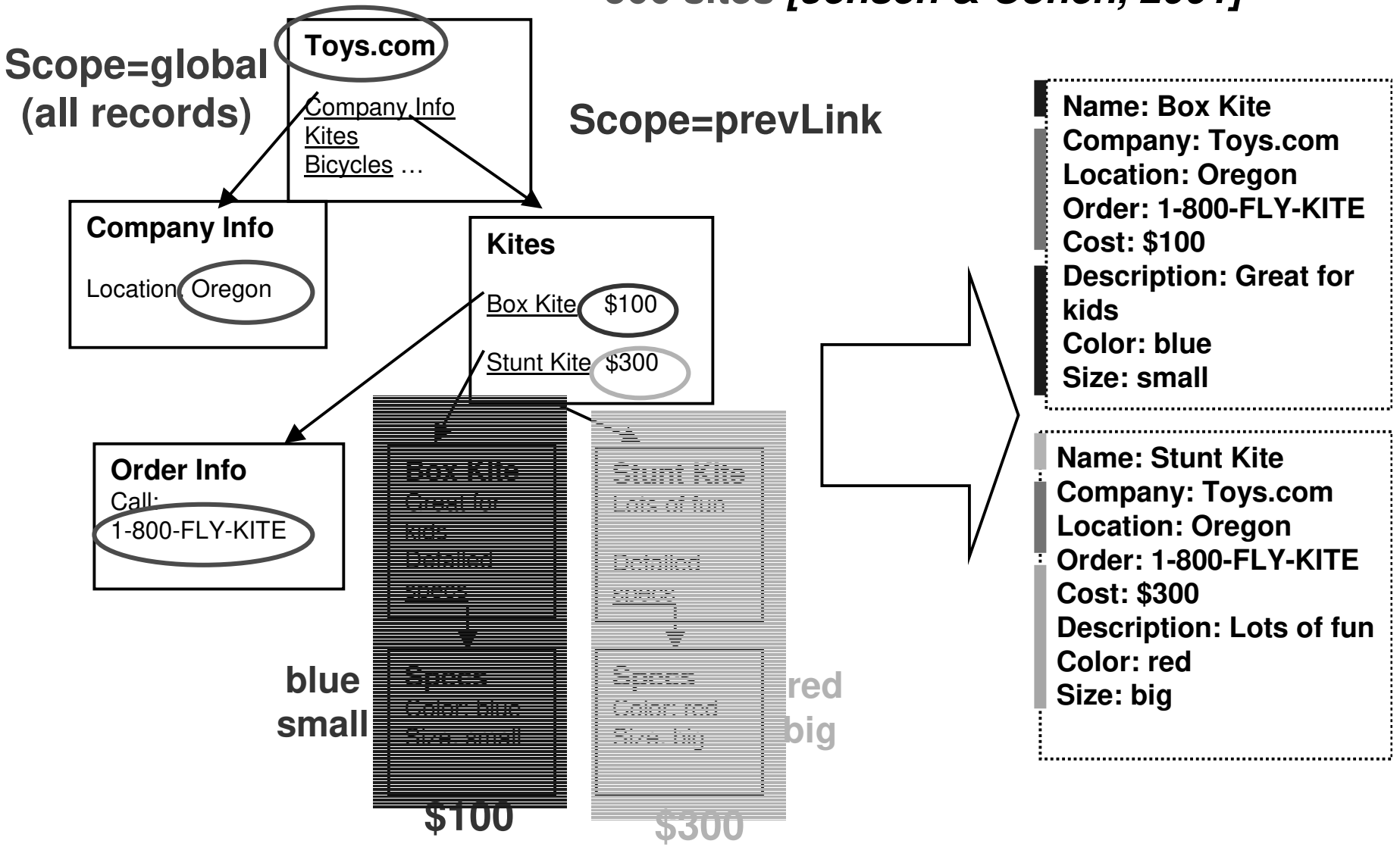
Local language models contribute evidence to relation classification.

Dependencies between classes of entities and relations!

Inference with loopy belief propagation.

