

# Information Extraction

CIS, LMU München  
Winter Semester 2015-2016

Dr. Alexander Fraser, CIS

# Information Extraction – Administravia - I

- Vorlesung
  - Learn the basics of Information Extraction (IE)
- Seminar
  - Each student will present a Referat on IE (Powerpoint, LaTeX, Mac)
    - The group will discuss it
  - Also: three or so practical sessions (hopefully we have time)
  - There are two seminars! **You come to just one of the two sessions**, either Thursdays (starting tomorrow), or Wednesdays (starting next week)

# Information Extraction – Administravia - II

- Registration:
  - If you are a CIS Student: check whether you are registered for *\*both\** the Vorlesung and the Seminar (these are **two things** in LSF!)
  - There are a good number of people only in the Vorlesung
  - There are a few people only in the Seminar
- A word about space:
  - The seminars are very full in LSF
  - This may be because people registered who will not actually do a Referat – if this applies to you, please let me know (for the sake of your colleagues!)

# Information Extraction – Administravia - III

- Vorlesung and Seminar are two separate courses (in same module for CIS people)
  - However, there may be some shifting around of slots depending on time constraints
- Vorlesung (Grade):
  - Klausur (probably 03.02, no discussion of this today please)
- Seminar (Grade):
  - Referat
  - Hausarbeit (write-up of the Referat) (6 pages, due **3 weeks** after you hold your Referat)
  - The Hausarbeit can also include practical exercises (optional, extra points)
- CIS-ler: No Notenverbesserung (everyone else: ask in your Fachschaft!)

# Information Extraction - Administravia - IV

- NEXT SEMINAR - COME TOMORROW \*OR\* ON COMING WEDNESDAY!
  - Ungraded quiz (so that I can see what you already know)
    - Optionally anonymous (you either put your name, or you don't)
  - I will also collect information on who you are and your interests – PUT YOUR NAME ON THIS PAGE! (this page will be collected separately!)
  - **And I want to know what you want to learn in this class!**

# Information Extraction – Administravia - V

- Syllabus: updated dynamically on my web page (see also WS last year, but there will be some differences)
  - Brief idea at end of this slide deck (if we finish, then today)
- List of Referatsthemen
  - This will be presented soon in the Seminar, probably in two weeks
- Literature:
  - Required: **Sunita Sarawagi. Information Extraction.** Foundations and Trends in Databases, 1(3):261–377, 2008. (good survey paper, somewhat brief)
    - **Please read the introduction for next week (it is available on the web page!)**
  - Optional: Christopher D. Manning, Prabhakar Raghavan and Hinrich Schuetze, *Introduction to Information Retrieval*, Cambridge University Press. 2008. (good information retrieval textbook, preview copies available from the book website: <http://nlp.stanford.edu/IR-book/>)

- Questions?

# Information Extraction

- An introduction to the course
  - The topic "Information Extraction" means different things to different people
  - In this course we will look at several different perspectives
  - There is unfortunately no comprehensive textbook that includes all of these perspectives



# My Biases

- As you may have noticed by now: I am from the US (PhD in Computer Science from USC/ISI AI division)
- I am on permanent staff here at CIS
- I do research in the broad area of **statistical NLP**
  - I mostly work on **statistical machine translation**, and related structured prediction problems (e.g., treebank-based syntactic parsing, generation using sequence (tagging) models)
  - I also work on other multilingual problems such as cross-language information retrieval
- With respect to **rule-based NLP** (with manually written rules), I'll try to be as fair as (humanly) possible, I do use these techniques sometimes too

# Outline for today

- Motivation
  - Problems requiring information extraction
  - Basic idea of the output
- Abstract idea of the core of an information extraction pipeline
- Course topics

# A problem




### [Bakery Jobs on CareerBuilder.com](#)

[www.careerbuilder.com/jobs/keyword/bakery](http://www.careerbuilder.com/jobs/keyword/bakery) 


Jobs 1 - 25 of 579 – Looking for **Bakery Jobs**? See currently available job **openings** on CareerBuilder.com. Browse the current listings and fill out job applications.

### [Baker Jobs, Employment | Indeed.com](#)

[www.indeed.com/q-Baker-jobs.html](http://www.indeed.com/q-Baker-jobs.html) 

Jobs 1 - 10 of 16047 – 16047 **Baker Jobs** available on Indeed.com. one search. all jobs.

### [Job Openings - Baker University](#)

[www.bakeru.edu](http://www.bakeru.edu) > Jobs 

If you are seeking employment in any of these areas, contact **Baker** University.

### [Baker, LA Jobs on CareerBuilder.com](#)

[www.careerbuilder.com/Jobs/Baker/](http://www.careerbuilder.com/Jobs/Baker/) 

Jobs 1 - 25 of 948 – Looking for **Baker, LA Jobs**? See currently available **job openings** on CareerBuilder.com. Browse the current listings and fill out **job ...**

### [Down Under Bakery Pies: Job Openings at DUB Pies](#)

[www.dubpies.com/jobs.php](http://www.dubpies.com/jobs.php) 

Listing of **job openings** at DUB Pies. Down Under **Bakery** (DUB) Pies is looking for more staff - check out our list of vacancies.

### [Field Engineers | Geoscience | Jobs and Careers at Baker Hughes](#)

[jobs.bakerhughes.com/](http://jobs.bakerhughes.com/) 


... Oil and Natural Gas? **Baker Hughes** has career information for you on these, more. ... Search **Jobs**. **Baker Hughes Jobs** ... Recent **Job Openings**. Completion ...

