Chapter 8: Information Extraction (IE)

8.1 Motivation and Overview

8.2 Rule-based IE

8.3 Hidden Markov Models (HMMs) for IE

8.4 Linguistic IE

8.5 Entity Reconciliation

8.6 IE for Knowledge Acquisition

8.1 Motivation and Overview

Goals:

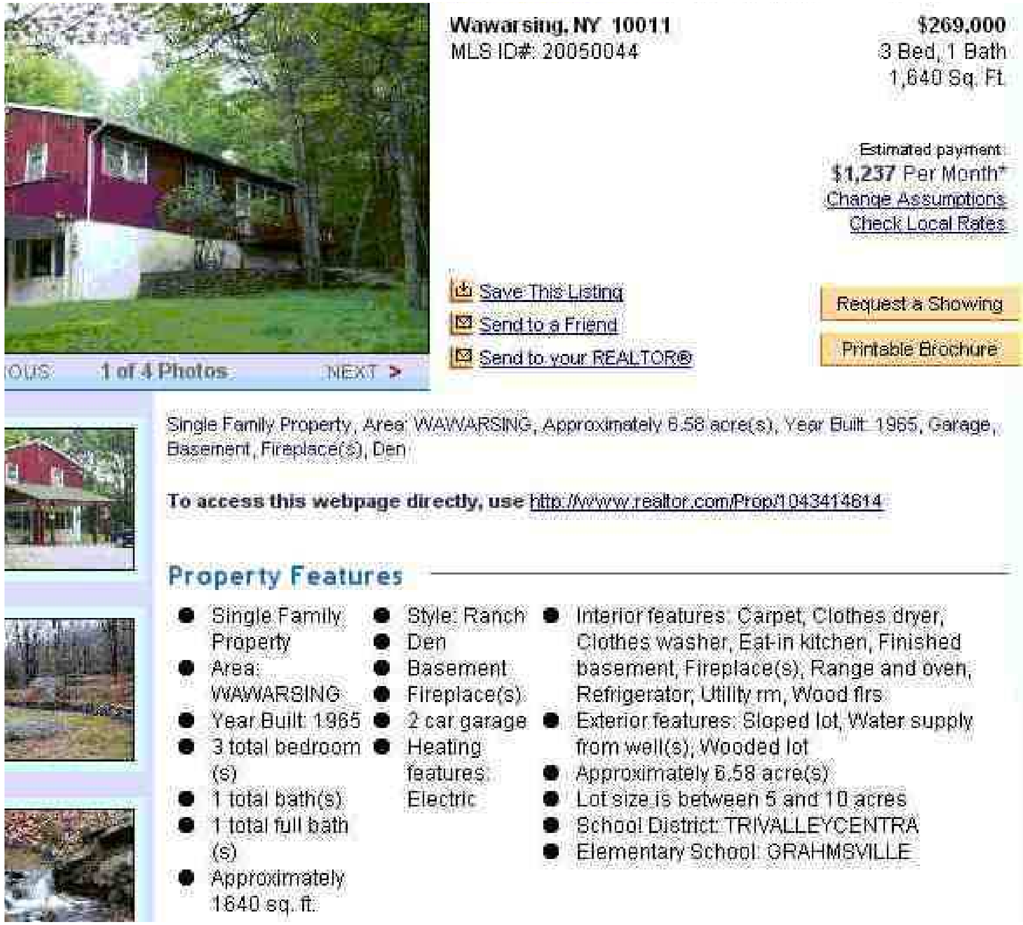
- **annotate text documents or Web pages (named entity recognition, html2xml, etc.)**
- **extract facts from text documents or Web pages (relation learning)**
- **find facts on the Web (or in Wikipedia) to populate thesaurus/ontology relations**
- **information enrichment (e.g. for business analytics)**

Technologies:

- **NLP (PoS tagging, chunk parsing, etc.)**
- **Pattern matching & rule learning (regular expressions, FSAs)**
- **Statistical learning (HMMs, MRFs, etc.)**
- **Lexicon lookups (name dictionaries, geo gazetteers, etc.)**
- **Text mining in general**

“Semantic” Data Production

Most data is (exposed as) HTML (or PDF or RSS or ...) or comes from data sources with unknown schema



Wawarsing, NY 10011
MLS ID#: 20050044

$269,000
3 Bed, 1 Bath
1,640 Sq. Ft.

Estimated payment:
$1,237 Per Month*
Change Assumptions
Check Local Rates

Save This Listing
Send to a Friend
Send to your REALTOR®

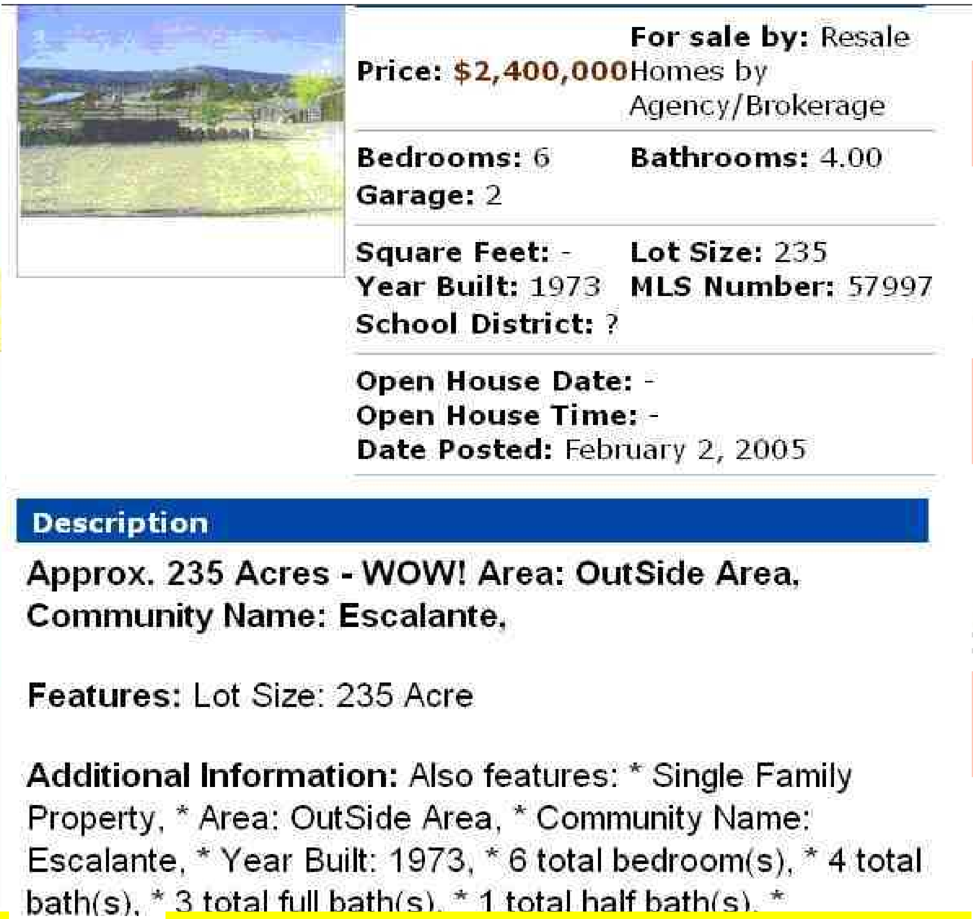
Request a Showing
Printable Brochure

Single Family Property, Area: WAWARSING, Approximately 6.58 acre(s), Year Built: 1965, Garage, Basement, Fireplace(s), Den

To access this webpage directly, use http://www.realtor.com/Prop/1043414614

Property Features

- Single Family Property
- Area: WAWARSING
- Year Built: 1965
- 3 total bedroom (s)
- 1 total bath(s)
- 1 total full bath (s)
- Approximately 1640 sq. ft.
- Style: Ranch
- Den
- Basement
- Fireplace(s)
- 2 car garage
- Heating features: Electric
- Interior features: Carpet, Clothes dryer, Clothes washer, Eat-in kitchen, Finished basement, Fireplace(s), Range and oven, Refrigerator, Utility rm, Wood flrs
- Exterior features: Sloped lot, Water supply from well(s), Wooded lot
- Approximately 6.58 acre(s)
- Lot size is between 5 and 10 acres
- School District: TRIVALLEYCENTRA
- Elementary School: GRAHMSVILLE



Price: **$2,400,000**
For sale by: Resale Homes by Agency/Brokerage

Bedrooms: 6
Bathrooms: 4.00
Garage: 2

Square Feet: -
Lot Size: 235
Year Built: 1973
MLS Number: 57997
School District: ?

Open House Date: -
Open House Time: -
Date Posted: February 2, 2005

Description

Approx. 235 Acres - WOW! Area: OutSide Area. Community Name: Escalante.

Features: Lot Size: 235 Acre

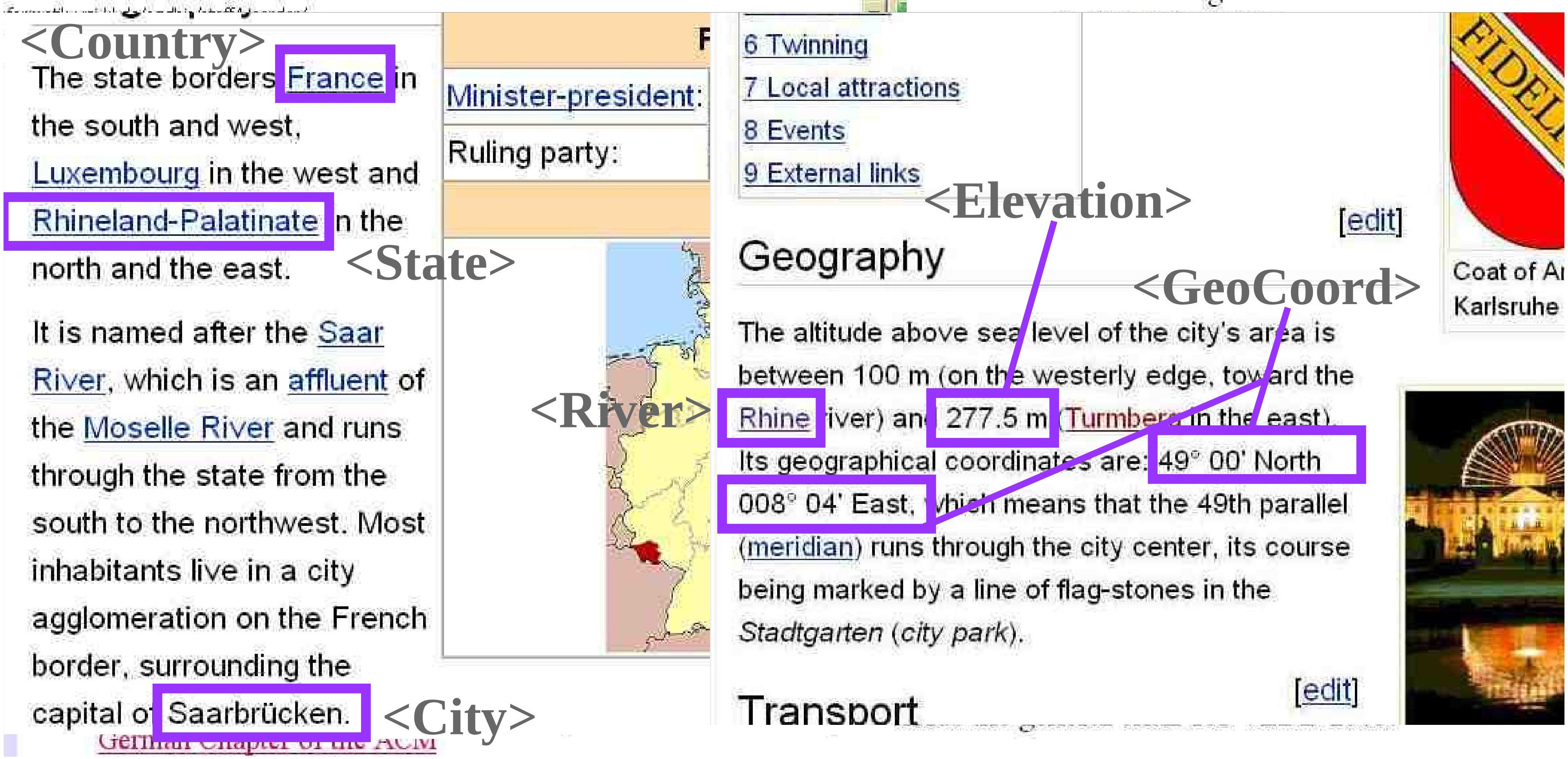
Additional Information: Also features: * Single Family Property, * Area: OutSide Area, * Community Name: Escalante, * Year Built: 1973, * 6 total bedroom(s), * 4 total bath(s), * 3 total full bath(s), * 1 total half bath(s), *

accessible by wrappers
(or, perhaps, Web Service)
→ rules, FSAs (reg. expr.), ...

what about „free-form“ data?
→ HMMs, MRFs, ...

“Semantic” Data Production

Most data is (exposed as) HTML (or PDF or RSS or ...) or comes from data sources with unknown schema



The state borders France in the south and west, Luxembourg in the west and Rhineland-Palatinate in the north and the east.

It is named after the Saar River, which is an affluent of the Moselle River and runs through the state from the south to the northwest. Most inhabitants live in a city agglomeration on the French border, surrounding the capital of Saarbrücken.

German Chapter of the ACM

Minister-president:

Ruling party:

6 Twinning

7 Local attractions

8 Events

9 External links

[edit]

Geography

The altitude above sea level of the city's area is between 100 m (on the westerly edge, toward the Rhine river) and 277.5 m (Turmberg in the east). Its geographical coordinates are: 49° 00' North 008° 04' East, which means that the 49th parallel (meridian) runs through the city center, its course being marked by a line of flag-stones in the *Stadtgarten* (*city park*).