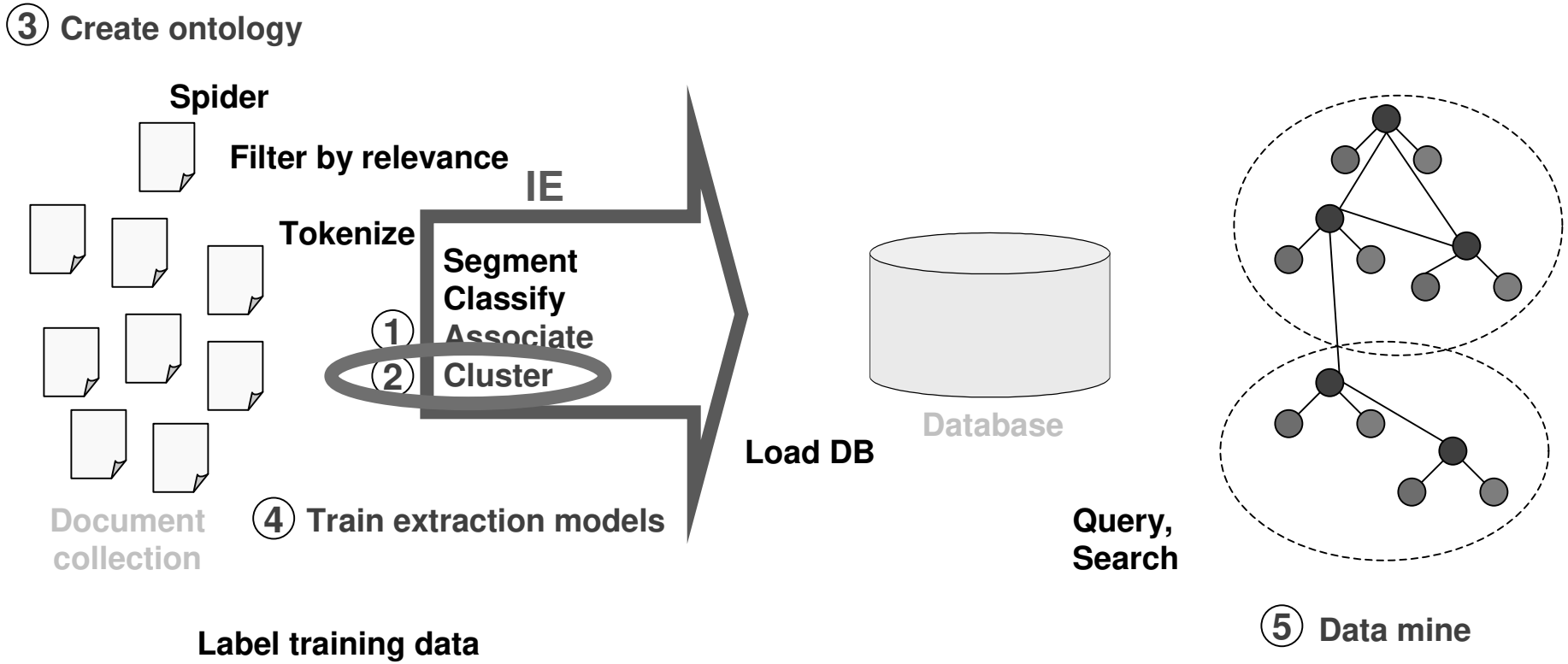
(1) Association of records from the web

5 label types sufficient for modeling 500 sites *[Jensen & Cohen, 2001]*



Broader View

Now touch on some other issues



(2) Clustering for Reference Matching and De-duplication

[Borthwick, 2000]

Learn *Pr ({duplicate, not-duplicate} | record1, record2)* with a Maximum Entropy classifier.

Field Name	Record		Feature	Weight	Prediction
	1	2			
Last name	Smith	Smith	Match	1.153	Merge
First name	Emily	Emely	No-match Soundex	1.350 4.708	No-merge Merge
DOB	[04/28/97]	[04/28/97]	Match	1.138	Merge
Multiple birth	N	N			
Mom's Maiden Name		CRUZ			
Mother's DOB		12/04/76			
Street	4528 3rd Ave	4528 3rd Ave	Match	4.342	Merge
City	Bronx	Bronx	Match	1.103	Merge
State	NY	NY			
Zip	10462	10462	Match	3.013	Merge
Phone	718-123-4567	718-123-6789	No-match	2.130	No-merge
Med Rec Number	11856437503	11856437503	Match	6.587	Merge