### [Corner Bakery Job Openings | Glassdoor](#)

[www.glassdoor.com/Job/Corner-Bakery-Job-Openings-E297310\\_P2...](http://www.glassdoor.com/Job/Corner-Bakery-Job-Openings-E297310_P2...) 

45 Corner **Bakery job openings**. Search job openings, see if they fit - company salaries, reviews, and more posted by Corner Bakery employees.

### [Jobs - Baker University](#)

[www.bakeru.edu/jobs](http://www.bakeru.edu/jobs) 

See links at left for a complete list of **Baker** University **job openings**. It is the policy of **Baker** University to afford equal opportunity for all persons without distinction ...

# A solution




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

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

**FlipDog.com**

Home Find Jobs Your Account Resource Center Support Employers

Job Search at FlipDog.com: Employment & Career Management



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

[Find a Job!](#)

[Post Your Resume](#)

**Employers**  
click here for  
Products & Services

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

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

**Pigskin Places**

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

**Jobs for Sports Fans**

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

**Job Seeker Newsletter**

Enter your e-mail address:

[Sign Me Up!](#)

**Showcase Jobs**

**MRI**  
Management Recruiters  
of Charlotte North

We provide total staffing solutions in the areas of Human Resources, Compensation, Web-based HR, self-service, and Customer Management Systems. [Learn More](#)



Looking for a Vice President of Academic Affairs to oversee planning, operation and evaluation of the college's academic programs. [Learn More](#)

powered by  
**WhizBang!**

Top 100 Web Sites\*  
PC Magazine, Nov. 2000

Top 10 Career Web Site\*  
Media Matrix, Sept. 2000

Top 10 Job Site\*

Internet



FlipDog.com

Fetch Your Next Job Here™

Employers Support

Home

Find Jobs

Your Account

Resource Center

Return to Results | Modify Search | New Search



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

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



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

1 - 25 of 47 jobs shown below

1 2 Next >

Search these results for:



Search tips

Show Jobs Posted:

For all time periods

View: Brief | Detailed

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

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

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

# Extracting Job Openings from the Web

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

File Edit View Favorites

Back Forward Stop

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

Links AMEX Rewards Time DogHouse My Yahoo!

Welcome

About OPUS

Executive Staff

**Job Listings**

Résumé Form

Job Hunt Hints

Academic Links

Science Fair Help

Industry Assocs.

FAQs

Contact Us

Site Map

e-mail

OPUS INTERNATIONAL INC.

About | Staff | Job

OPUS: Job Listings - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites

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

Links AMEX Rewards Time DogHouse My Yahoo!

Welcome

About OPUS

Executive Staff

**Job Listings**

Résumé Form

Job Hunt Hints

Academic Links

Science Fair Help

Industry Assocs.

FAQs

Contact Us

Site Map

e-mail

Test Kitchen-Consumer Food Relations

Major food manufacturer in Chicago area seeks a consumer food professional to write and test recipes. Will make presentations; will be a key player in a cross-functional team. Requires a BS in human ecology, nutrition, Food Science, or related field, plus a minimum three years' applicable experience.

Contact: Moira: e-mail: 1-800-488-2611

**Ice Cream Guru**

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

Contact: Susan: e-mail: 1-800-488-2611

**Title:** Ice Cream Guru

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

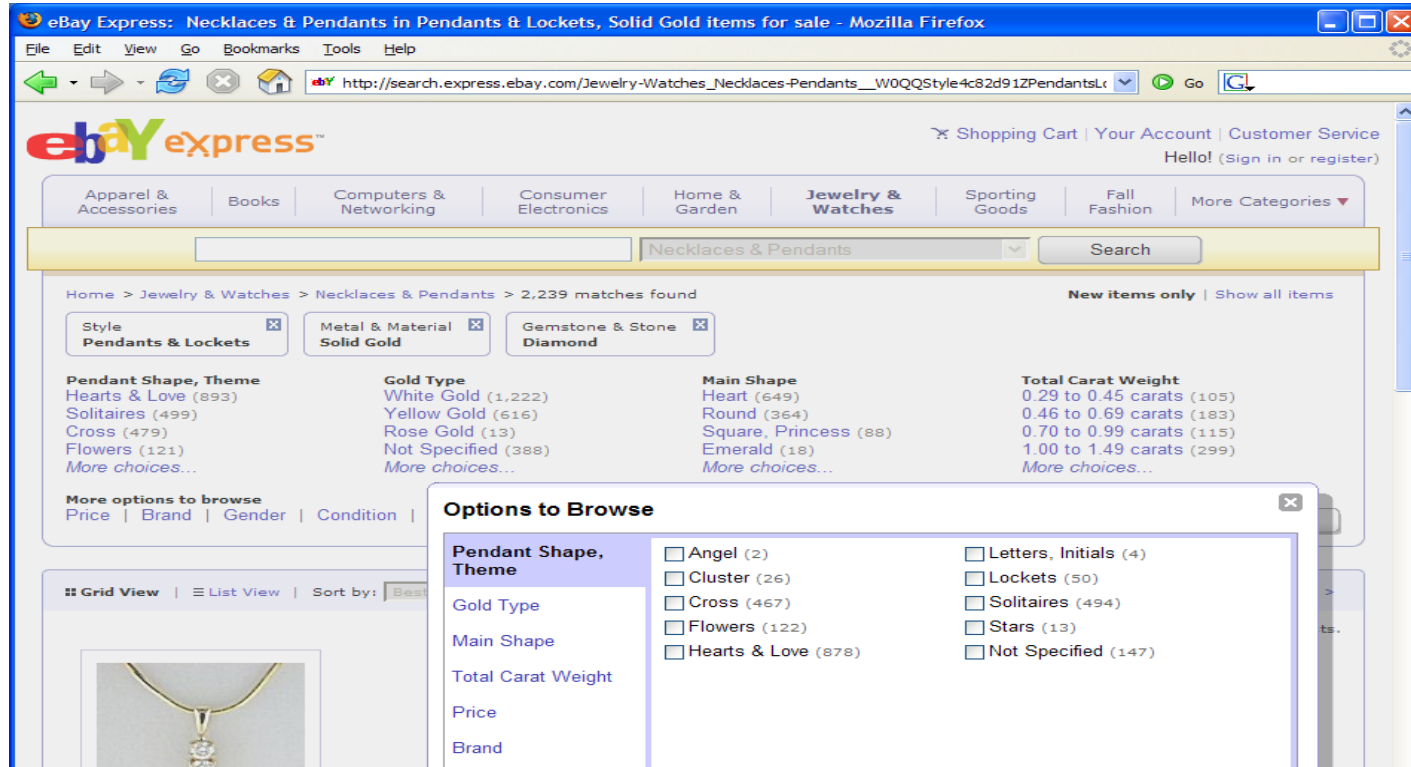
**Contact:** [susan@foodscience.com](mailto:susan@foodscience.com)