[edit]

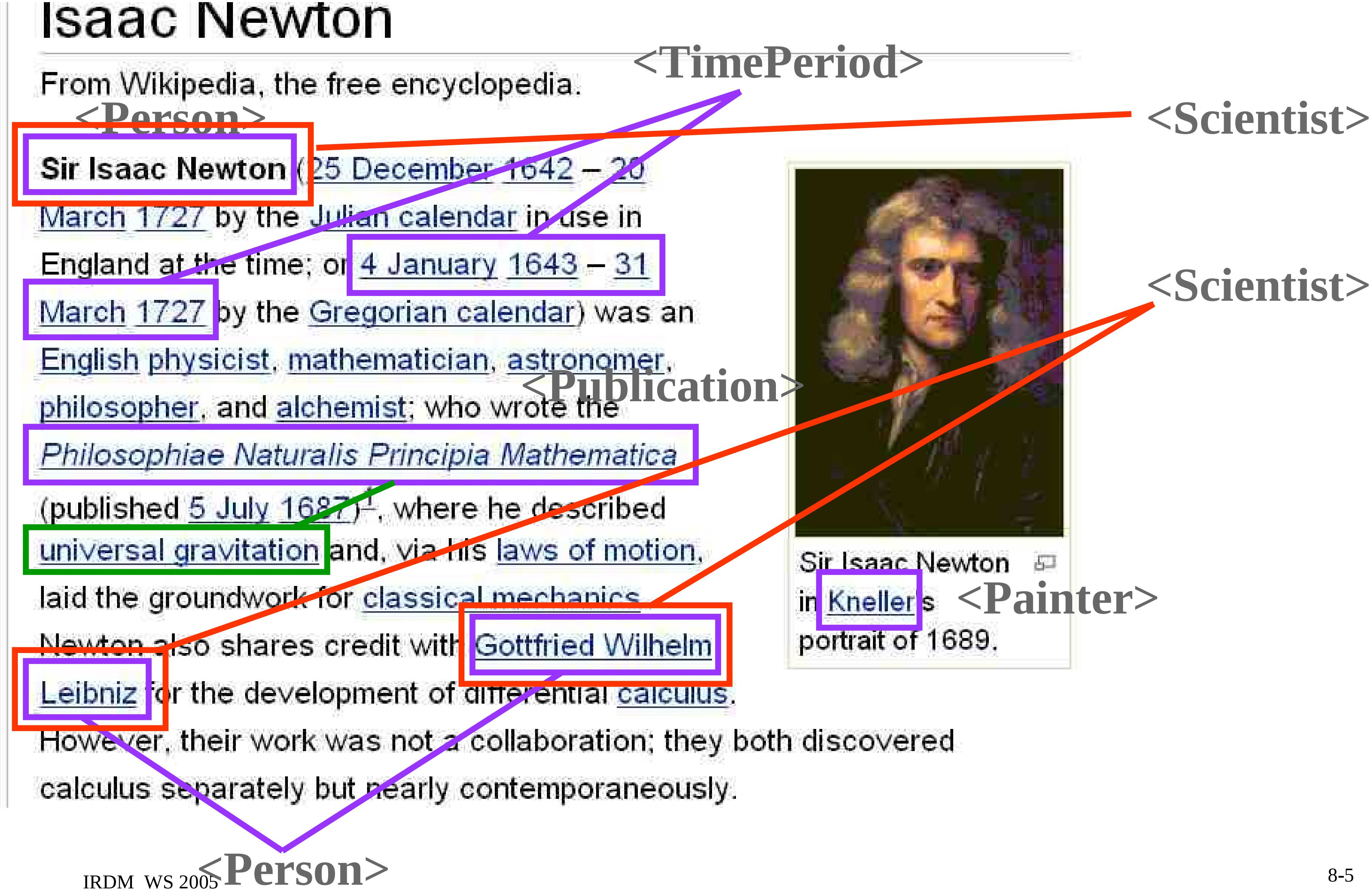
Transport

Coat of Ar

Karlsruhe

“Semantic” Data Production

Most data is (exposed as) HTML (or PDF or RSS or ...) or comes from data sources with unknown schema



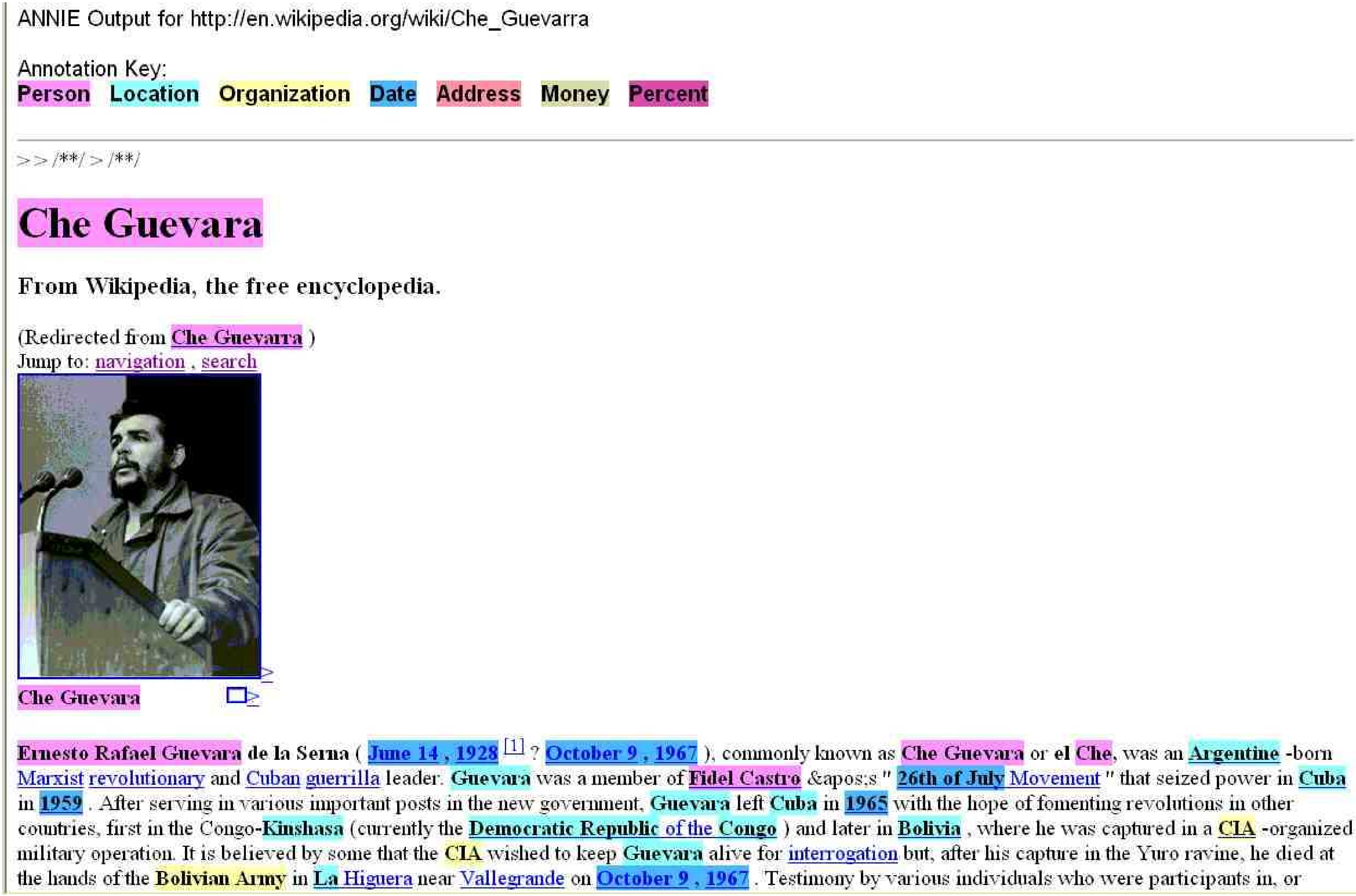
Isaac Newton

From Wikipedia, the free encyclopedia.

Sir Isaac Newton (25 December 1642 – 20 March 1727 by the Julian calendar in use in England at the time; or 4 January 1643 – 31 March 1727 by the Gregorian calendar) was an English physicist, mathematician, astronomer, philosopher, and alchemist; who wrote the *Philosophiae Naturalis Principia Mathematica* (published 5 July 1687)[1], where he described universal gravitation and, via his laws of motion, laid the groundwork for classical mechanics. Newton also shares credit with Gottfried Wilhelm Leibniz for the development of differential calculus. However, their work was not a collaboration; they both discovered calculus separately but nearly contemporaneously.

Sir Isaac Newton in Kneller's portrait of 1689.

NLP-based IE from Web Pages



ANNIE Output for http://en.wikipedia.org/wiki/Che_Guevarra

Annotation Key:
Person Location Organization Date Address Money Percent

>> /**/ > /**/

Che Guevara

From Wikipedia, the free encyclopedia.

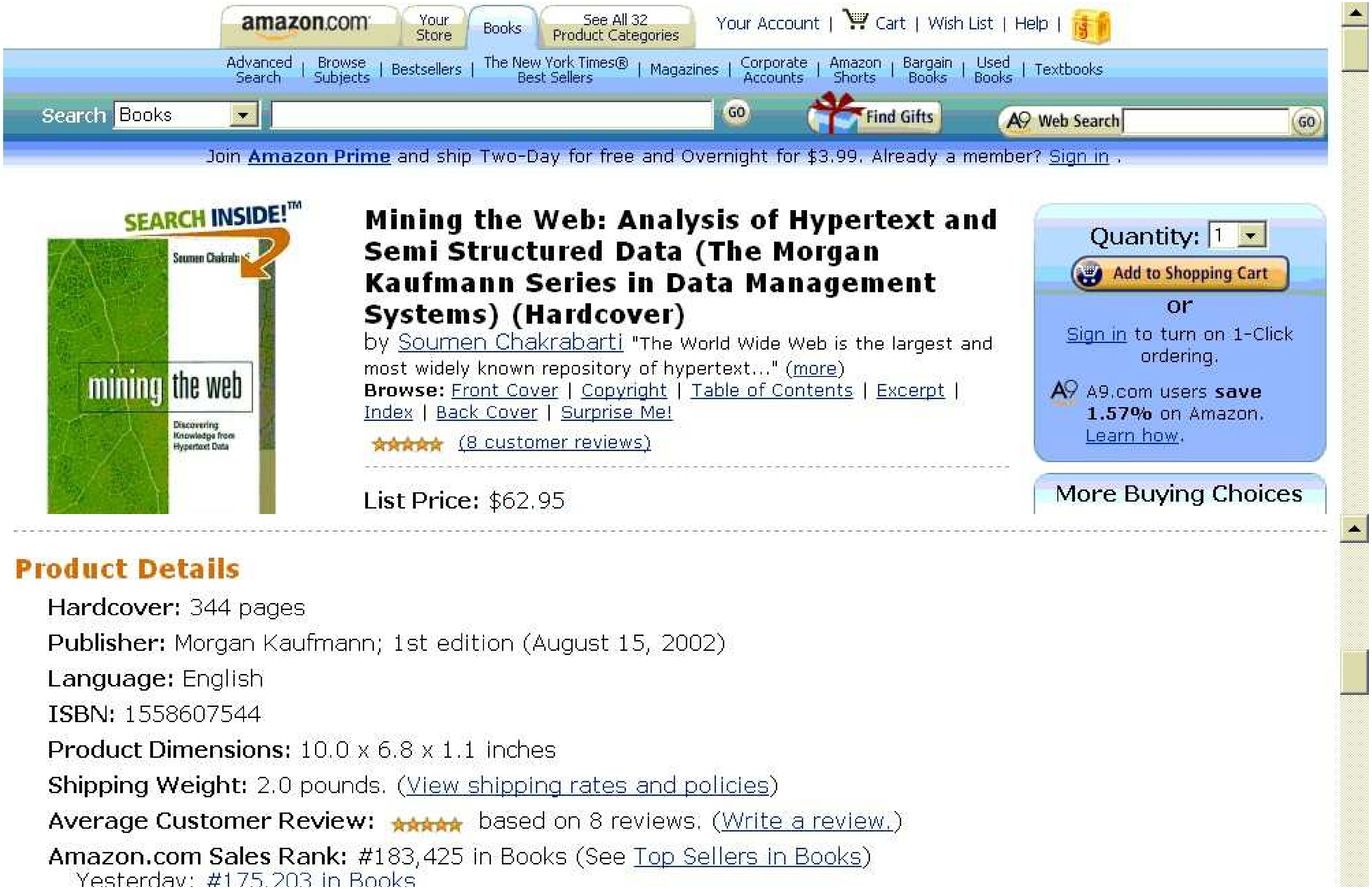
(Redirected from Che Guevarra)
Jump to: navigation , search

Che Guevara

Ernesto Rafael Guevara de la Serna (June 14 , 1928 [1] ? October 9 , 1967), commonly known as Che Guevara or el Che, was an Argentine -born Marxist revolutionary and Cuban guerrilla leader. Guevara was a member of Fidel Castro &apos;s " 26th of July Movement " that seized power in Cuba in 1959 . After serving in various important posts in the new government, Guevara left Cuba in 1965 with the hope of fomenting revolutions in other countries, first in the Congo-Kinshasa (currently the Democratic Republic of the Congo) and later in Bolivia , where he was captured in a CIA -organized military operation. It is believed by some that the CIA wished to keep Guevara alive for interrogation but, after his capture in the Yuro ravine, he died at the hands of the Bolivian Army in La Higuera near Vallegrande on October 9 , 1967 . Testimony by various individuals who were participants in, or

Leading open-source tool: GATE/ANNIE
http://www.gate.ac.uk/annie/

Extracting Structured Records from Deep Web Source (1)


amazon.com
Your Store
Books
See All 32 Product Categories
Your Account | Cart | Wish List | Help |
Advanced Search | Browse Subjects | Bestsellers | The New York Times® Best Sellers | Magazines | Corporate Accounts | Amazon Shorts | Bargain Books | Used Books | Textbooks
Search Books
GO
Find Gifts
A9 Web Search
GO
SEARCH INSIDE!™
Soumen Chakrabarti
mining the web
Discovering Knowledge from Hypertext Data
Quantity: 1
Add to Shopping Cart
or
More Buying Choices

Join Amazon Prime and ship Two-Day for free and Overnight for $3.99. Already a member? Sign in .

Mining the Web: Analysis of Hypertext and Semi Structured Data (The Morgan Kaufmann Series in Data Management Systems) (Hardcover)

by Soumen Chakrabarti "The World Wide Web is the largest and most widely known repository of hypertext..." (more)

Browse: Front Cover | Copyright | Table of Contents | Excerpt | Index | Back Cover | Surprise Me!

★★★★★ (8 customer reviews)

List Price: $62.95

Sign in to turn on 1-Click ordering.

A9.com users **save 1.57%** on Amazon. Learn how.

Product Details

Hardcover: 344 pages

Publisher: Morgan Kaufmann; 1st edition (August 15, 2002)

Language: English

ISBN: 1558607544

Product Dimensions: 10.0 x 6.8 x 1.1 inches

Shipping Weight: 2.0 pounds. (View shipping rates and policies)

Average Customer Review: ★★★★★ based on 8 reviews. (Write a review.)

Amazon.com Sales Rank: #183,425 in Books (See Top Sellers in Books)

Yesterday: #175,203 in Books

Extracting Structured Records from Deep Web Source (2)

```
<div class="buying"><b class="sans">Mining the Web: Analysis of Hypertext and
Semi Structured Data (The Morgan Kaufmann Series in Data Management Systems)
(Hardcover)</b><br />by

<a href="/exec/obidos/search-handle-url/index=books&field-author-exact=Soumen%20Chak
5490548">Soumen Chakrabarti</a>

<div class="buying" id="priceBlock">
<style type="text/css">
  td.productLabel { font-weight: bold; text-align: right; white-space: nowrap; vertical-align: t
  table.product   { border: 0px; padding: 0px; border-collapse: collapse; }
</style>
<table class="product">
  <tr>
    <td class="productLabel">List Price:</td>
    <td>$62.95</td>
  </tr>
  <tr>
    <td class="productLabel">Price:</td>
    <td><b class="price">$62.95</b>
& this item ships for <b>FREE with Super Saver Shipping</b>.


...
```

Extracting Structured Records from Deep Web Source (3)

```
<a name="productDetails" id="productDetails"></a>
<hr noshade="noshade" size="1" class="bucketDivider
<table cellpadding="0" cellspacing="0" border="0">
  <tr>
    <td class="bucket">
<b class="h1">Product Details</b><br />
  <div class="content">
<ul>
<li><b>Hardcover:</b> 344 pages</li>
<li><b>Publisher:</b> Morgan Kaufmann; 1st edition (
<li><b>Language:</b> English</li>
<li><b>ISBN:</b> 1558607544</li>
<li><b>Product Dimensions:</b> 10.0 x 6.8 x 1.1 inches<
<li><b>Shipping Weight:</b> 2.0 pounds. (<a href="htt
shipping rates and policies</a>)</li>
<li><b>Average Customer Review:</b> <img src="http
border="0" /> based on 8 reviews.
 (<a href="http://www.amazon.com/gp/customer-reviews/write-a-review.html/102-8395894-5
<li>
<b>Amazon.com Sales Rank:</b>  #183,425 in Books (See <a href="/exec/obidos/tg/new-for-
```

extract record:

Title: Mining the Web: Analysi
Author: Soumen Chakrabarti,
Hardcover: 344 pages,
Publisher: Morgan Kaufmann,
Language: English,
ISBN: 1558607544.
...
AverageCustomerReview: 4
NumberOfReviews: 8,
SalesRank: 183425
...