$$\text{Merge Total} = 587.2$$

$$\text{No-merge total} = 2.9$$

$$\frac{587.2}{587.2 + 2.9} = 0.995$$

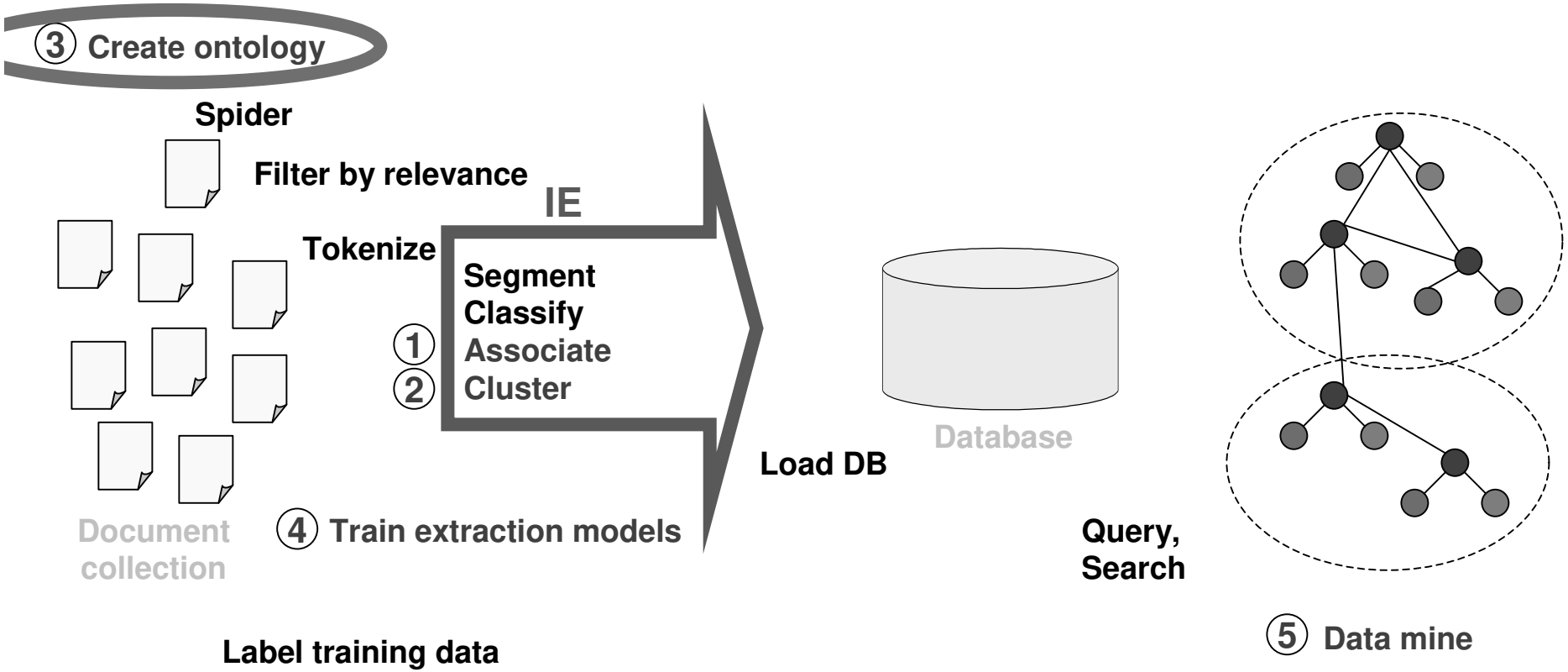
Do greedy agglomerative clustering using this Probability as a distance metric.