**Category:** Travel/Hospitality

**Function:** Food Services

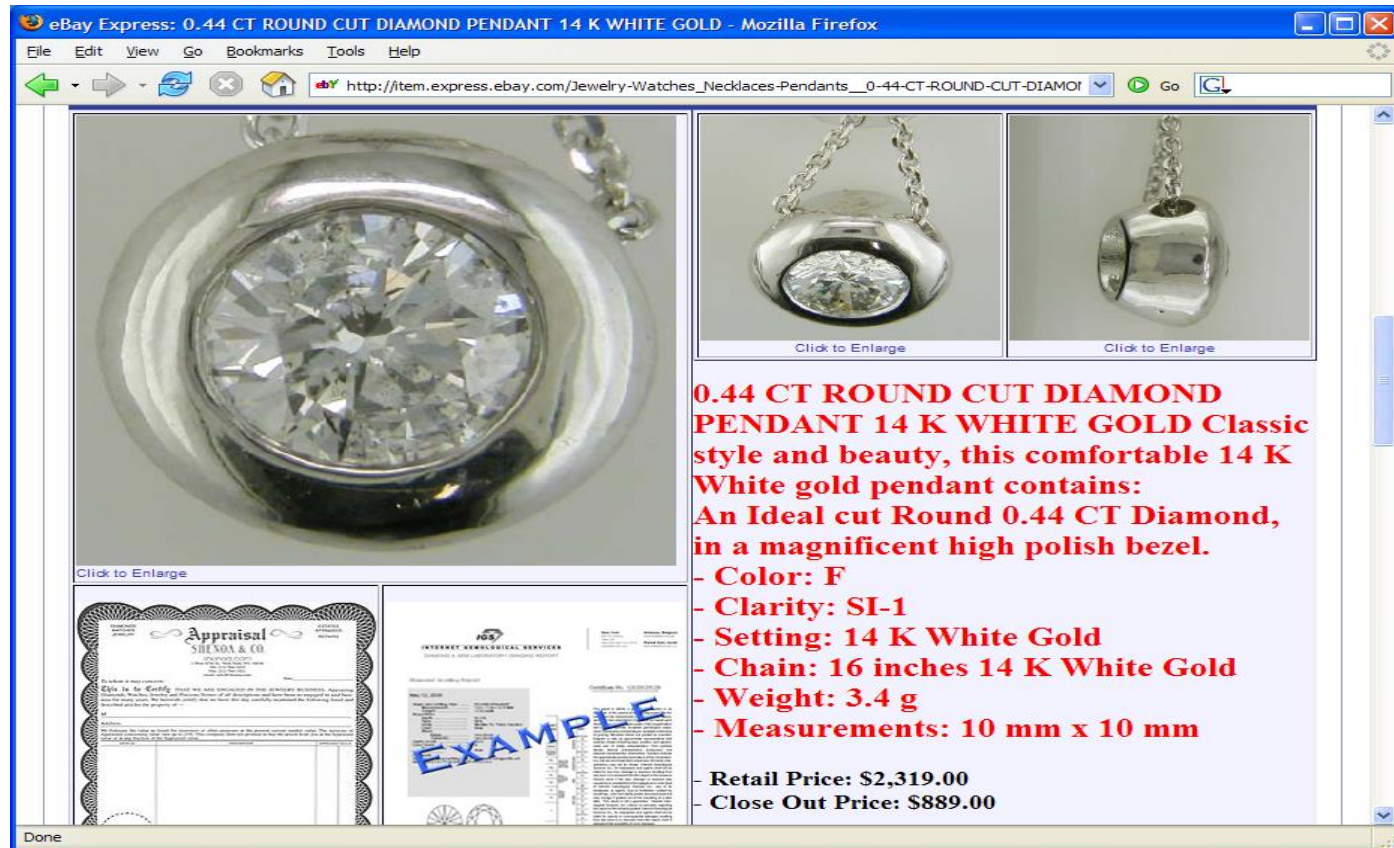


# Another Problem





# Often structured information in text



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

File Edit View Go Bookmarks Tools Help

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

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

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

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

Appraisal SHESSON & CO

INTERNET JEWELLERICAL SERVICES

EXAMPLE

# Another Problem

The screenshot shows a Microsoft Internet Explorer browser window with the following content:

- Address bar:** <http://citeseer.nj.nec.com/peter90critical.html>
- Title bar:** A Critical Evaluation of Commensurable Abduction Models for Semantic Interpretation - Peter, Wi - Microsoft Internet Explorer p
- Page Header:**
  - Left:** **A Critical Evaluation of Commensurable Abduction Models for Semantic Interpretation (1990)** (Correct) (5 citations)  
Peter Norvig Robert Wilensky University of California, Berkeley Computer...  
Thirteenth International Conference on Computational Linguistics, Volume 3
  - Right:** Download: [norvig.com/coling.ps](#)  
Cached: [PS.gz](#) [PS](#) [PDF](#) [DjVu](#) [Image](#) [Update](#) [Help](#)  
From: [norvig.com/resume \(more\)](#)  
Home: [R.Wilensky](#) [HPSearch](#) [\(Correct\)](#)
- Navigation:** [Bookmark](#) [Context](#) [Related](#)
- Buttons:** [\(Enter summary\)](#) [Rate this article: 1 2 3 4 5 \(best\)](#) [Comment on this article](#)
- Abstract:** this paper we critically evaluate three recent abductive interpretation models, those of Charniak and Goldman (1989); Hobbs, Stickel, Martin and Edwards (1988); and Ng and Mooney (1990). These three models add the important property of commensurability: all types of evidence are represented in a common currency that can be compared and combined. While commensurability is a desirable property, and there is a clear need for a way to compare alternate explanations, it appears that a single scalar measure is not enough to account for all types of processing. We present other problems for the abductive approach, and some tentative solutions. [\(Update\)](#)
- Context of citations to this paper:** [More](#)
- Text:** .... (break slight modification of the one given in [Ng and Mooney, 1990] The new definition remedies the anomaly reported in [Norvig and Wilensky, 1990] of occasionally preferring spurious interpretations of greater depths. Table 1: Empirical Results Comparing Coherence and...  
.... costs as probabilities, specifically within the context of using abduction for text interpretation, are discussed in [Norvig and Wilensky \(1990\)](#). The use of abduction in disambiguation is discussed in Kay et al. 1990) We will assume the following: 13) a. Only literals...
- Cited by:** [More](#)
  - [Translation Mismatch in a Hybrid MT System - Gawron \(1999\)](#) (Correct)
  - [Abduction and Mismatch in Machine Translation - Gawron \(1999\)](#) (Correct)
  - [Interpretation as Abduction - Hobbs, Stickel, Appelt, Martin \(1990\)](#) (Correct)
- Active bibliography (related documents):** [More](#) [All](#)
  - 0.1: [Critiquing Effective Decision Support in Time-Critical Domains - Gertner \(1995\)](#) (Correct)
  - 0.1: [Decision Analytic Networks in Artificial Intelligence - Matzkevich, Abramson \(1995\)](#) (Correct)
  - 0.1: [A Probabilistic Network of Predicates - DeRose, Liu \(1992\)](#) (Correct)

# Definition of IE

**Information Extraction** (IE) is the process of extracting structured information (e.g., database tables) from unstructured machine-readable documents (e.g., Web documents).

Elvis Presley was a famous rock singer.

...

Mary once remarked that the only attractive thing about the painter Elvis Hunter was his first name.

**Information  
Extraction**



GName	FName	Occupation
Elvis	Presley	singer
Elvis	Hunter	painter
...	...	

*“Seeing the Web as a table”*

# Defining an IE problem



- In what I will refer to as "classic" IE, we are converting documents to one or more table entries
  - There are other kinds of IE, we will talk about those later
- The **design** of these tables is usually determined by some business need
- Let's look at the table entries for a similar set of examples to the ones we just saw

# Motivating Examples

**579 Jobs in Northern California**

Refine your search

Keyword(s)

(Pipeline) Business Strategy Associate - Release Engineer - Quality Assurance

Senior Flash Memory Technologist - Storage Architect - SSD

---

**Search Results**

Job Title / Description ( show titles only )	Company
<p><b>RN-Registered Nurse/LVN-Licensed Vocational Nurse - View similar jobs</b></p> <p>Job type: Full-Time/Part-Time</p> <p>Maxim's office in Sherman Oaks is seeking compassionate Registered Nurses (RN) and Licensed ... Maxim's office in Sherman Oaks is seeking...</p> <p> <a href="#">View full job description</a> <a href="#">Save to MyCareerBuilder</a> <a href="#">Email to a friend</a> </p>	Maxim Healthcare Services, Inc
<p><b>Nurse Practitioner - Acute Care Nurse Practitioner - View similar jobs</b></p> <p>Job type: Full-Time</p> <p>Vanderbilt University Medical Center is currently hiring Nurse Practitioners to join our team ... Vanderbilt University Medical Center is...</p> <p> <a href="#">View full job description</a> <a href="#">Save to MyCareerBuilder</a> <a href="#">Email to a friend</a> </p>	Vanderbilt University Medical Center (VUMC)

Title	Type	Location
Business strategy Associate	Part time	Palo Alto, CA
Registered Nurse	Full time	Los Angeles
...	...	