IE Applications

- Comparison shopping & recommendation portals
  e.g. consumer electronics, used cars, real estate, pharmacy, etc.
- Business analytics on customer dossiers, financial reports, etc.
  e.g.: How was company X (the market Y) performing in the last 5 years?
- Market/customer, PR impact, and media coverage analyses
  e.g.: How are our products perceived by teenagers (girls)?
  How good (and positive?) is the press coverage of X vs. Y?
  Who are the stakeholders in a public dispute on a planned airport?
- Job brokering (applications/resumes, job offers)
  e.g.: Ho well does the candidate match the desired profile?

- Knowledge management in consulting companies
  e.g.: Do we have experience and competence on X, Y, and Z in Brazil?
- Mining E-mail archives
  e.g.: Who knew about the scandal on X before it became public?
- Knowledge extraction from scientific literature
  e.g.: Which anti-HIV drugs have been found ineffective in recent papers?
- General-purpose knowledge acquisition
  Can we learn encyclopedic knowledge from text & Web corpora?

IE Viewpoints and Approaches

IE as learning (restricted) regular expressions
(wrapping pages with common structure from Deep-Web source)

IE as learning relations
(rules for identifying instances of n-ary relations)

IE as learning fact boundaries

IE as learning text/sequence segmentation (HMMs etc.)

IE as learning contextual patterns (graph models etc.)

IE as natural-language analysis (NLP methods)

IE as large-scale text mining for knowledge acquisition
(combination of tools incl. Web queries)

IE Quality Assessment

fix IE task (e.g. extracting all book records
from a set of bookseller Web pages)
manually extract all correct records

now use standard IR measures:

- precision
- recall
- F1 measure

benchmark settings:

- MUC (Message Understanding Conference), no longer active
- ACE (Automatic Content Extraction), http://www.nist.gov/speech/tests/ace/
- TREC Enterprise Track, http://trec.nist.gov/tracks.html
- Enron e-mail mining, http://www.cs.cmu.edu/~enron

Landscape of IE Tasks and Methods

next 6 slides are from:

William W. Cohen:
Information Extraction and Integration: an Overview,
Tutorial Slides,
http://www.cs.cmu.edu/~wcohen/ie-survey.ppt

IE is different in different domains!

Example: on web there is less grammar, but more formatting & linking

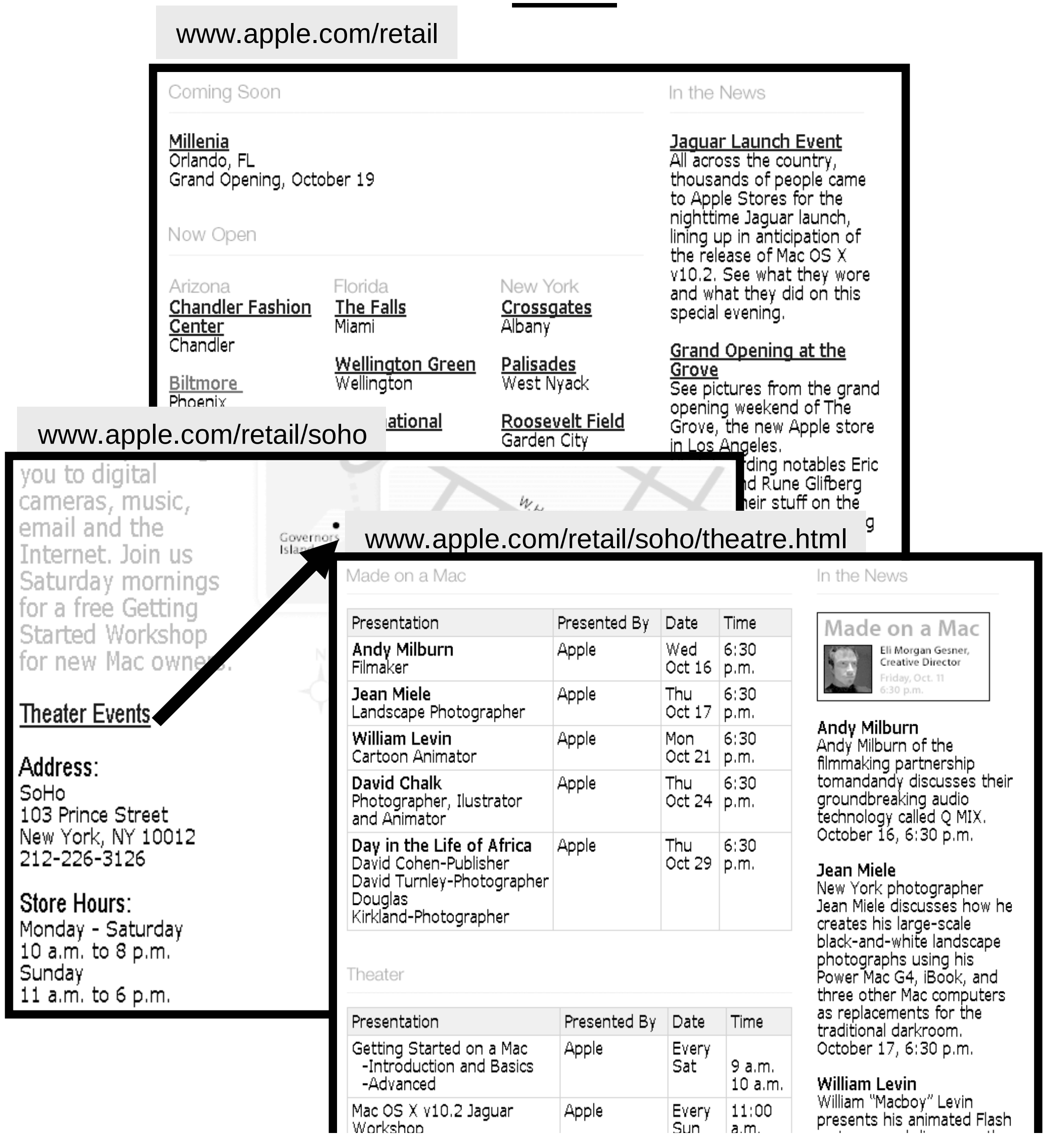
Newswire

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."

Web



www.apple.com/retail

Coming Soon

Millenia
Orlando, FL
Grand Opening, October 19

Now Open

Arizona
Chandler Fashion Center
Chandler

Biltmore
Phoenix

Florida
The Falls
Miami

Wellington Green
Wellington

ational

New York
Crossgates
Albany

Palisades
West Nyack

Roosevelt Field
Garden City

In the News

Jaguar Launch Event
All across the country, thousands of people came to Apple Stores for the nighttime Jaguar launch, lining up in anticipation of the release of Mac OS X v10.2. See what they wore and what they did on this special evening.

Grand Opening at the Grove
See pictures from the grand opening weekend of The Grove, the new Apple store in Los Angeles.
rding notables Eric
d Rune Glifberg
heir stuff on the
g

www.apple.com/retail/soho

you to digital cameras, music, email and the Internet. Join us Saturday mornings for a free Getting Started Workshop for new Mac owners.

Theater Events

Address:
SoHo
103 Prince Street
New York, NY 10012
212-226-3126

Store Hours:
Monday - Saturday
10 a.m. to 8 p.m.
Sunday
11 a.m. to 6 p.m.

www.apple.com/retail/soho/theatre.html

Made on a Mac

Presentation	Presented By	Date	Time
Andy Milburn Filmaker	Apple	Wed Oct 16	6:30 p.m.
Jean Miele Landscape Photographer	Apple	Thu Oct 17	6:30 p.m.
William Levin Cartoon Animator	Apple	Mon Oct 21	6:30 p.m.
David Chalk Photographer, Ilustrator and Animator	Apple	Thu Oct 24	6:30 p.m.
Day in the Life of Africa David Cohen-Publisher David Turnley-Photographer Douglas Kirkland-Photographer	Apple	Thu Oct 29	6:30 p.m.

Theater

Presentation	Presented By	Date	Time
Getting Started on a Mac -Introduction and Basics -Advanced	Apple	Every Sat	9 a.m. 10 a.m.
Mac OS X v10.2 Jaguar Workshop	Apple	Every Sun	11:00 a.m.

In the News

Made on a Mac
Eli Morgan Gesner, Creative Director
Friday, Oct. 11
6:30 p.m.

Andy Milburn
Andy Milburn of the filmmaking partnership tomandandy discusses their groundbreaking audio technology called Q MIX.
October 16, 6:30 p.m.

Jean Miele
New York photographer Jean Miele discusses how he creates his large-scale black-and-white landscape photographs using his Power Mac G4, iBook, and three other Mac computers as replacements for the traditional darkroom.
October 17, 6:30 p.m.

William Levin
William "Macboy" Levin presents his animated Flash

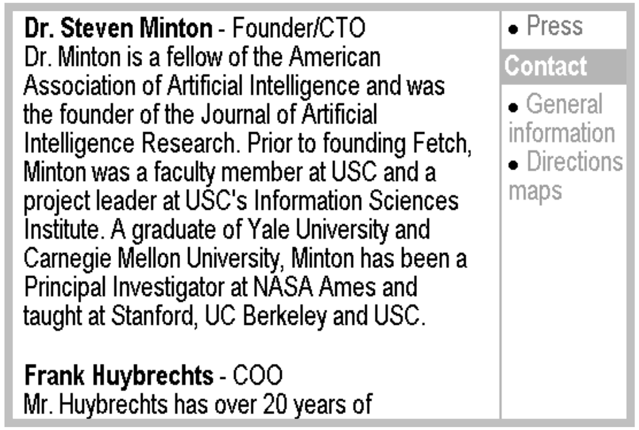
The directory structure, link structure, formatting & layout of the Web is its own new grammar.

Landscape of IE Tasks (1/4): Degree of Formatting

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.

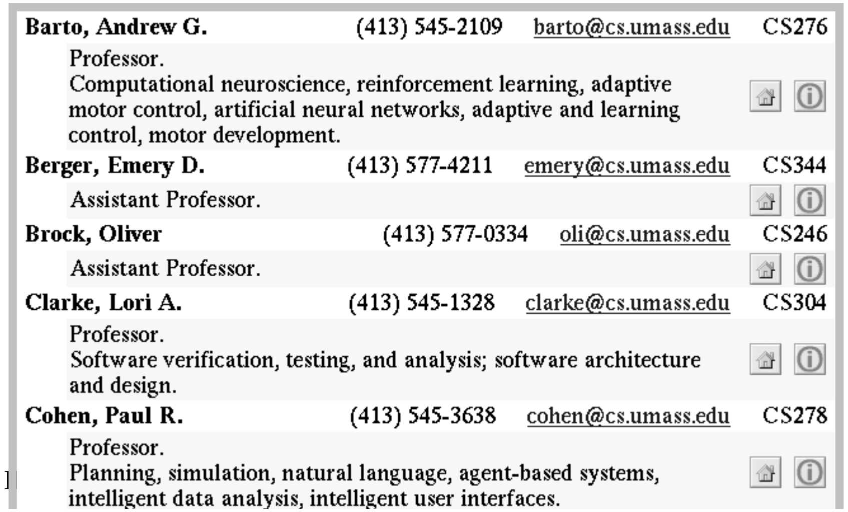
Grammatical sentences and some formatting & links



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.

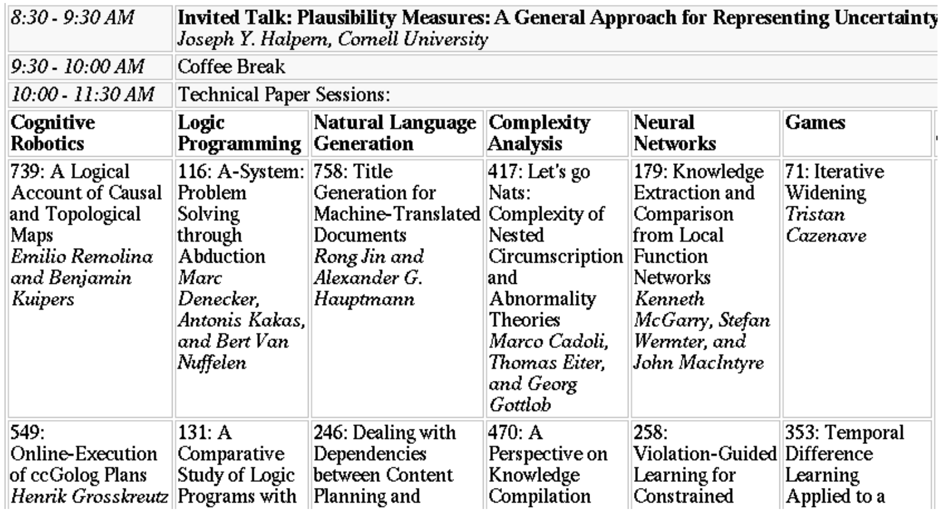
Frank Huybrechts - COO
Mr. Huybrechts has over 20 years of

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.			

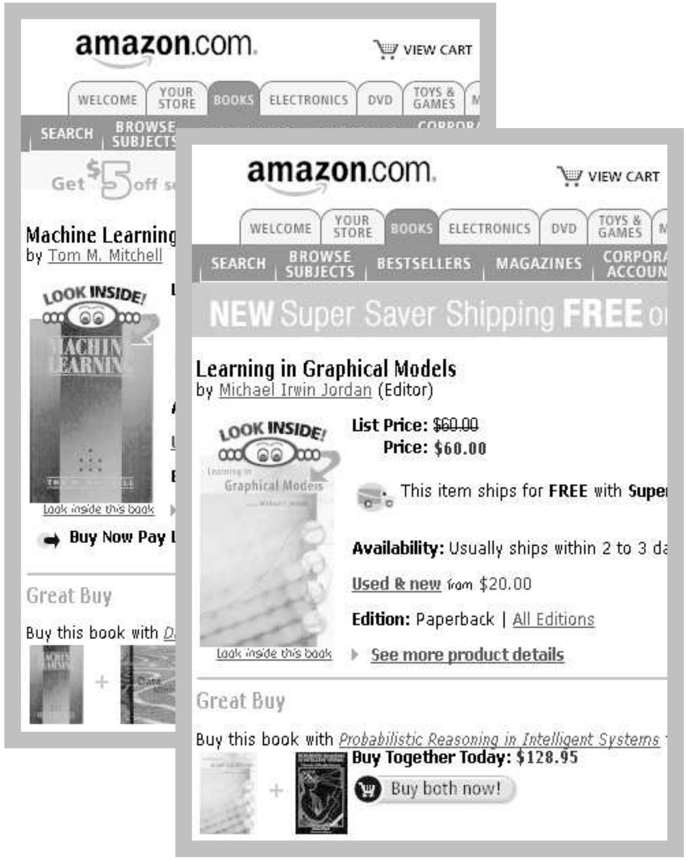
Tables



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): Intended Breadth of Coverage

Web site specific	**Genre specific**	**Wide, non-specific**
Formatting	Layout	Language
Amazon.com Book Pages	Resumes	University Names



Machine Learning
by Tom M. Mitchell

Buy Now Pay

Great Buy

Buy this book with

Learning in Graphical Models
by Michael Irwin Jordan (Editor)