(2) Clustering for Reference Matching and De-duplication

- Efficiently clustering large data sets by pre-clustering with a cheap distance metric.
  - *[McCallum, Nigam & Ungar, 2000]*
- Learn a better distance metric.
  - *[Cohen & Richman, 2002]*
- Don't simply merge greedily: capture dependencies among multiple merges.
  - *[Pasula, Marthi, Milch, Russell, Shpitser, NIPS 2002]*

Broader View

Now touch on some other issues



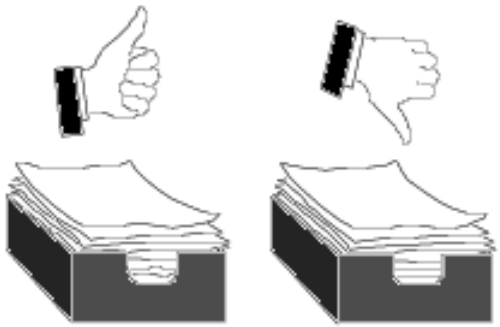
(3) Automatically Inducing an Ontology

[Riloff, '95]

Two inputs:

(1)

preclassified texts



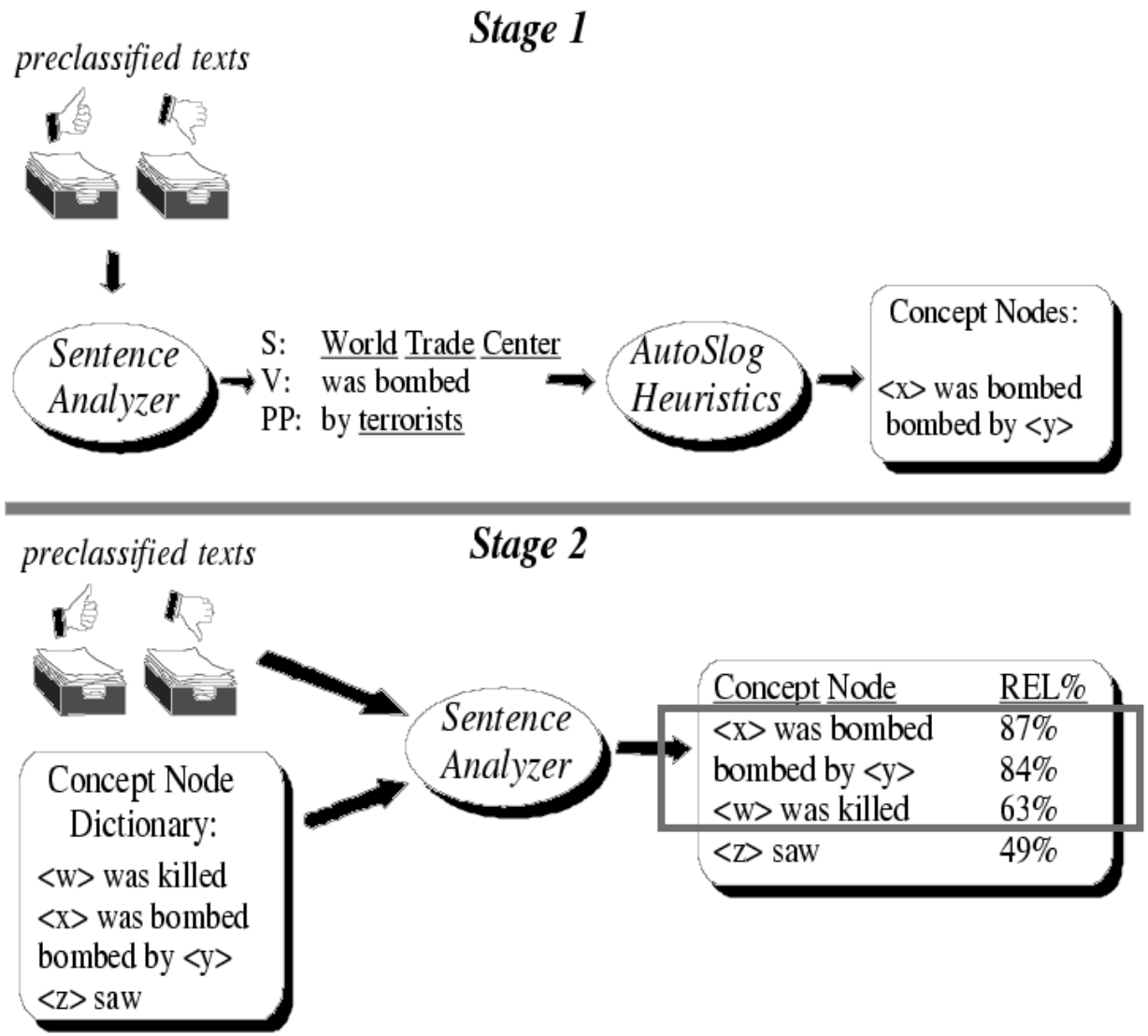
(2)

Heuristic "interesting" meta-patterns.