# Motivating Examples

Biography for

**Elvis Presley** [More at IMDb](#)

**Date of Birth**

[8 January 1935](#), [Tupelo, Mississippi, USA](#)

**Date of Death**

[16 August 1977](#), [Memphis, Tennessee, USA](#) (cardiac arrhythmia)

**Birth Name**

Elvis Aron Presley

**Nickname**

The Pelvis  
The King  
The King Of Rock 'n'

**Height**

6' (1.83 m)

**Mini Biography**

Elvis Aaron Presley

Name	Birthplace	Birthdate
Elvis Presley	Tupelo, MI	1935-01-08
...	...	...



**Biography**

[Overview](#) / [1935-1957](#) / [1958-1965](#) / [1966-1969](#) / [1970-1977](#)

**Overview**

**Elvis Aaron Presley**, in the humblest of circumstances, was born to Vernon and Gladys Presley in a [two-room house in Tupelo, Mississippi](#) on January 8, 1935. His twin brother, Jessie Garon, was stillborn, leaving Elvis to grow up as an only child. He and his parents moved to [Memphis, Tennessee](#) in 1948, and Elvis graduated from Humes High School there in 1953.

# Motivating Examples

## Information Extraction: Techniques and Challenges

Ralph Grishman

### Information Integration Papers

[Answering Queries Using Templates With Binding Patterns](#). In PODS 1995, specify binding patterns.

[The TSIMMIS Approach to Mediation: Data Models and Languages](#). A survey appears in *J. Intelligent Information Systems* 8:2, pp. 117-132, March, 1997.

Author	Publication	Year
Grishman	Information Extraction...	2006
...	...	...

# Motivating Examples



Ballroom Dance Shoe  
1 new from \$49.95  
 ★★☆☆☆ (5)  
 > Show only So Danca items

X-Strap Ballroom Dance Shoe  
1 new from \$49.95  
 ★★★★★ (5)  
 > Show only So Danca items



**Dynex™ - 32" Class / 720p / 60Hz / LCD HDTV**  
 Model: DX-32L150A11 | SKU: 9558089  
 ★★★★★ 3.8 of 5 (180 reviews)  
 Check Shipping & Availability >

Compare



**Dynex™ - 24" Class / 1080p / 60Hz / LCD HDTV**  
 Model: DX-24L150A11 | SKU: 9848048  
 ★★★★★ 4.3 of 5 (54 reviews)  
 Check Shipping & Availability >