List Price: ~~$60.00~~
Price: $60.00

This item ships for **FREE** with **Super**

Availability: Usually ships within 2 to 3 da

Used & new from $20.00

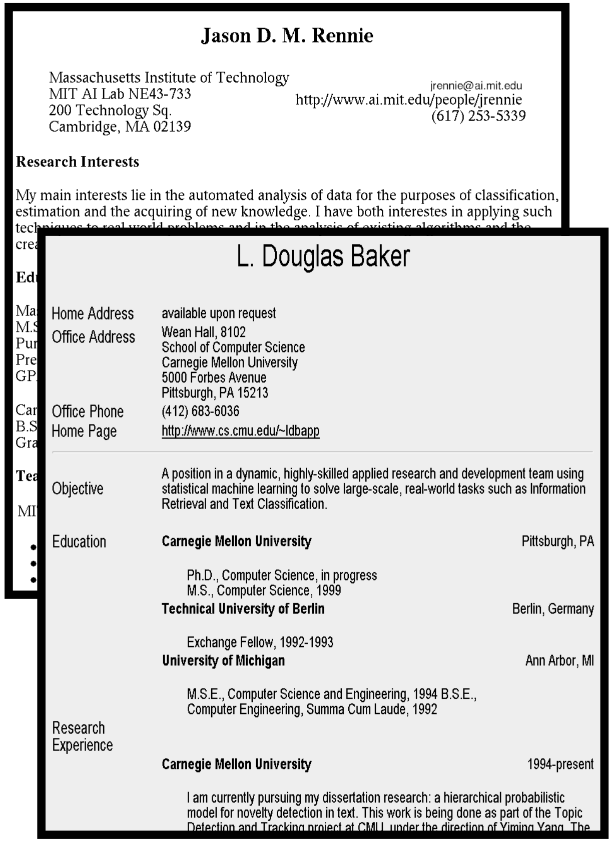
Edition: Paperback | All Editions

See more product details

Great Buy

Buy this book with *Probabilistic Reasoning in Intelligent Systems*

Buy Together Today: $128.95



Jason D. M. Rennie

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

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

Research Interests

My main interests lie in the automated analysis of data for the purposes of classification, estimation and the acquiring of new knowledge. I have both interestes in applying such

L. Douglas Baker

Home Address	available upon request
Office Address	Wean Hall, 8102 School of Computer Science Carnegie Mellon University 5000 Forbes Avenue Pittsburgh, PA 15213
Office Phone	(412) 683-6036
Home Page	http://www.cs.cmu.edu/~ldbapp

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

Education

Carnegie Mellon University Pittsburgh, PA

Ph.D., Computer Science, in progress
M.S., Computer Science, 1999

Technical University of Berlin Berlin, Germany

Exchange Fellow, 1992-1993

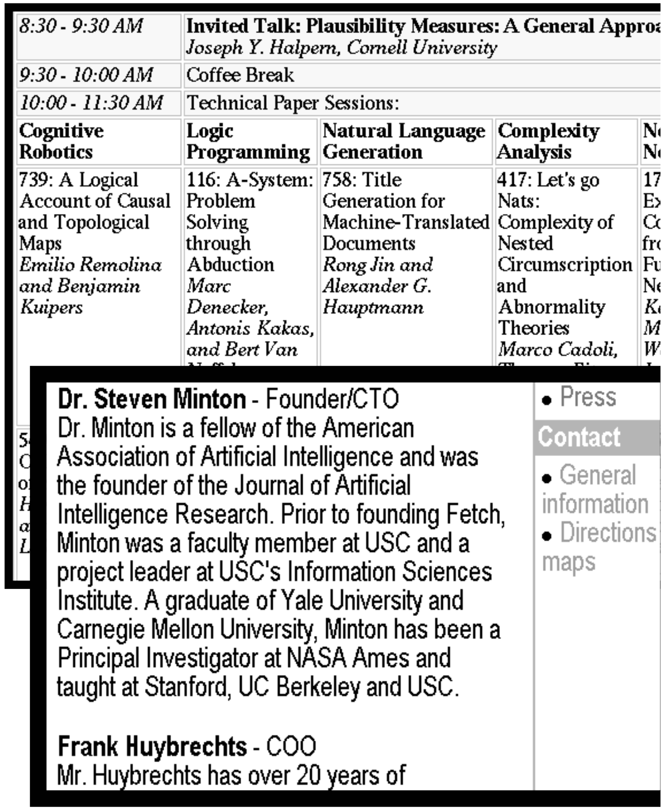
University of Michigan Ann Arbor, MI

M.S.E., Computer Science and Engineering, 1994 B.S.E., Computer Engineering, Summa Cum Laude, 1992

Research Experience

Carnegie Mellon University 1994-present

I am currently pursuing my dissertation research: a hierarchical probabilistic model for novelty detection in text. This work is being done as part of the Topic



8:30 - 9:30 AM	**Invited Talk: Plausibility Measures: A General Approa** *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**	
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*	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,*	

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

Landscape of IE Tasks (3/4): 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): Single Field vs. Record

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

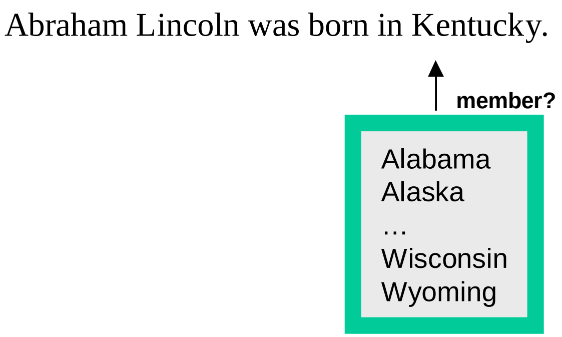
Relation:	**Succession**
Company:	**General Electric**
Title:	**CEO**
Out:	**Jack Welsh**
In:	**Jeffrey Immelt**

“Named entity” extraction

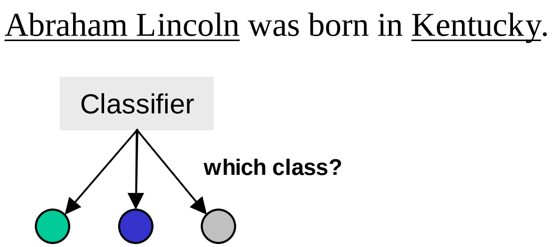
Relation extraction

Landscape of IE Techniques (1/1): Models

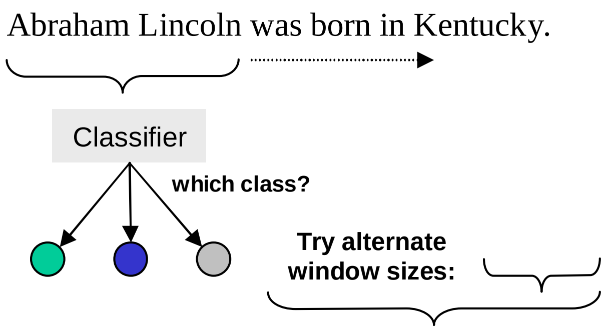
Lexicons


Abraham Lincoln was born in Kentucky.

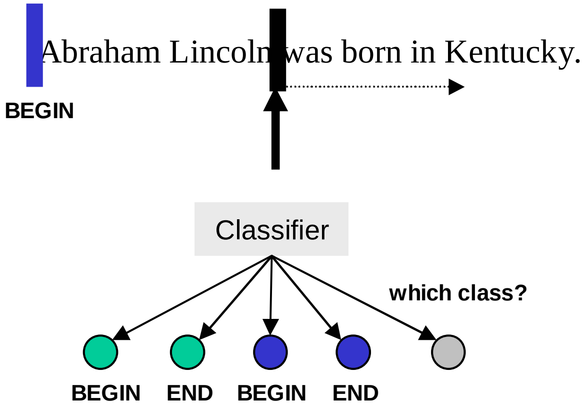
Classify Pre-segmented Candidates


Abraham Lincoln was born in Kentucky.

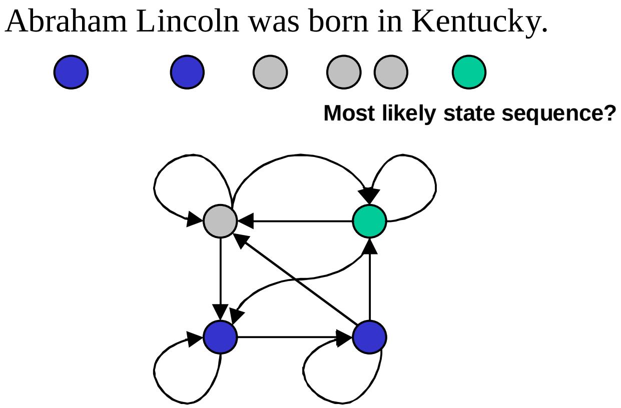
Sliding Window


Abraham Lincoln was born in Kentucky.

Boundary Models


Abraham Lincoln was born in Kentucky.

Finite State Machines


Abraham Lincoln was born in Kentucky.

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

8.2 Rule-based Information Extraction (Wrapper Induction)

Goal:

identify & extract unary, binary, and n-ary relations as facts embedded in regularly structured text, to generate entries in a schematized database

Approach:

rule-driven regular expression matching:

interpret docs from source (e.g. Web site to be wrapped) as regular language, and specify rules for matching specific types of facts

- Hand-annotate characteristic sample(s) for pattern
- Infer rules/patterns (e.g. using W4F (Sahuguet et al.) on IMDB):

```
movie = html
(.head.title.txt, match/(.*?) [(]/                    //title
 .head.title.txt, match/.*?[(]([0-9]+)[)]/            //year
 .body->td[i:0].a[*].txt                              //genre
where html.body->td[i].b[0].txt = „Genre“
and ...
```

LR Rules and Their Generalization

- Annotation of delimiters produces many small rules
- Generalize by combining rules (via inductive logic programming)
- Simplest rule type: **LR rule**

L token (left neighbor)	***fact token***	R token (right neighbor)
pre-filler pattern	*filler pattern*	*post-filler pattern*

Example:

<HTML> <TITLE> Some Country Codes </TITLE> <BODY>
<B> Congo </B> <I> 242 </I> <BR>
<B> Egypt </B> <I> 20 </I> <BR>
<B> France </B> <I> 30 </I> <BR>
</BODY> </HTML>

should produce binary relation with 3 tuples
{<Congo, 242>, <Egypt, 20>, <France, 30>}

Rules are:
L=<B>, R=</B> → Country
L=<I>, R=</I> → Code

Generalize rules by combinations (or even FOL formulas)

e.g. (L=<B> $\vee$ L=<td>) $\wedge$ IsNumeric(token) $\wedge$ … → Code

Implemented in RAPIER (Califf/Mooney) and other systems

Advanced Rules: HLRT, OCLR, NHLRT, etc.

Limit application of LR rules to proper contexts
(e.g. to skip over Web page header
<HTML> <TITLE> <B> List of Countries </B> </TITLE> <BODY> <B> Congo ...)

- **HLRT rules** (head left token right tail):
  apply LR rule only if inside H … T
- **OCLR rules** (open (left token right)* close):
  O and C identify tuple, LR repeated for invidual elements
- **NHLRT rules** (nested HLRT):
  apply rule at current nesting level,
  or open additional level, or return to higher level

Incorporate HTML-specific functions and predicates into rules:
inTitleTag(token), tableRowHeader(token), tableNextCol(token), etc.

Learning Regular Expressions

input: hand-tagged examples of a regular language
learn: (restricted) regular expression for the language
or a finite-state transducer that reads sentences of the language
and outputs the tokens of interest

Example:

This appartment has 3 bedrooms. <BR> The monthly rent is $ 995.
This appartment has 3 bedrooms. <BR> The monthly rent is $ 995.
The number of bedrooms is 2. <BR> The rent is $ 675 per month.

Learned pattern: * *Digit* „<BR>“ * „$“ *Number* *
Input sentence: There are 2 bedrooms. <BR> The price is $ 500 for one month.
Output tokens: Bedrooms: 1, Price: 500

but: grammar inference for full-fledged regular languages is hard
→ focus on restricted fragments of the class of regular languages

implemented in WHISK (Soderland 1999) and a few other systems

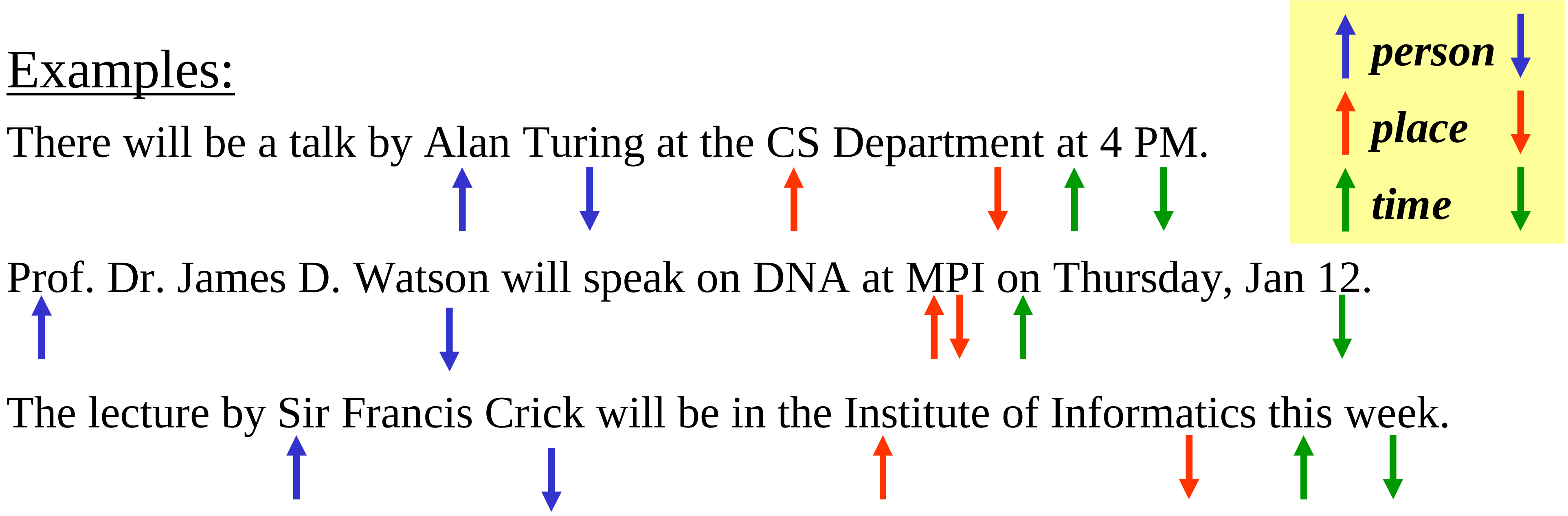
IE as Boundary Classification

Key idea:

Learn classifiers (e.g. SVMs) to recognize start token and end token for the facts under consideration

Combine multiple classifiers (ensemble learning) for robustness

Examples:



There will be a talk by Alan Turing at the CS Department at 4 PM.

Prof. Dr. James D. Watson will speak on DNA at MPI on Thursday, Jan 12.

The lecture by Sir Francis Crick will be in the Institute of Informatics this week.