Linguistic Pattern	Example
1. <subject> active-verb	<perpetrator> bombed
2. <subject> active-verb direct-object[3]	<perpetrator> claimed responsibility
3. <subject> passive-verb	<victim> was murdered
4. <subject> verb infinitive	<perpetrator> attempted to kill
5. <subject> auxiliary noun	<victim> was victim
6. active-verb <direct-object>	bombed <target>
7. passive-verb <direct-object>[4]	killed <victim>
8. infinitive <direct-object>	to kill <victim>
9. verb infinitive <direct-object>	threatened to attack <target>
10. gerund <direct-object>	killing <victim>
11. noun auxiliary <direct-object>	fatality was <victim>
12. noun preposition <noun-phrase>	bomb against <target>
13. active-verb preposition <noun-phrase>	killed with <instrument>
14. passive-verb preposition <noun-phrase>	was aimed at <target>
15. infinitive preposition <noun-phrase>[3]	to fire at <victim>

(3) Automatically Inducing an Ontology

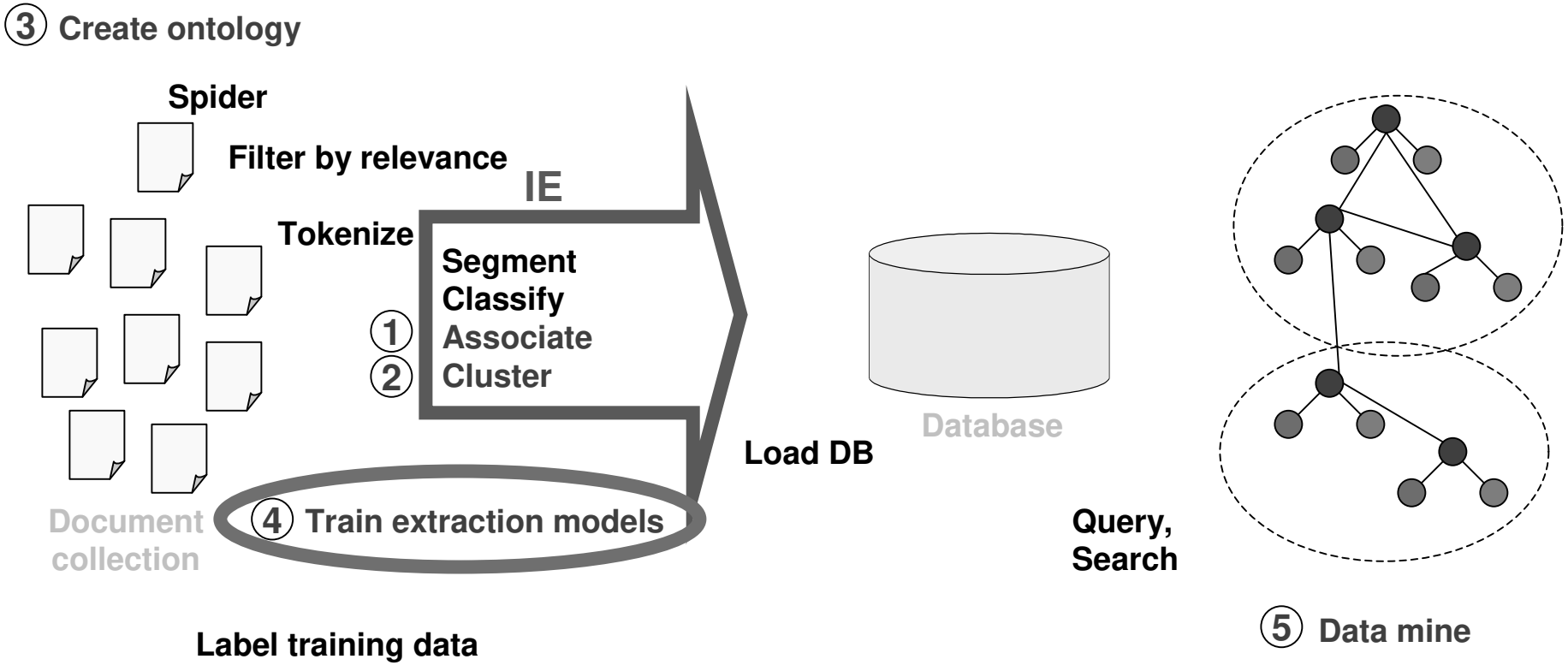
[Riloff, '95]



Subject/Verb/Object patterns that occur more often in the relevant documents than the irrelevant ones.