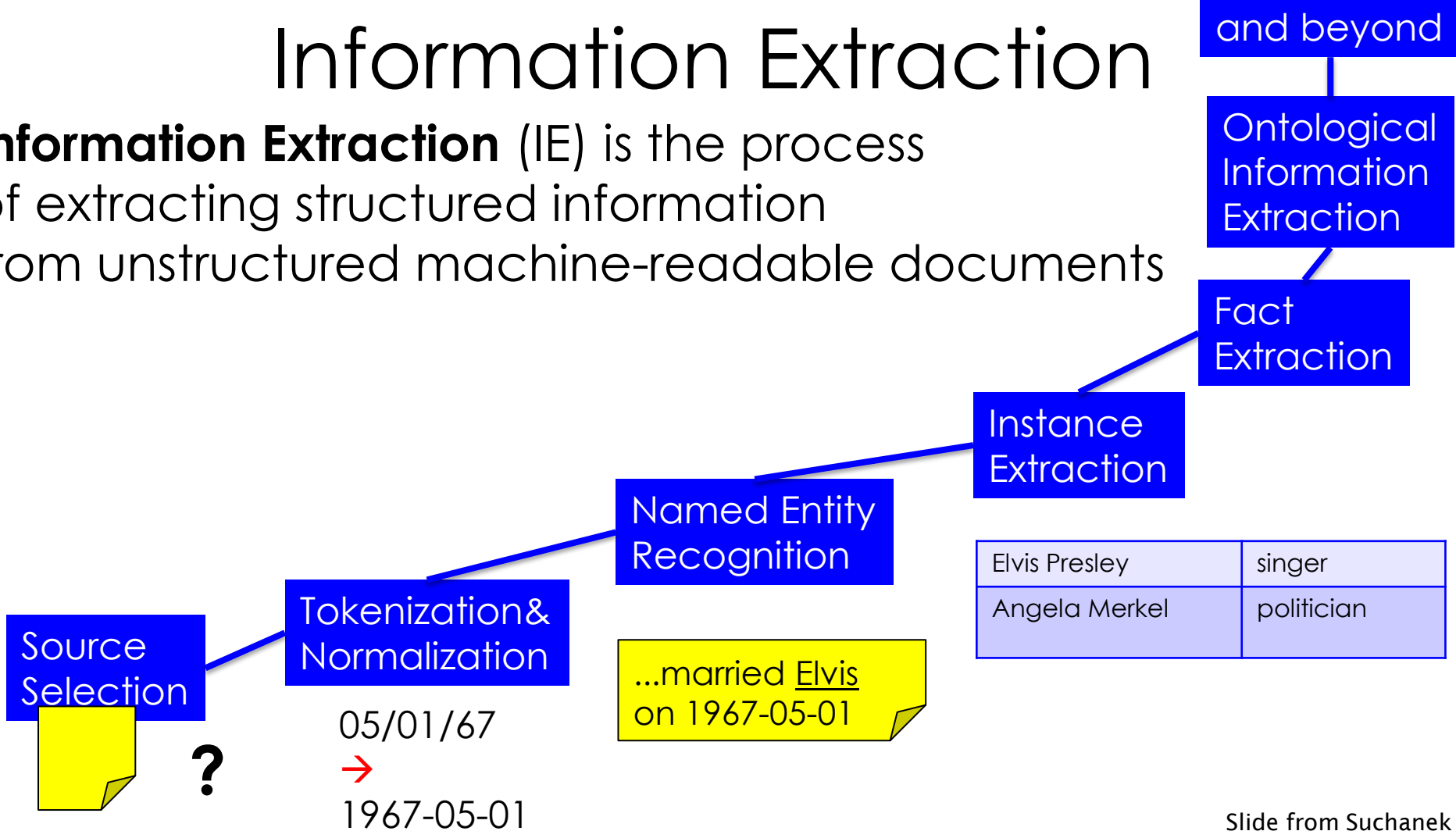
Compare

Product	Type	Price
Dynex 32"	LCD TV	\$1000
...	...	



# Information Extraction

**Information Extraction** (IE) is the process of extracting structured information from unstructured machine-readable documents



# Information Extraction

**Traditional definition:** Recovering structured data from text

**What are some of the sub-problems/challenges?**

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

## Board Members

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

# Information Extraction?

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

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

## Board Members

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

# Information Extraction?

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

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

## Board Members

◦ <b>Itzhak Fisher</b> Chairman of Nielsen BuzzMetrics
◦ <b>Thom Mastrelli</b> Executive Vice President/Corporate Development, VNU
◦ <b>Jonathan Carson</b> CEO of Nielsen BuzzMetrics
◦ <b>Mahendra Vora</b> CEO and Owner, Vora Technology Park

◦ <b>Ori Levy</b> President of Nielsen BuzzMetrics Israel
◦ <b>Ron Schneier</b> Senior Vice President and General Manager, Nielsen Ventures
◦ <b>James O'Hara</b> Senior Vice President and Chief Financial Officer, VNU's Media Measurement and Information Group

# Information Extraction?

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

**James O'Hara (I)**



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

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

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

Sometimes Credited As:  
James Lilburn / Jim O'Hara

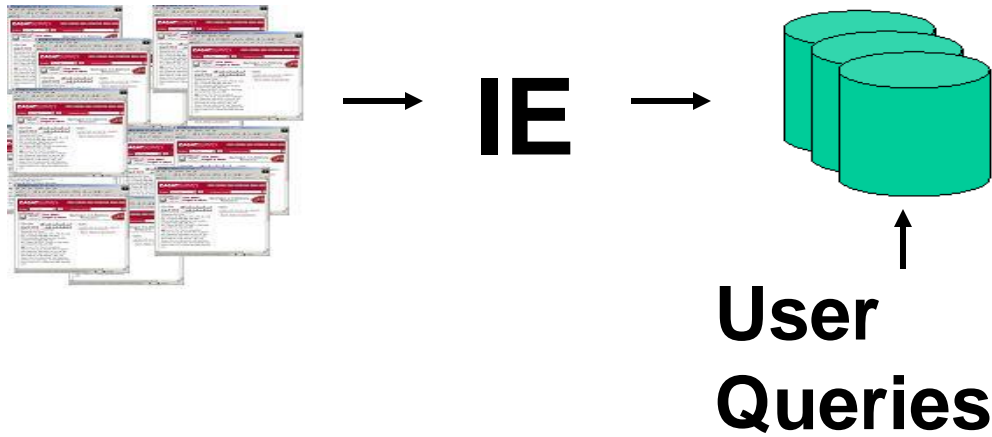
 [IMDbPro Details](#)  [Add IMDb Resume](#)

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

Herkovic	Jane is a member of the Nielsen senior leadership team and a senior member of the VNU MMI Finance team. She is based in New York and reports to both Susan Whiting, president and CEO of Nielsen Media Research, and <b>Jim O'Hara</b> , senior vice president and chief financial officer for VNU Media Measurement and Information.
Susan D. Whiting	
Douglas Darfield	
Paul J. Donato	
Sara Erichson	
Dave Harkness	
Jack Loftus	