Classifiers test each token (with PoS tag, LR neighbor tokens, etc. as features) for two classes: begin-fact, end-fact

Implemented in ELIE system (Finn/Kushmerick)

Properties and Limitations of Rule-based IE

- Powerful for wrapping regularly structured Web pages
  (typically from same Deep-Web site)
- Many complications on real-life HTML
  (e.g. misuse of HTML tables for layout)
  → use classifiers to distinguish good vs. bad HTML
- Flat view of input limits sample annotation
  → annotate tree patterns (and use tree automata for inferences)
  see e.g. Lixto (Gottlob et al.), Roadrunner (Crescenzi/Mecca)
- Regularities with exceptions difficult to capture
  → learn positive and negative cases (and use statistical models)

RAPIER in More Detail

slides on RAPIER are from:

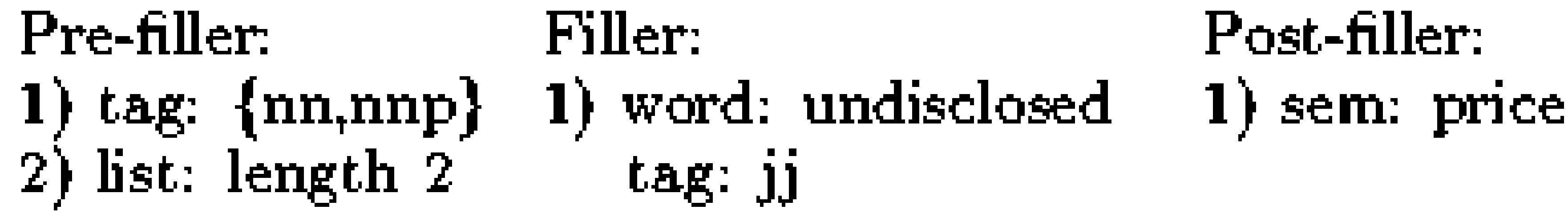
Christopher Manning, Prabhakar Raghavan, Hinrich Schütze,
Text Information Retrieval, Mining, and Exploitation
Course Material, Stanford University, Winter 2003
http://www.stanford.edu/class/cs276b/2003/syllabus.html

Rapier [Califf & Mooney, AAAI-99]

- Rapier learns three regex-style patterns for each slot:
  ▲Pre-filler pattern ▲ Filler pattern ▲ Post-filler pattern

- One of several recent trainable IE systems that incorporate linguistic constraints. (See also: **SIFT** [Miller *et al*, MUC-7]; **SRV** [Freitag, AAAI-98]; **Whisk** [Soderland, MLJ-99].)

"…paid $11M for the company…"
"…sold to the bank for an undisclosed amount…"
"…paid Honeywell an undisclosed price…"



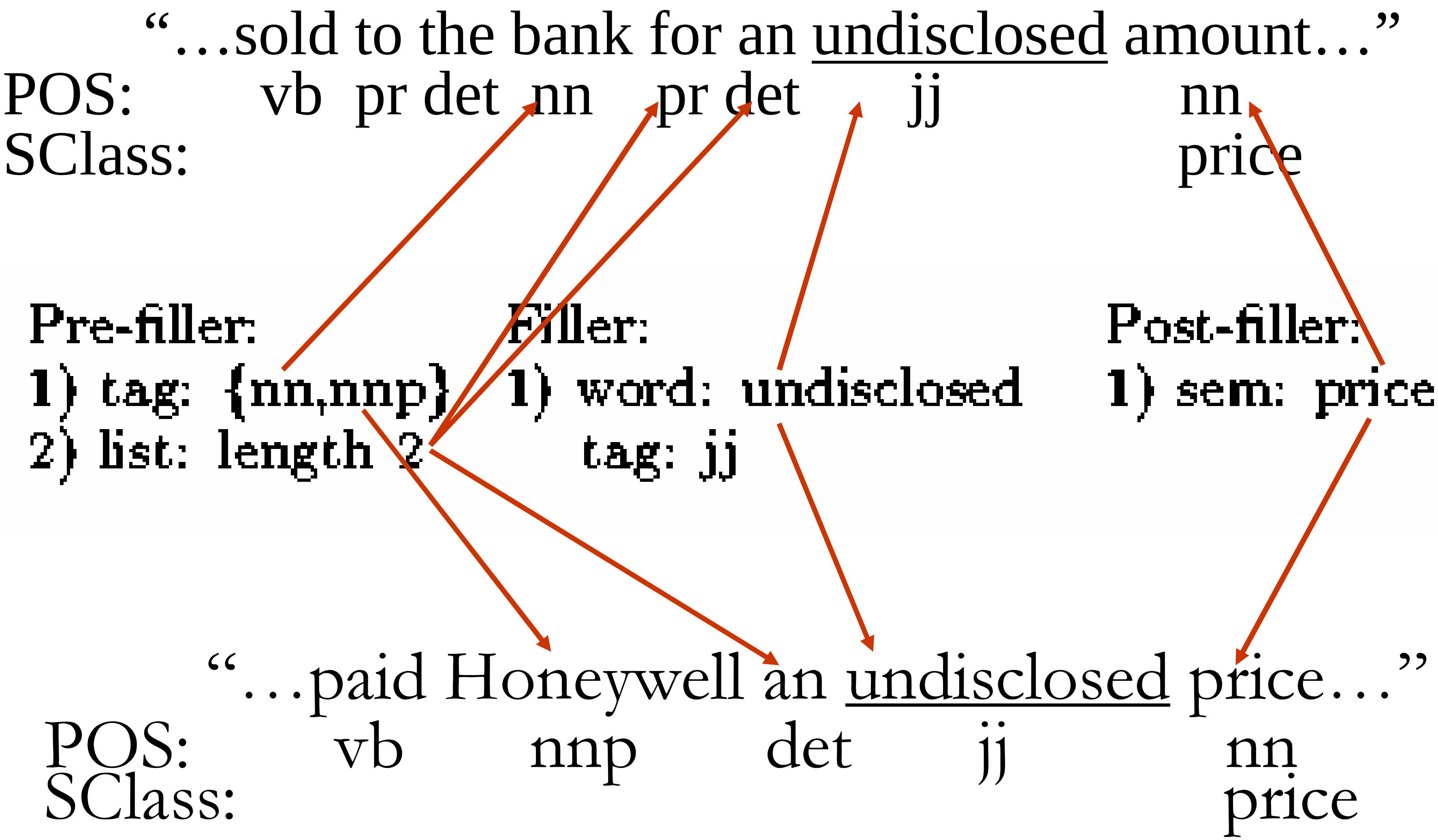
Pre-filler:	Filler:	Post-filler:
1) tag: {nn,nnp}	1) word: undisclosed	1) sem: price
2) list: length 2	tag: jj	

RAPIER rules for extracting "**transaction price**"

Part-of-speech tags & Semantic classes

- Part of speech: syntactic role of a specific word
  - noun (nn), proper noun (nnp), adjectve (jj), adverb (rb), determiner (dt), verb (vb), "." ("."), …
  - NLP: Well-known algorithms for automatically assigning POS tags to English, French, Japanese, … (>95% accuracy)
- Semantic Classes: Synonyms or other related words
  - "Price" class: price, cost, amount, …
  - "Month" class: January, February, March, …, December
  - "US State" class: Alaska, Alabama, …, Washington, Wyoming
  - WordNet: large on-line thesaurus containing (among other things) semantic classes

Rapier rule matching example



"…sold to the bank for an undisclosed amount…"

POS: vb pr det nn pr det jj nn

SClass: price

Pre-filler:	Filler:	Post-filler:
1) tag: {nn,nnp}	1) word: undisclosed	1) sem: price
2) list: length 2	tag: jj	

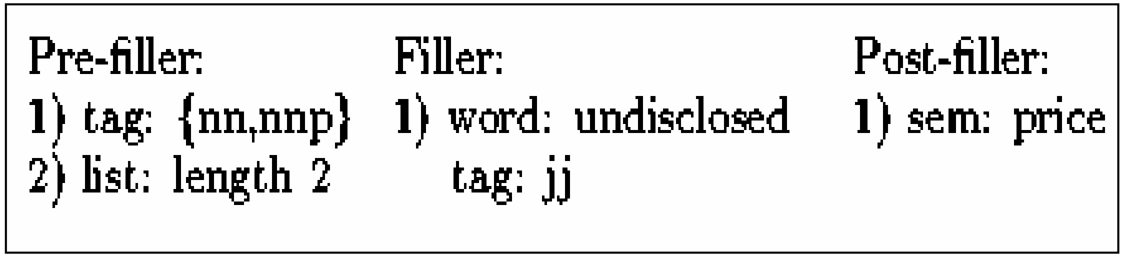
"…paid Honeywell an undisclosed price…"

POS: vb nnp det jj nn

SClass: price

Rapier Rules: Details

- **Rapier rule** :=
  - pre-filler **pattern**
  - filler **pattern**
  - post-filler **pattern**



Pre-filler:	Filler:	Post-filler:
1) tag: {nn,nnp}	1) word: undisclosed	1) sem: price
2) list: length 2	tag: jj	

- **pattern** := **subpattern +**
- **subpattern** := **constraint +**
- **constraint** :=
  - *Word* - exact word that must be present
  - *Tag* - matched word must have given POS tag
  - *Class* - semantic class of matched word
  - Can specify disjunction with “{…}”
  - *List length N* - between 0 and N words satisfying other constraints

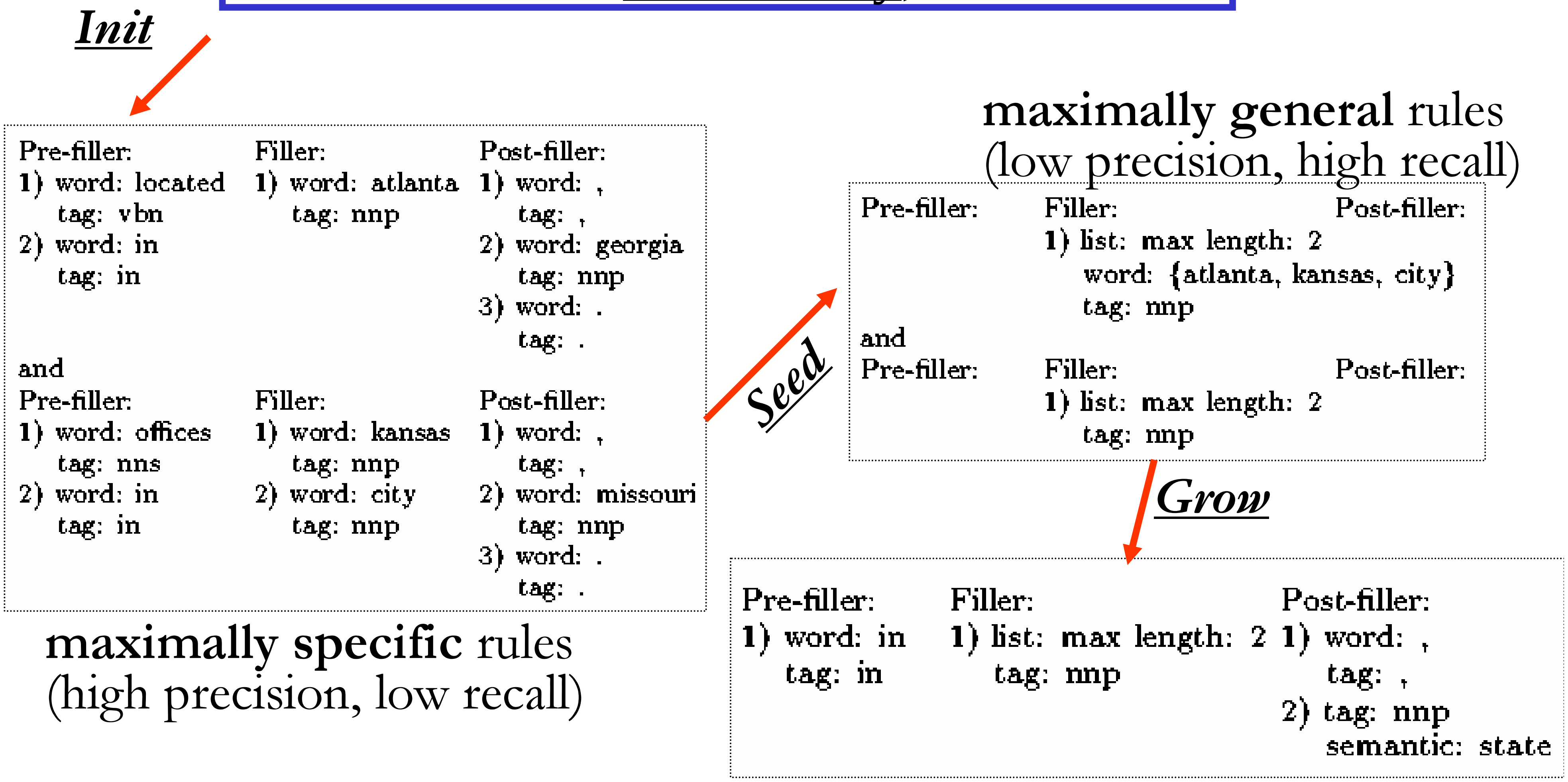
Rapier's Learning Algorithm

- **Input**: set of training examples (list of documents annotated with "extract this substring")
- **Output**: set of rules

- ***Init***: Rules = a rule that exactly matches each training example
- Repeat several times:
  - ***Seed***: Select M examples randomly and generate the K most-accurate maximally-general filler-only rules (prefiller = postfiller = "true").
  - ***Grow***:
    Repeat For N = 1, 2, 3, …
      Try to improve K best rules by adding N context words of prefiller or postfiller context
  - ***Keep***:
    Rules = Rules ∪ the best of the K rules – subsumed rules

Learning example (one iteration)

2 examples:
'… located in Atlanta, Georgia…"
'… offices in Kansas City, Missouri…'



Init

Pre-filler:	Filler:	Post-filler:
1) word: located	1) word: atlanta	1) word: ,
tag: vbn	tag: nnp	tag: ,
2) word: in		2) word: georgia
tag: in		tag: nnp
		3) word: .
		tag: .

and

Pre-filler:	Filler:	Post-filler:
1) word: offices	1) word: kansas	1) word: ,
tag: nns	tag: nnp	tag: ,
2) word: in	2) word: city	2) word: missouri
tag: in	tag: nnp	tag: nnp
		3) word: .
		tag: .

maximally specific rules
(high precision, low recall)

Seed

maximally general rules
(low precision, high recall)

Pre-filler:	Filler:	Post-filler:
	1) list: max length: 2	
	word: {atlanta, kansas, city}	
	tag: nnp	

and

Pre-filler:	Filler:	Post-filler:
	1) list: max length: 2	
	tag: nnp	

Grow

Pre-filler:	Filler:	Post-filler:
1) word: in	1) list: max length: 2	1) word: ,
tag: in	tag: nnp	tag: ,
		2) tag: nnp
		semantic: state

appropriately general rule (high precision, high recall)

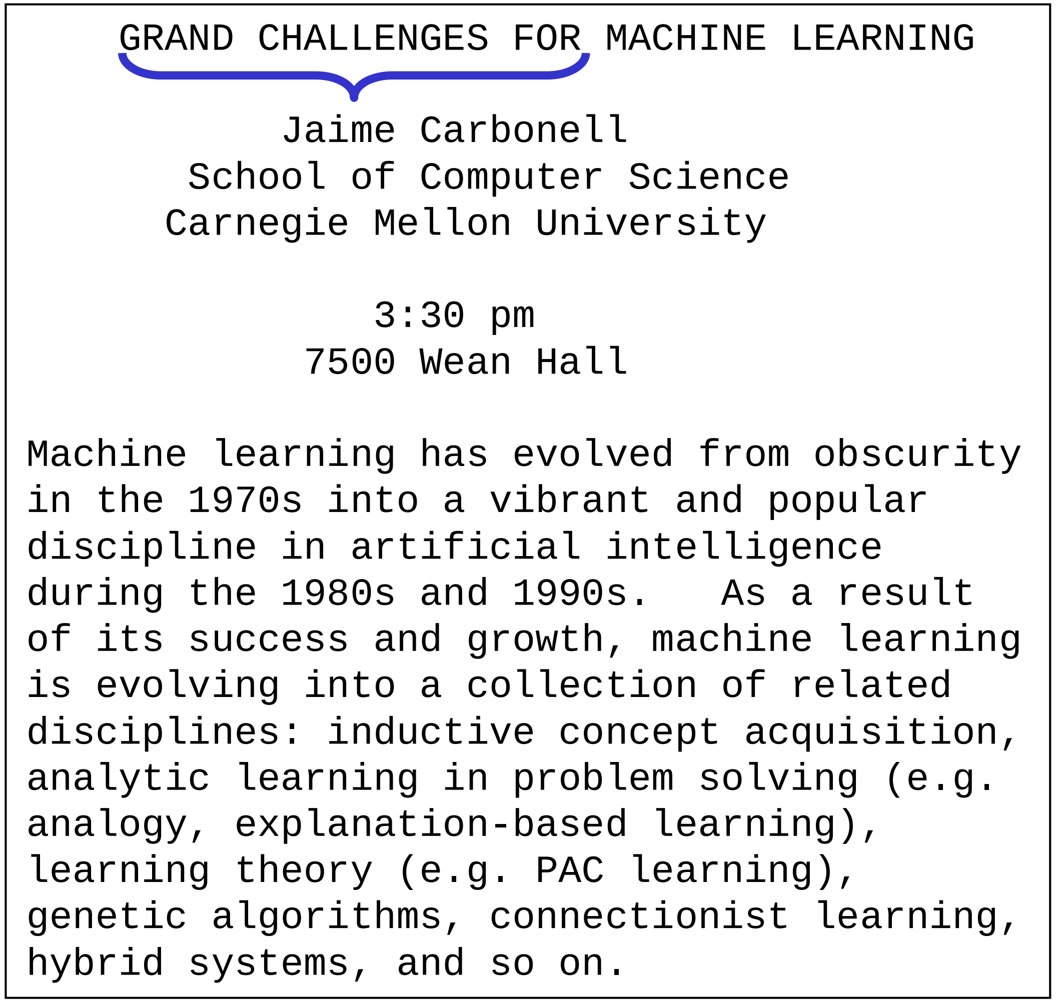
Sliding Windows

slides on Sliding-Windows IE are from:

William W. Cohen:
Information Extraction and Integration: an Overview,
Tutorial Slides,
http://www.cs.cmu.edu/~wcohen/ie-survey.ppt

Extraction by Sliding Window

E.g.
Looking for
seminar
location



GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell
School of Computer Science
Carnegie Mellon University

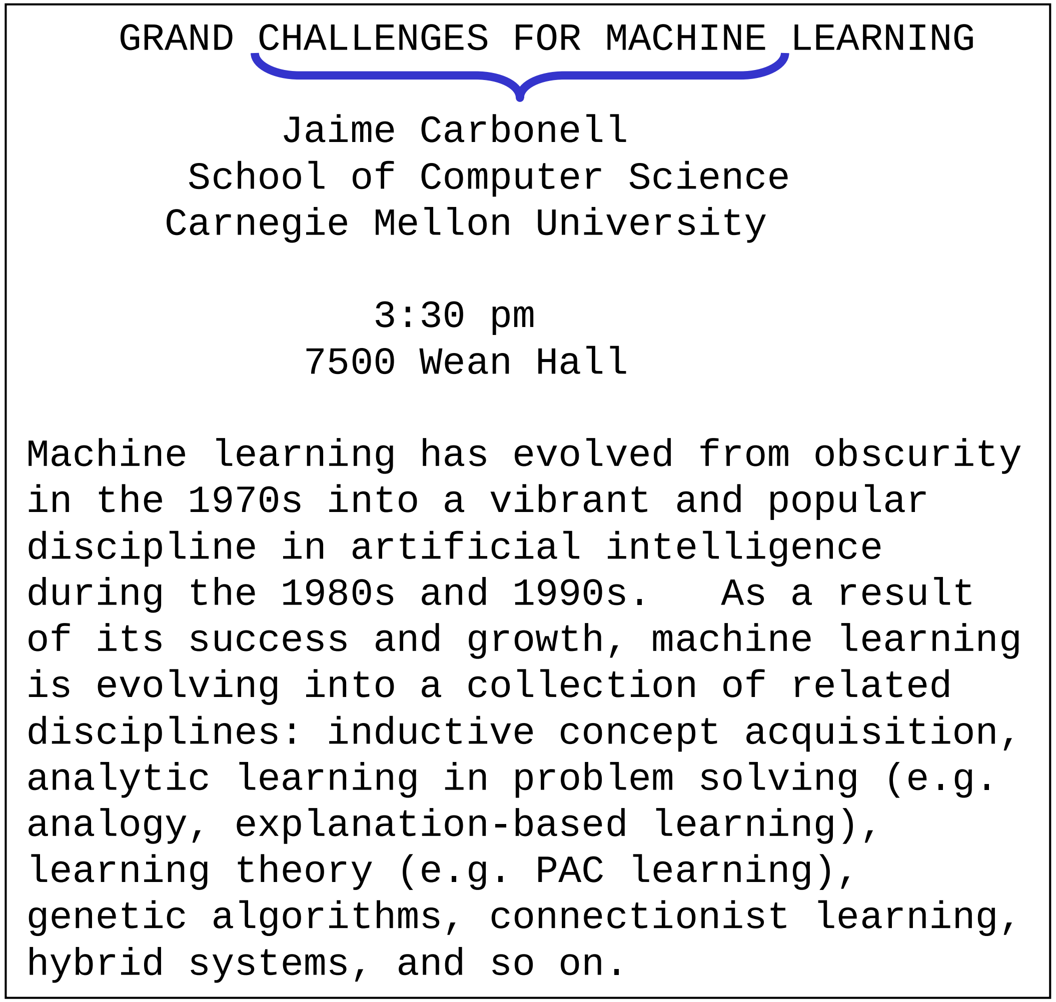
3:30 pm
7500 Wean Hall

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

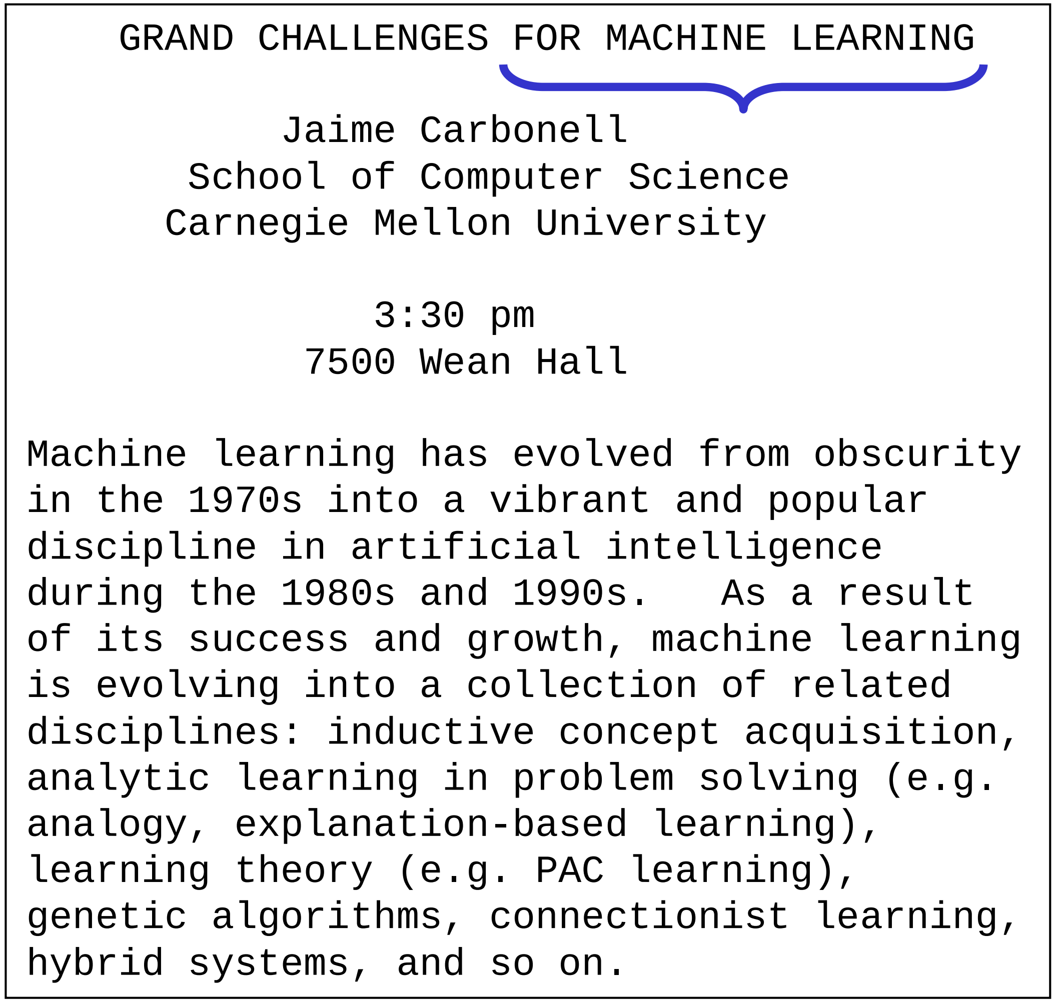


```
     GRAND CHALLENGES FOR MACHINE LEARNING

                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

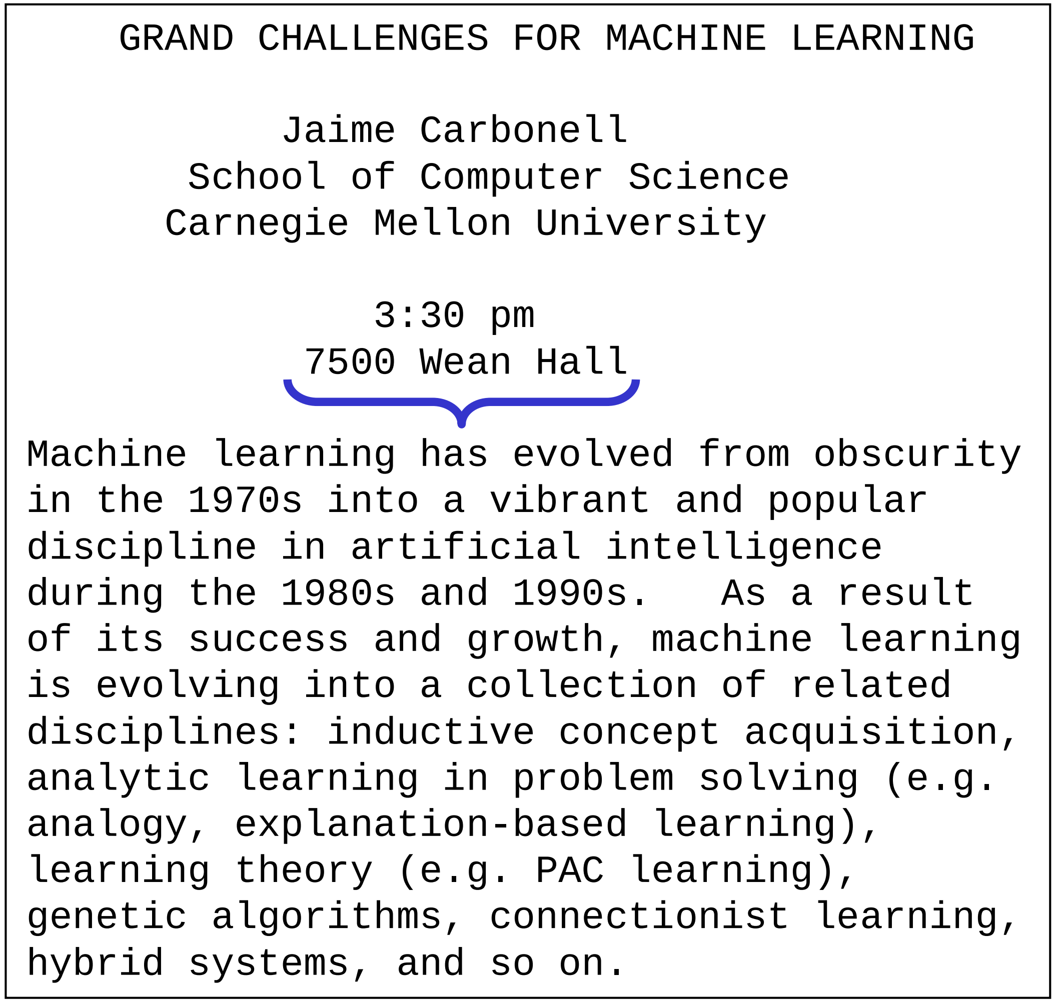
3:30 pm
7500 Wean Hall

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

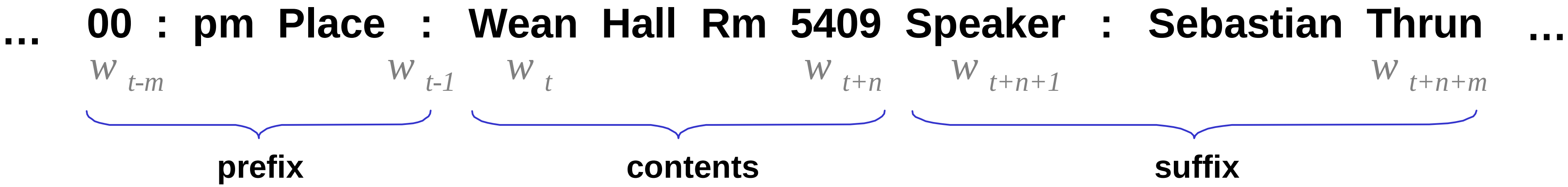
3:30 pm
7500 Wean Hall

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**

Estimate Pr(LOCATION|window) using Bayes rule

Try all “reasonable” windows (vary length, position)

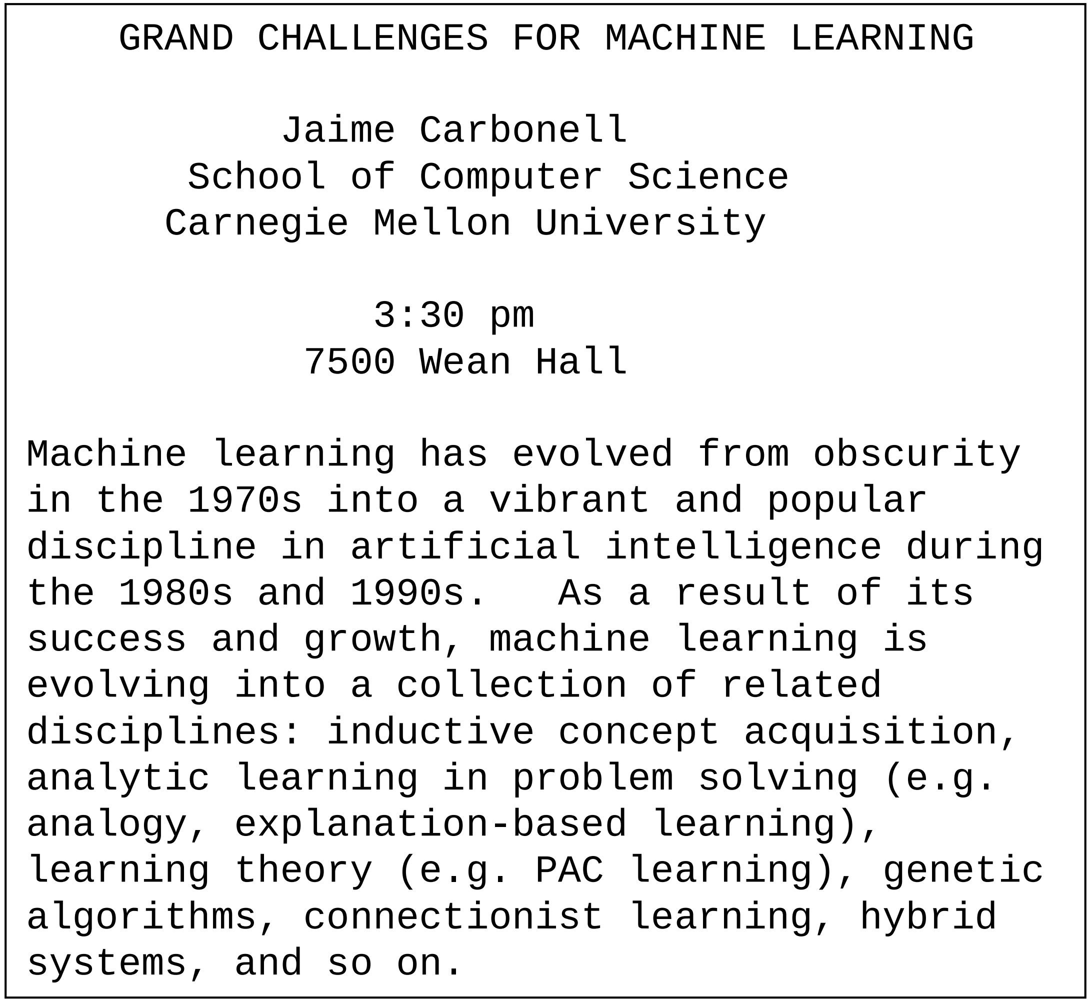
Assume independence for length, prefix words, suffix words, content words

Estimate from data quantities like: Pr(“Place” in prefix|LOCATION)

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

“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

[Freitag 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;
  - Use inductive logic programming techniques to express all these…