Broader View

Now touch on some other issues



(4) Training IE Models using Unlabeled Data

[Collins & Singer, 1999]

…says Mr. Cooper, a vice president of …

NNP NNP appositive phrase, head=president

Use two independent sets of features:

Contents: full-string=*Mr._Cooper*, contains(*Mr.*), contains(*Cooper*)
Context: context-type=*appositive*, appositive-head=*president*

1. Start with just seven rules: and ~1M sentences of NYTimes

full-string=New_York	➔ Location
fill-string=California	➔ Location
full-string=U.S.	➔ Location
contains(Mr.)	➔ Person
contains(Incorporated)	➔ Organization
full-string=Microsoft	➔ Organization
full-string=I.B.M.	➔ Organization

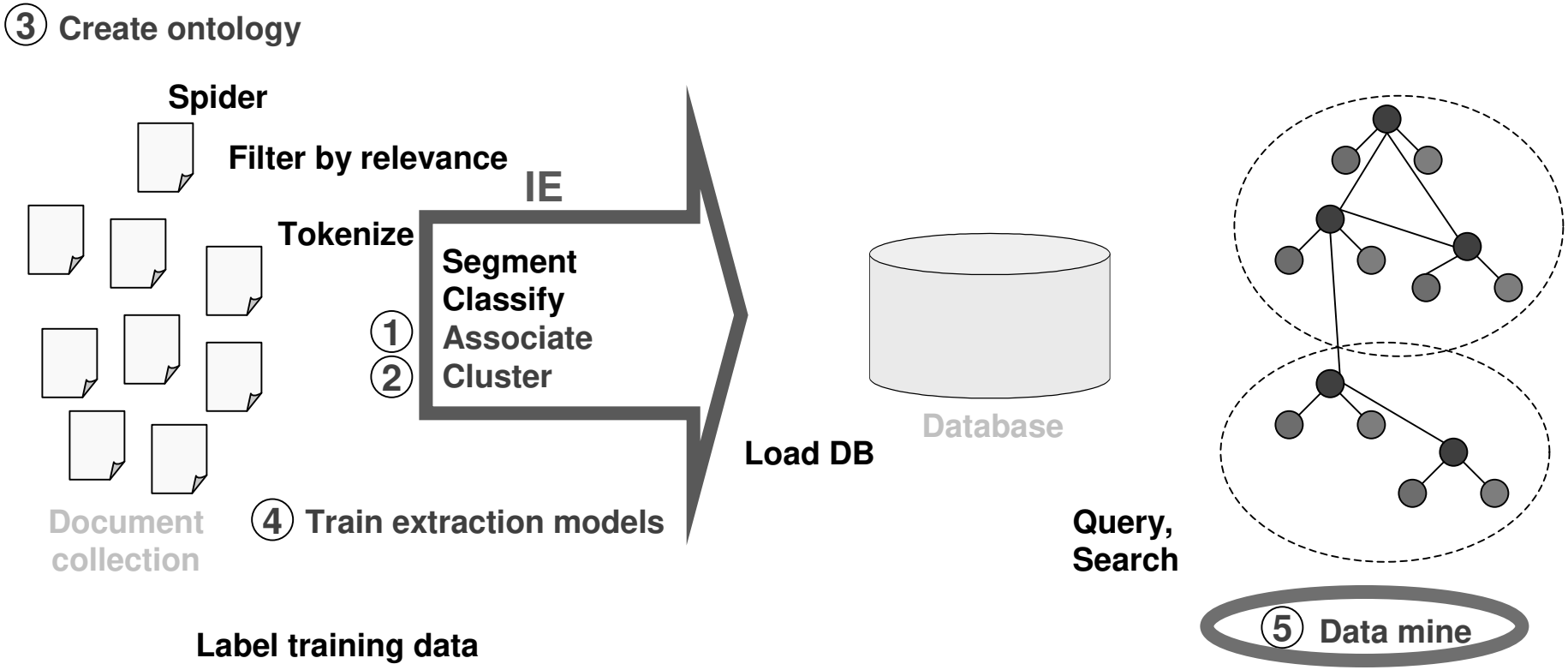
2. Alternately train & label using each feature set.