# Information extraction

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



# Not all documents are created equal...



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

# Natural Text: MEDLINE

## Journal Abstracts

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

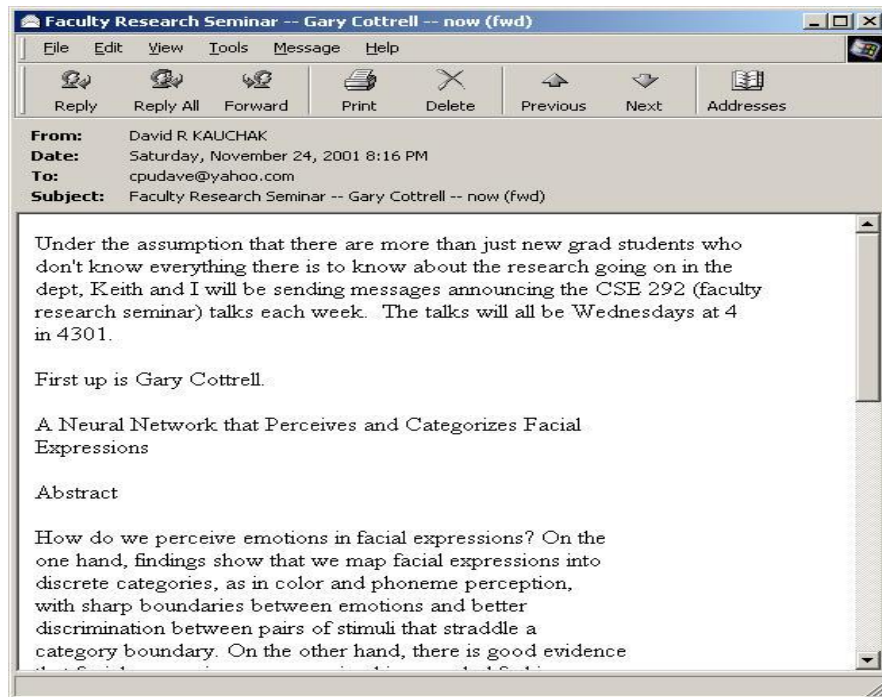
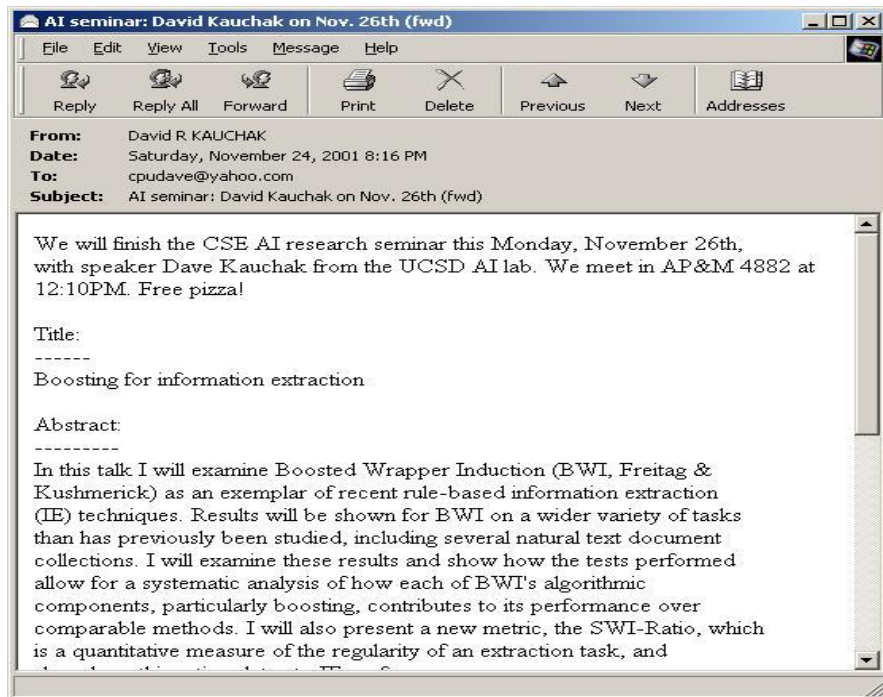
**BACKGROUND:** The most challenging aspect of revision hip surgery is the management of bone loss. A reliable and valid measure of bone loss is important since it will aid in future studies of hip revisions and in preoperative planning. We developed a measure of femoral and acetabular bone loss associated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

There was also significant agreement ( $p < 0.0001$ ) between each rater's assessment and the intraoperative assessment.



# Partially Structured: Seminar Announcements

Extract time, location, speaker, etc.





# Information extraction pipeline

For years, Microsoft Corporation CEO Bill Gates was against open source. But today he appears to have changed his mind. "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...

Name	Title	Organization
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	Founder	Free Soft..

# The Full Task of Information Extraction

As a family of techniques:

Information Extraction =  
segmentation + classification + association + clustering

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.

Now 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."

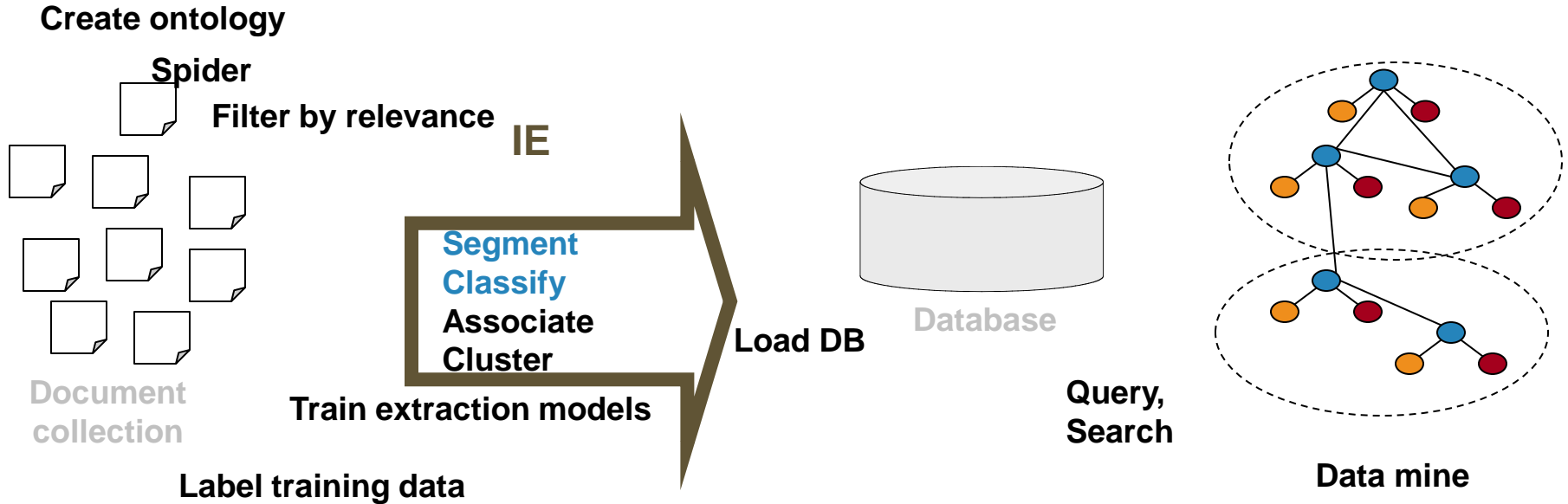
Richard Stallman, founder of the Free Software Foundation, countered saying...