```
<title>Course Information for CS 213</title>
<h1>CS 213 C++ Programming</h1>
```

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, $\underline{f}$, $\underline{c}$) = {≀∇⊣⇕⇕𝒲⟩|𝒳: f(W)=c*
  - *some(+X, W, <$\underline{f_1,\ldots,f_k}$>, $\underline{g}$, $\underline{c}$)= ⌉§⟩∫⊔∫W:*

    $g(f_k(\ldots(f_1(W)\ldots))=c$
  - *tokenLength(+X, $\underline{\text{relop,}}$ $\underline{c}$):*
  - *position(+W,direction,$\underline{\text{relop}}$, $\underline{c}$):*
    - *e.g., tokenLength(X,>,4), position(W,fromEnd,<,2)*

courseNumber(X) ¬↖

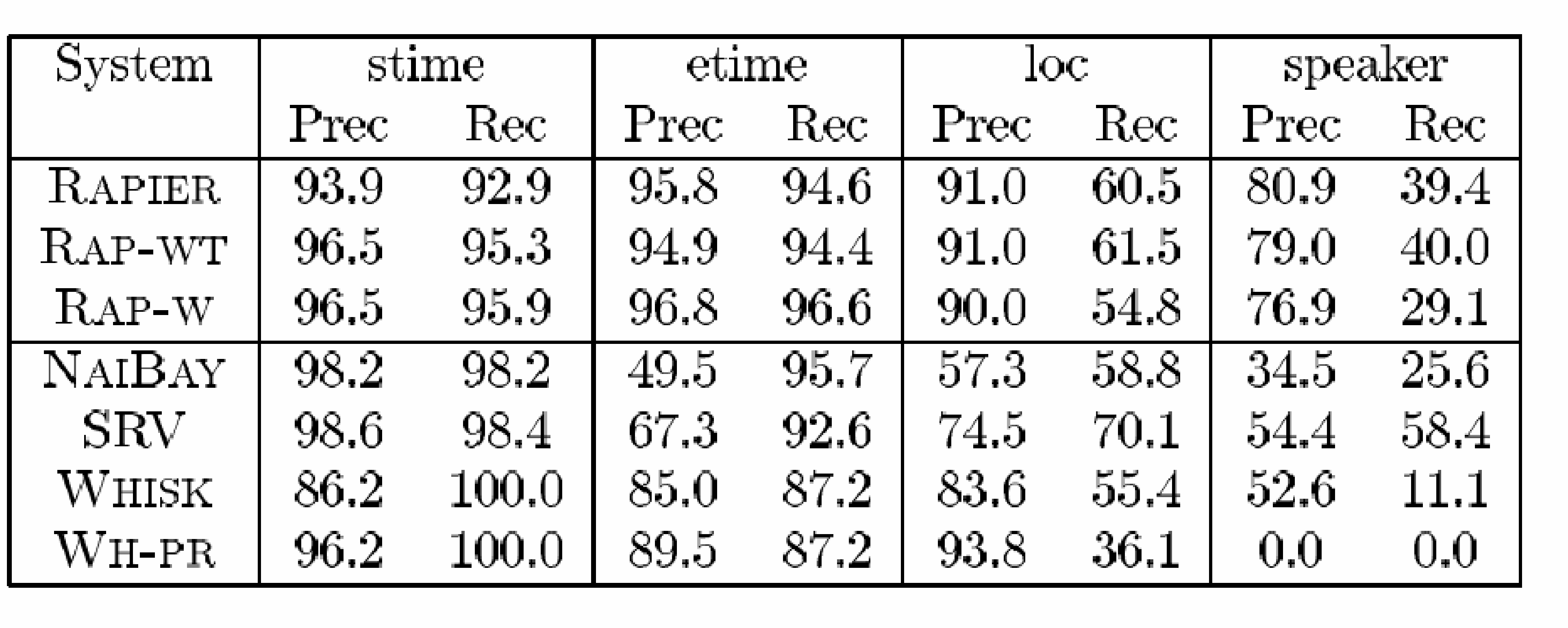
tokenLength(X,=,2),
every(X, inTitle, false),
some(X, A, <previousToken>, inTitle, true),
some(X, B, <>. tripleton, true)

Non-primitive conditions make greedy search easier

<title>Course Information for CS 213</title>

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

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

BWI: Learning to detect boundaries

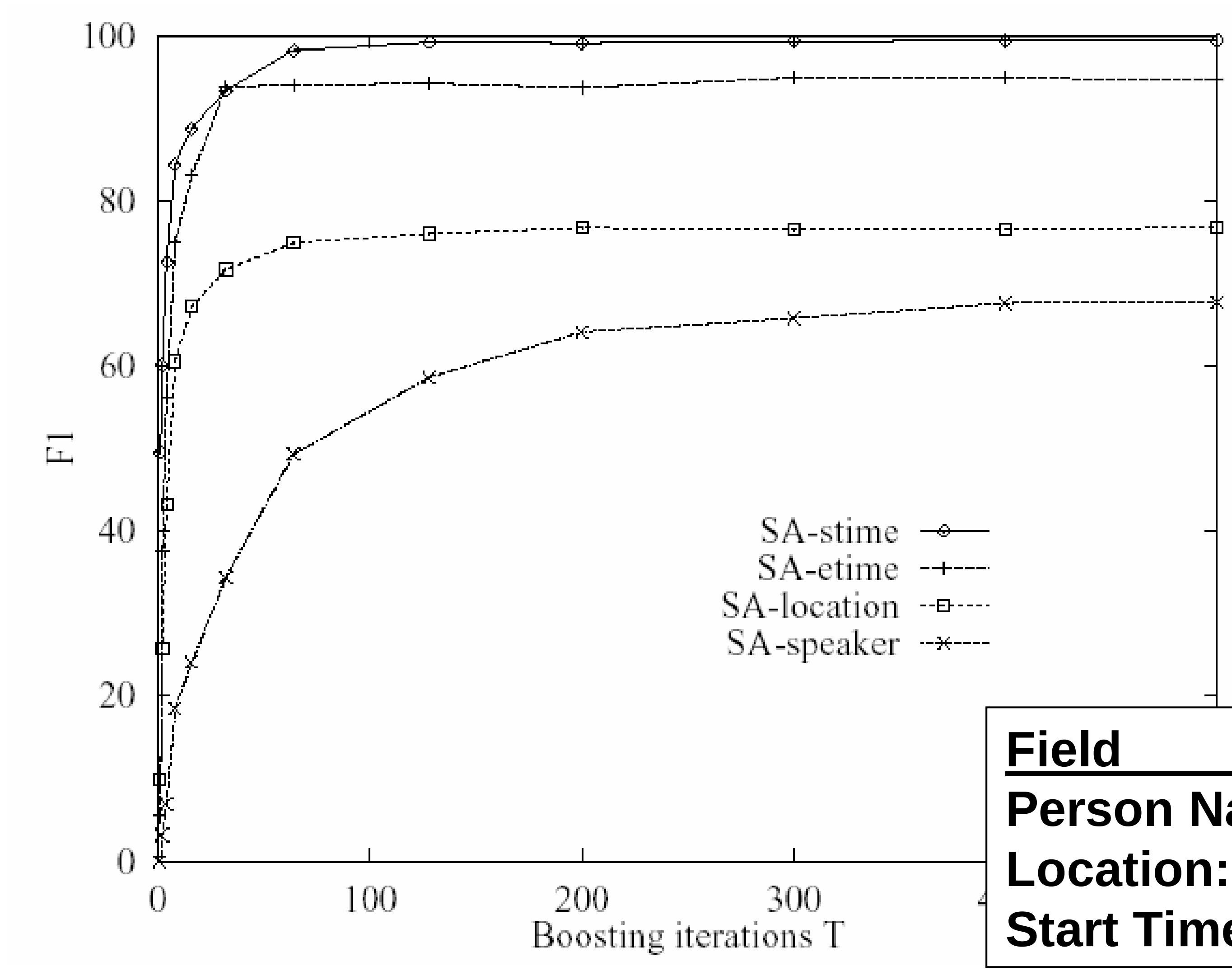
[Freitag & Kushmerick, AAAI 2000]

- 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



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.
  - Expensive for long entity names
  - 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.

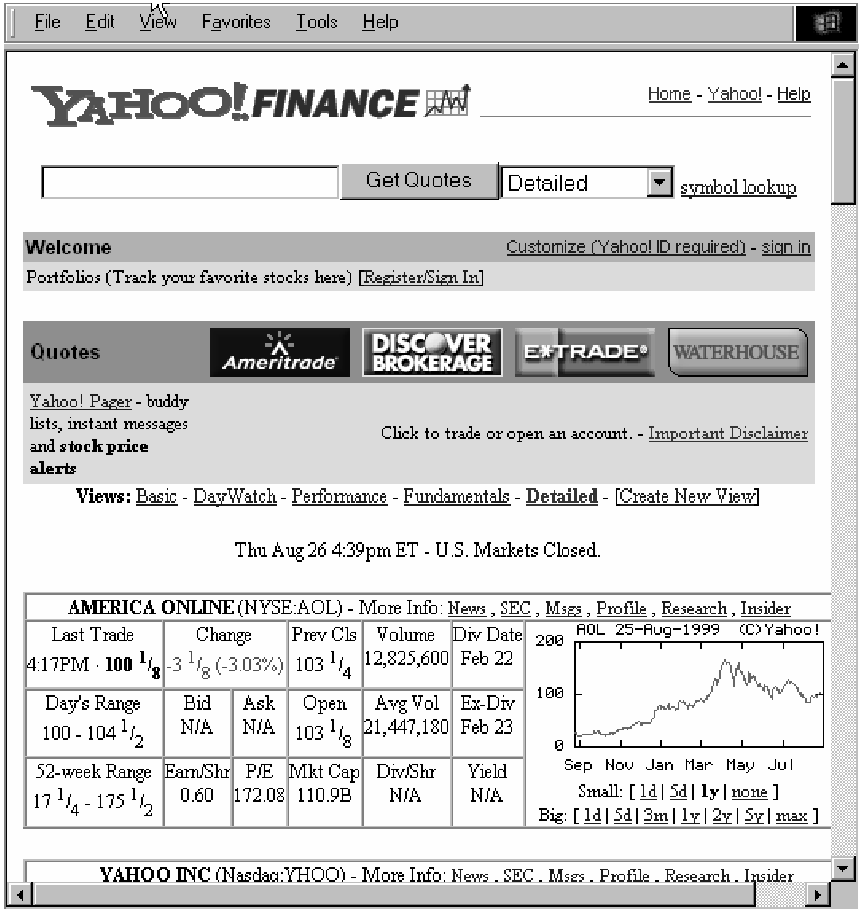
Tree-based Pattern Matcher: Example W4F (World Wide Web Wrapper Factory)

W4F (Sahuguet/Azavant 1999):
converts HTML to XML based on DOM Trees
based on hand-crafted rules
(+ GUI to simplify rule specification)

Following slides are from:

Arnaud Sahuguet, Fabien Azavant:
Looking at the Web through <XML> Glasses,
Talk at CoopIS 1999,
http://db.cis.upenn.edu/research/w4f.html

Put the glasses on



Home - Yahoo! - Help

symbol lookup

Welcome Customize (Yahoo! ID required) - sign in

Portfolios (Track your favorite stocks here) [Register/Sign In]

Quotes

Yahoo! Pager - buddy lists, instant messages and **stock price alerts**

Click to trade or open an account. - Important Disclaimer

Views: Basic - DayWatch - Performance - Fundamentals - **Detailed** - [Create New View]

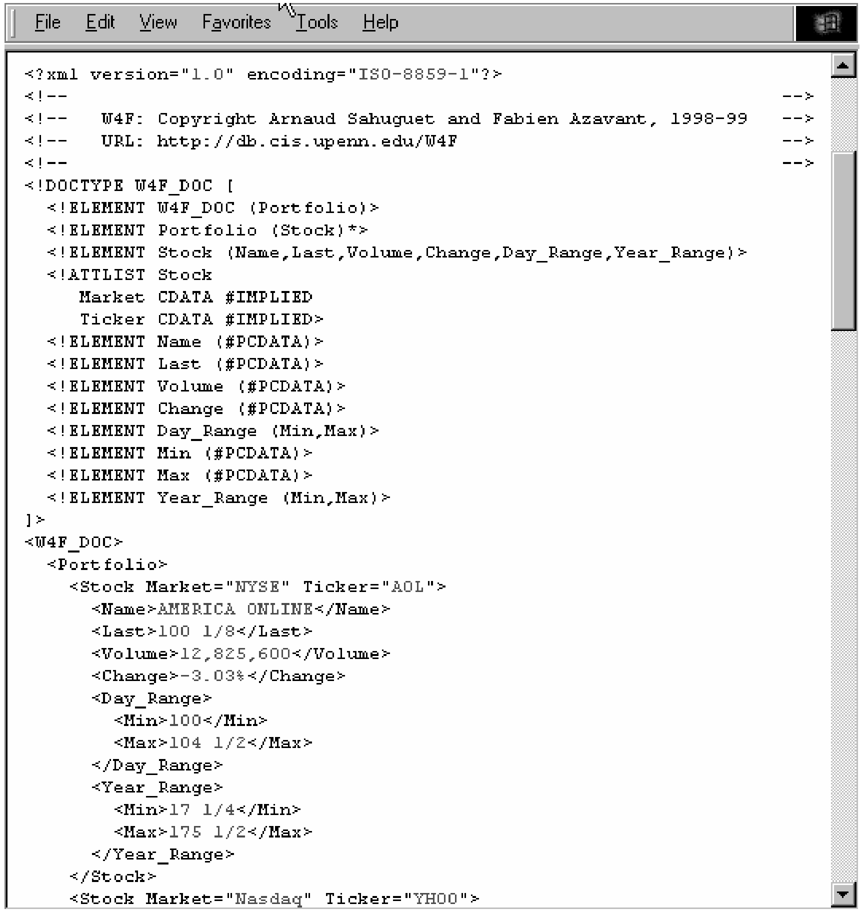
Thu Aug 26 4:39pm ET - U.S. Markets Closed.

AMERICA ONLINE (NYSE:AOL) - More Info: News , SEC , Msgs , Profile , Research , Insider

Last Trade 4:17PM · **100 $^{1}/_{8}$**	Change -3 $^{1}/_{8}$ (-3.03%)		Prev Cls 103 $^{1}/_{4}$	Volume 12,825,600	Div Date Feb 22
Day's Range 100 - 104 $^{1}/_{2}$	Bid N/A	Ask N/A	Open 103 $^{1}/_{8}$	Avg Vol 21,447,180	Ex-Div Feb 23
52-week Range 17 $^{1}/_{4}$ - 175 $^{1}/_{2}$	Earn/Shr 0.60	P/E 172.08	Mkt Cap 110.9B	Div/Shr N/A	Yield N/A

Small: [1d | 5d | **1y** | none]
Big: [1d | 5d | 3m | 1y | 2y | 5y | max]

YAHOO INC (Nasdaq:YHOO) - More Info: News , SEC , Msgs , Profile , Research , Insider

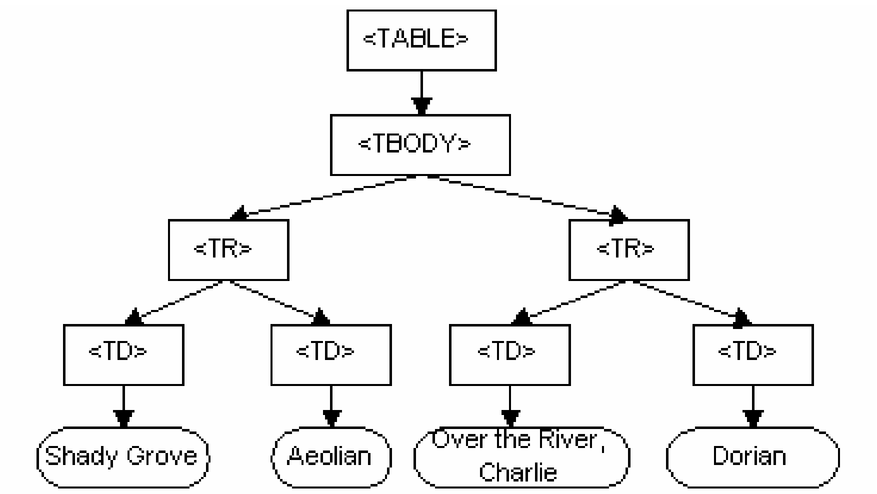