3. Obtain 83% accuracy at finding *person*, *location*, *organization* & *other* in appositives and prepositional phrases!

See also *[Brin 1998], [Riloff & Jones 1999]*

Broader View

Now touch on some other issues



(5) Data Mining: Working with IE Data

- Some special properties of IE data:
  - It is based on extracted text
  - It is "dirty", (missing extraneous facts, improperly normalized entity names, etc.
  - May need cleaning before use

- What operations can be done on dirty, unnormalized databases?
  - Query it directly with a language that has "soft joins" across similar, but not identical keys. *[Cohen 1998]*
  - Construct features for learners *[Cohen 2000]*
  - Infer a "best" underlying clean database *[Cohen, Kautz, MacAllester, KDD2000]*

(5) Data Mining: Mutually supportive IE and Data Mining

[Nahm & Mooney, 2000]

Extract a large database
Learn rules to predict the value of each field from the other fields.
Use these rules to increase the accuracy of IE.

Example DB record

Filled Job Template
title: Senior DBMS Consultant
salary: Up to $55K
state: TX
city: Dallas
country: US
language: Powerbuilder, Progress, C, C++, Visual Basic
platform: UNIX, NT
application: SQL Server, Oracle
area: Electronic Commerce, Customer Service
required years of experience: 3
desired years of experience: 5
required degree: BS

Sample Learned Rules

platform:AIX & !application:Sybase & application:DB2
➔application:Lotus Notes

language:C++ & language:C & application:Corba & title=SoftwareEngineer
➔ platform:Windows

language:HTML & platform:WindowsNT & application:ActiveServerPages
➔ area:Database

Language:Java & area:ActiveX & area:Graphics
➔ area:Web

Wrap-up

IE Resources

- Data
  - RISE, **http://www.isi.edu/~muslea/RISE/index.html**
  - Linguistic Data Consortium (LDC)
    - Penn Treebank, Named Entities, Relations, etc.
  - **http://www.biostat.wisc.edu/~craven/ie**
  - **http://www.cs.umass.edu/~mccallum/data**
- Code
  - TextPro, **http://www.ai.sri.com/~appelt/TextPro**
  - MALLET, **http://www.cs.umass.edu/~mccallum/mallet**
- Both
  - **http://www.cis.upenn.edu/~adwait/penntools.html**
  - **http://www.cs.umass.edu/~mccallum/ie**

Where from Here?

- Science
  - Higher accuracy, integration with IE's consumers.
  - Scoped Learning, Minimizing labeled data needs, unified models of all four of IE's components.
  - Multi-modal IE: text, images, video, audio. Multi-lingual.
- Profit
  - SRA, Inxight, Fetch, Mohomine, Cymfony,… *you?*
  - Bio-informatics, Intelligent Tutors, Information Overload, Anti-terrorism
- Fun
  - Search engines that return "things" instead of "pages" (people, companies, products, universities, courses…)
  - New insights by mining previously untapped knowledge.

References

- [Bikel et al 1997] Bikel, D.; Miller, S.; Schwartz, R.; and Weischedel, R. Nymble: a high-performance learning name-finder. In *Proceedings of ANLP'97*, p194-201.
- [Califf & Mooney 1999], Califf, M.E.; Mooney, R.: Relational Learning of Pattern-Match Rules for Information Extraction, in *Proceedings of the Sixteenth National Conference on Artificial Intelligence* (AAAI-99).
- [Cohen, Hurst, Jensen, 2002] Cohen, W.; Hurst, M.; Jensen, L.: A flexible learning system for wrapping tables and lists in HTML documents. *Proceedings of The Eleventh International World Wide Web Conference* (WWW-2002)
- [Cohen, Kautz, McAllester 2000] Cohen, W; Kautz, H.; McAllester, D.: Hardening soft information sources. *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining* (KDD-2000).
- [Cohen, 1998] Cohen, W.: Integration of Heterogeneous Databases Without Common Domains Using Queries Based on Textual Similarity, in *Proceedings of ACM SIGMOD-98*.
- [Cohen, 2000a] Cohen, W.: Data Integration using Similarity Joins and a Word-based Information Representation Language, *ACM Transactions on Information Systems*, 18(3).
- [Cohen, 2000b] Cohen, W. Automatically Extracting Features for Concept Learning from the Web, *Machine Learning: Proceedings of the Seventeeth International Conference (ML-2000).*
- [Collins & Singer 1999] Collins, M.; and Singer, Y. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.
- [De Jong 1982] De Jong, G. An Overview of the FRUMP System. In: Lehnert, W. & Ringle, M. H. (eds), *Strategies for Natural Language Processing*. Larence Erlbaum, 1982, 149-176.
- [Freitag 98] Freitag, D: Information extraction from HTML: application of a general machine learning approach, *Proceedings of the Fifteenth National Conference on Artificial Intelligence* (AAAI-98).
- [Freitag, 1999], Freitag, D. *Machine Learning for Information Extraction in Informal Domains*. Ph.D. dissertation, Carnegie Mellon University.
- [Freitag 2000], Freitag, D: Machine Learning for Information Extraction in Informal Domains, *Machine Learning* 39(2/3): 99-101 (2000).
- Freitag & Kushmerick, 1999] Freitag, D; Kushmerick, D.: Boosted Wrapper Induction. *Proceedings of the Sixteenth National Conference on Artificial Intelligence* (AAAI-99)
- [Freitag & McCallum 1999] Freitag, D. and McCallum, A. Information extraction using HMMs and shrinakge. In Proceedings AAAI-99 Workshop on Machine Learning for Information Extraction. AAAI Technical Report WS-99-11.
- [Kushmerick, 2000] Kushmerick, N: Wrapper Induction: efficiency and expressiveness, *Artificial Intelligence*, 118(pp 15-68).
- [Lafferty, McCallum & Pereira 2001] Lafferty, J.; McCallum, A.; and Pereira, F., Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *In Proceedings of ICML-2001*.
- [Leek 1997] Leek, T. R. *Information extraction using hidden Markov models*. Master's thesis. UC San Diego.
- [McCallum, Freitag & Pereira 2000] McCallum, A.; Freitag, D.; and Pereira. F., Maximum entropy Markov models for information extraction and segmentation, In *Proceedings of ICML-2000*
- [Miller et al 2000] Miller, S.; Fox, H.; Ramshaw, L.; Weischedel, R. A Novel Use of Statistical Parsing to Extract Information from Text. *Proceedings of the 1st Annual Meeting of the North American Chapter of the ACL (NAACL),* p. 226 - 233.

References

- [Muslea et al, 1999] Muslea, I.; Minton, S.; Knoblock, C. A.: *A Hierarchical Approach to Wrapper Induction*. Proceedings of Autonomous Agents-99.
- [Muslea et al, 2000] Musclea, I.; Minton, S.; and Knoblock, C. Hierarhical wrapper induction for semistructured information sources. *Journal of Autonomous Agents and Multi-Agent Systems*.
- [Nahm & Mooney, 2000] Nahm, Y.; and Mooney, R. A mutually beneficial integration of data mining and information extraction. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence*, pages 627--632, Austin, TX.
- [Punyakanok & Roth 2001] Punyakanok, V.; and Roth, D. The use of classifiers in sequential inference. *Advances in Neural Information Processing Systems 13*.
- [Ratnaparkhi 1996] Ratnaparkhi, A., A maximum entropy part-of-speech tagger, in *Proc. Empirical Methods in Natural Language Processing Conference*, p133-141.
- [Ray & Craven 2001] *Ray, S.; and Craven, Ml.* Representing Sentence Structure in Hidden Markov Models for Information Extraction. *Proceedings of the 17th International Joint Conference on Artificial Intelligence,* Seattle, WA. Morgan Kaufmann.
- [Soderland 1997]: Soderland, S.: Learning to Extract Text-Based Information from the World Wide Web. *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (KDD-97).
- [Soderland 1999] Soderland, S. Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34(1/3):233-277.