```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!--                                                               -->
<!--   W4F: Copyright Arnaud Sahuguet and Fabien Azavant, 1998-99  -->
<!--   URL: http://db.cis.upenn.edu/W4F                            -->
<!--                                                               -->
<!DOCTYPE W4F_DOC [
  <!ELEMENT W4F_DOC (Portfolio)>
  <!ELEMENT Portfolio (Stock)*>
  <!ELEMENT Stock (Name,Last,Volume,Change,Day_Range,Year_Range)>
  <!ATTLIST Stock
     Market CDATA #IMPLIED
     Ticker CDATA #IMPLIED>
  <!ELEMENT Name (#PCDATA)>
  <!ELEMENT Last (#PCDATA)>
  <!ELEMENT Volume (#PCDATA)>
  <!ELEMENT Change (#PCDATA)>
  <!ELEMENT Day_Range (Min,Max)>
  <!ELEMENT Min (#PCDATA)>
  <!ELEMENT Max (#PCDATA)>
  <!ELEMENT Year_Range (Min,Max)>
]>
<W4F_DOC>
  <Portfolio>
    <Stock Market="NYSE" Ticker="AOL">
      <Name>AMERICA ONLINE</Name>
      <Last>100 1/8</Last>
      <Volume>12,825,600</Volume>
      <Change>-3.03%</Change>
      <Day_Range>
        <Min>100</Min>
        <Max>104 1/2</Max>
      </Day_Range>
      <Year_Range>
        <Min>17 1/4</Min>
        <Max>175 1/2</Max>
      </Year_Range>
    </Stock>
    <Stock Market="Nasdaq" Ticker="YHOO">
```

HTML Extraction Language (HEL)

- Tree-based data-model
  - an HTML page is seen as a labeled tree (DOM^Document Object Model^)
- Tree navigation via path-expressions (with conditions)
  - extraction rules are described as paths along the tree
  - path expressions always return text values
- Regular expression
  - regular expressions (à la Perl) can be applied on text values to capture finer granularity

```
<TABLE> <TBODY>
<TR>
<TD>Shady Grove</TD>
<TD>Aeolian</TD>
</TR>
<TR>
<TD>Over the River, Charlie</TD>
<TD>Dorian</TD>
</TR>
</TBODY>
</TABLE>
```



HTML DOM Tree

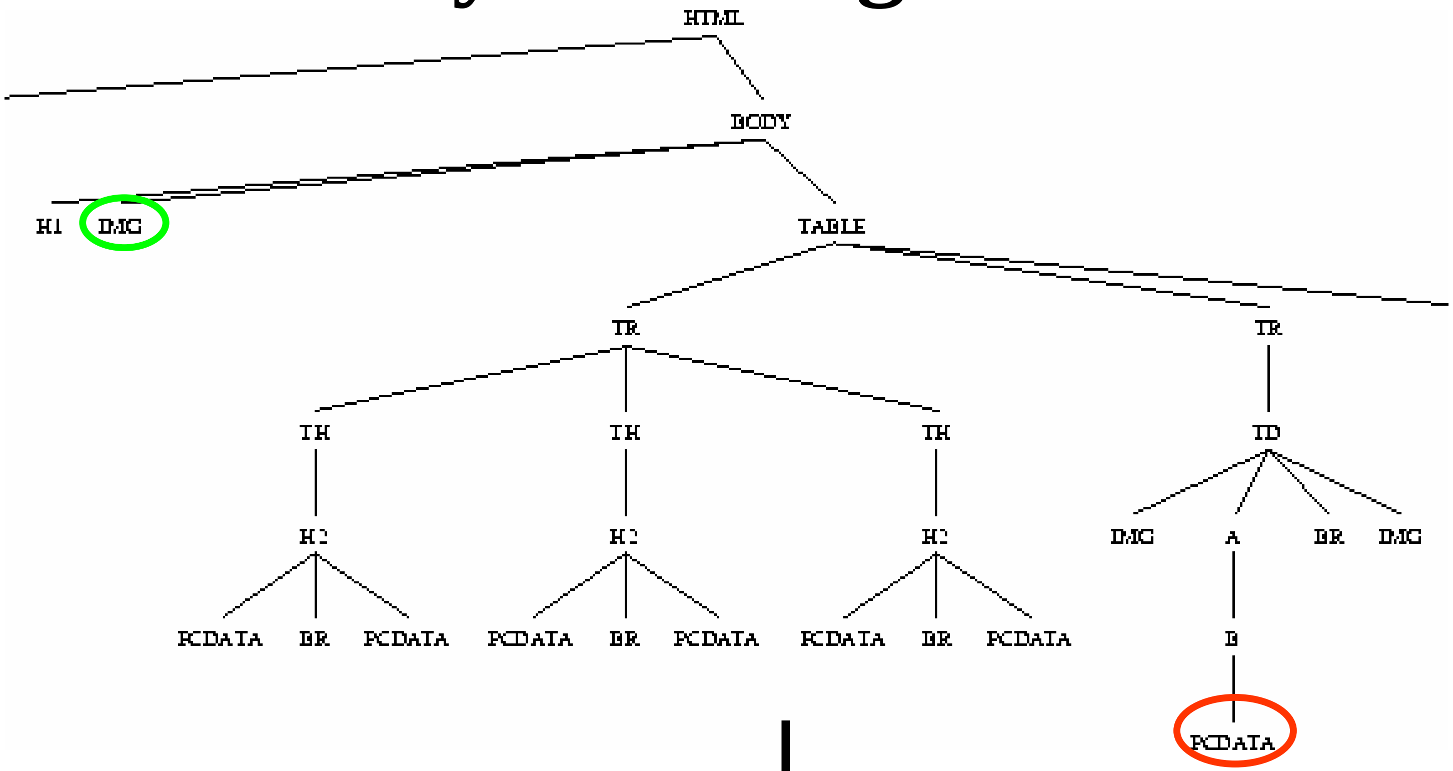
Tree navigation

- Following the document hierarchy: “.”
  - “.” explores the immediate children of a node
  - useful for limited nested structures

- Following the document flow: “->”
  - “->” explores the nodes found along a depth-first search
  - useful to create shortcuts
  - “->” only stops when it reaches the end

- When accessing nodes, index ranges can be used
  - e.g.. html.body->a[*].txt
  - e.g.. html.body.table[0].tr[1-].td[0].txt
  - returns a collection of nodes

2 ways to navigate the tree



HIERARCHICAL NAVIGATION

html.body.img[0].getAttr(src)

html.body.
table[0].tr[1].td[0].a[0].b[0].pcdata[0].txt

FLOW NAVIGATION

Using "->", there are more than 1 way to get to a node

html->img[0].getAttr(src)

html.h1[0]->img[0].getAttr(src)

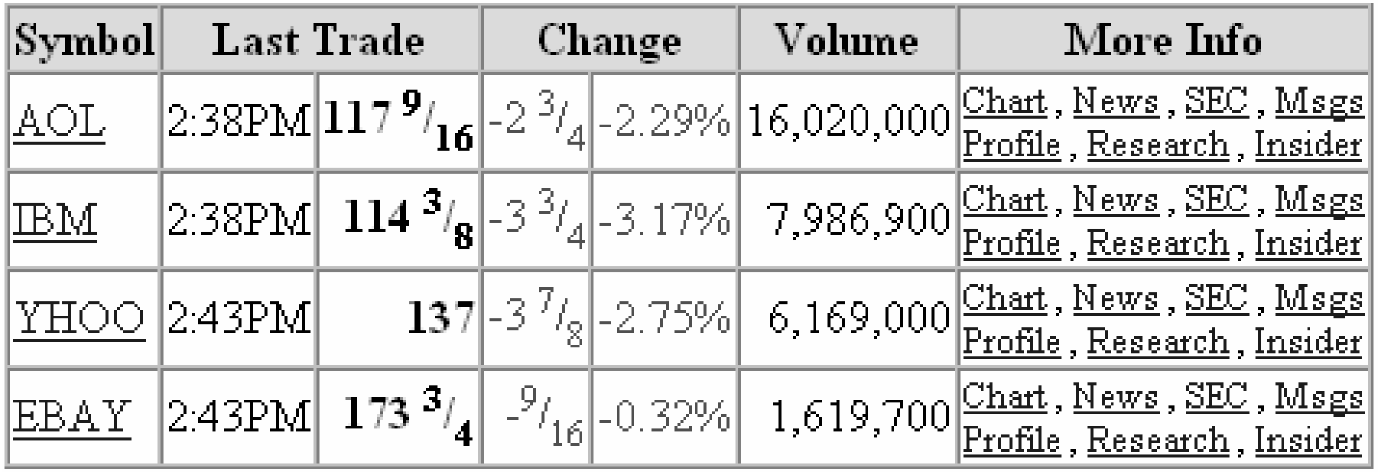
html->tr[1]->pcdata[0].txt

html->pcdata[7].txt

Using conditions

- Sometimes, we do not know ahead of time where exactly the information is located. Take the example of the IBM stock.

Let us assume that this table corresponds to table[5] inside the HTML page.



Symbol	Last Trade		Change		Volume	More Info
AOL	2:38PM	**117 9/16**	-2 3/4	-2.29%	16,020,000	Chart , News , SEC , Msgs Profile , Research , Insider
IBM	2:38PM	**114 3/8**	-3 3/4	-3.17%	7,986,900	Chart , News , SEC , Msgs Profile , Research , Insider
YHOO	2:43PM	**137**	-3 7/8	-2.75%	6,169,000	Chart , News , SEC , Msgs Profile , Research , Insider
EBAY	2:43PM	**173 3/4**	-9/16	-0.32%	1,619,700	Chart , News , SEC , Msgs Profile , Research , Insider

- You can write the following extraction rule:

```
            html->table[5].tr[i].td[2].txt
      where html->table[5].tr[i].td[0].txt = “IBM”
```

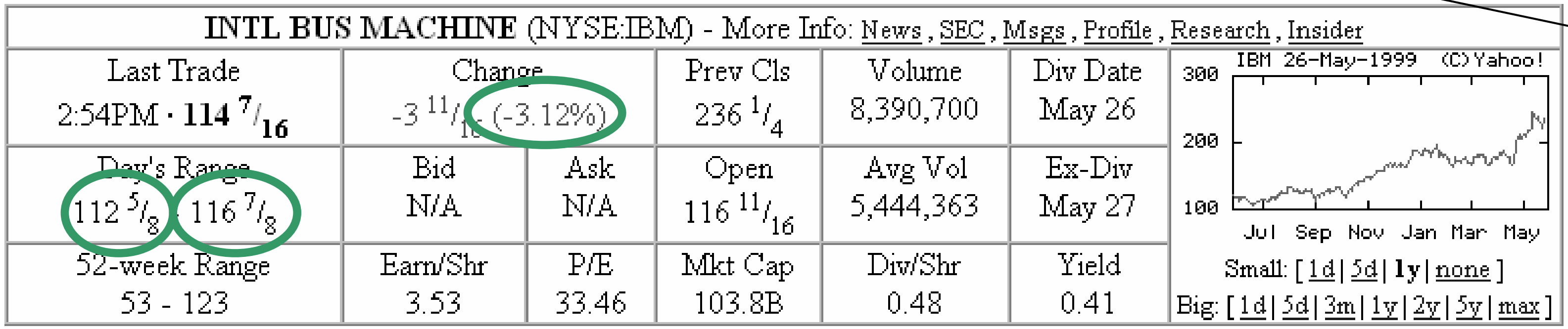
- Conditions involve index ranges only.
- Conditions are resolved against node properties, not nodes themselves.

Using regular expressions (à la Perl)

- In some cases, we want to go deeper than the tag structure.

- We want to extract the % change
  - table.tr[1].td[1].txt, **match /[(](.*?)[)]/**
- We want to extract the day’s range for the stock:
  - table.tr[2].td[0].txt, **match/Day’s Range (.*)/, split /-/**

regular expression operators can be used in cascade



INTL BUS MACHINE (NYSE:IBM) - More Info: News , SEC , Msgs , Profile , Research , Insider

Last Trade 2:54PM · **$114\ {}^{7}/_{16}$**	Change $-3\ {}^{11}/_{16}$ (-3.12%)		Prev Cls $236\ {}^{1}/_{4}$	Volume 8,390,700	Div Date May 26	
Day's Range $112\ {}^{5}/_{8}$ - $116\ {}^{7}/_{8}$	Bid N/A	Ask N/A	Open $116\ {}^{11}/_{16}$	Avg Vol 5,444,363	Ex-Div May 27	
52-week Range 53 - 123	Earn/Shr 3.53	P/E 33.46	Mkt Cap 103.8B	Div/Shr 0.48	Yield 0.41	Small: [1d | 5d | **1y** | none] Big: [1d | 5d | 3m | 1y | 2y | 5y | max]

- Semantics
  - match /(.....)/ returns a string
  - match /(...) (...)/ returns a list of strings
  - split /...../ returns a list of strings

Building Complex Structures

- Atomic values are not enough.
- The fork operator “#” permits to follow a path along various subpaths. Results are put together into a list.
- Following the previous example, we can extract the entire stock inform and put it in one structure.

```
html.body.center.table[i:*]
    ( .tr[0].td[0].b[0].txt                                // name
    # .tr[0].td[0].b[0]->pcdata[1].txt, match /[(](.*?):/  // trading plac
    # .tr[0].td[0].b[0]->pcdata[1].txt, match /:(.*?)[)]/  // ticker
    # .tr[1].td[0].b[0].txt                                // last trade
    # .tr[1].td[3].pcdata[1].txt                           // volume
    # .tr[1].td[1].txt, match /[(](.*?)[)]/                // change %
    # .tr[2].td[0].txt, match /Range(.*)/, split /-/       // Day range
    # .tr[3].td[0].txt, match /Range(.*)/, split /-/       // Year range
    )
where html.body.center.table[i].tr[0].td[0].getAttr(colspan) = "7";
```