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



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

# An Even Broader View



# Landscape of IE Tasks: Document 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.

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.

- Press
- Contact**
- General information
- Directions maps

Non-grammatical snippets,  
rich formatting & links

<b>Barto, Andrew G.</b> Professor. Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.	(413) 545-2109	<a href="mailto:barto@cs.umass.edu">barto@cs.umass.edu</a>	CS276	 
<b>Berger, Emery D.</b> Assistant Professor.	(413) 577-4211	<a href="mailto:emery@cs.umass.edu">emery@cs.umass.edu</a>	CS344	 
<b>Brock, Oliver</b> Assistant Professor.	(413) 577-0334	<a href="mailto:oli@cs.umass.edu">oli@cs.umass.edu</a>	CS246	 
<b>Clarke, Lori A.</b> Professor. Software verification, testing, and analysis; software architecture and design.	(413) 545-1328	<a href="mailto:clarke@cs.umass.edu">clarke@cs.umass.edu</a>	CS304	 

## Tables

8:30 - 9:30 AM	<b>Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty</b> <i>Joseph Y. Halpern, Cornell University</i>				
9:30 - 10:00 AM	Coffee Break				
10:00 - 11:30 AM	Technical Paper Sessions:				
<b>Cognitive Robotics</b>	<b>Logic Programming</b>	<b>Natural Language Generation</b>	<b>Complexity Analysis</b>	<b>Neural Networks</b>	<b>Games</b>
739: A Logical Account of Causal and Topological Maps <i>Emilio Remolina and Benjamin Kuipers</i>	116: A-System: Problem Solving through Abduction <i>Marc Donecker, Antonis Kakas, and Bert Van Nuffelen</i>	758: Title Generation for Machine-Translated Documents <i>Rong Jin and Alexander G. Hauptmann</i>	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories <i>Marco Cadoli, Thomas Eiter, and Georg Gottlob</i>	179: Knowledge Extraction and Comparison from Local Function Networks <i>Konrath McGarry, Stefan Wermter, and John MacIntyre</i>	71: Iterative Widening <i>Tristan Cazenave</i>
549: Online-Execution of ccGolog Plans <i>Henrik Grosskreutz and Gerhard Lakemeyer</i>	131: A Comparative Study of Logic Programs with Preference <i>Torsten Schaub and Kewen</i>	246: Dealing with Dependencies between Content Planning and Surface Realisation in a Pipeline Generation	470: A Perspective on Knowledge Compilation <i>Adnan Darwiche and Pierre Marquis</i>	258: Violation-Guided Learning for Constrained Formulations in Neural-Network Time-Series	353: Temporal Difference Learning Applied to a High Performance Game-Playing

# Landscape of IE Tasks

## Intended Breadth of Coverage

### Web site specific

#### Formatting

Amazon.com Book Pages

amazon.com. VIEW CART

WELCOME BROWSE SUBJECTS

SEARCH BROWSE SUBJECTS BESTSELLERS MAGAZINES CORPORATE ACCOUNT

NEW Super Saver Shipping FREE on orders over \$35

**Learning in Graphical Models**  
by Michael Irwin Jordan (Editor)

List Price: \$60.00  
Price: \$60.00

This item ships for FREE with Super Saver Shipping

Availability: Usually ships within 2 to 3 days

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

Buy both now!

### Genre specific

#### Layout

Resumes

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 research interests lie in the automated analysis of data for the purposes of identifying...  
est...  
test...  
cre...

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  
(412) 683-8036

Office Phone  
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 Detection and Tracking project at CMU, under the direction of Yoram Singer. The

### Wide, non-specific

#### Language

University Names

8:30 - 9:30 AM	Invited Talk: <b>Plausibility Measures: A General Approach</b> <i>Joseph Y. Halpern, Cornell University</i>		
9:30 - 10:00 AM	Coffee Break		
10:00 - 11:30 AM	Technical Paper Sessions:		
<b>Cognitive Robotics</b>	<b>Logic Programming</b>	<b>Natural Language Generation</b>	<b>Complexity Analysis</b>
739: A Logical Account of Causal and Topological Maps <i>Emilio Remolina and Benjamin Kuipers</i>	116: A-System: Problem Solving through Abduction <i>Marc Denecker,</i>	758: Title Generation for Machine-Translated Documents <i>Rong Jin and Alexander G. Hauptmann</i>	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality

**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
- General information
- Directions maps



# Landscape of IE Tasks :

## Complexity of entities/relations

### 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 is 412-268-1299

### Complex pattern

#### U.S. postal addresses

University of Arkansas

P.O. Box 140

Hope, Al

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:

## Arity of relation

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 Welch

*In:* Jeffrey Immelt

*"Named entity" extraction*

# Association task = Relation Extraction

- Checking if groupings of entities are instances of a relation
  1. Manually engineered rules
    - Rules defined over words/entities: “<company> located in <location>”
    - Rules defined over parsed text:
      - “((Obj <company>) (Verb located) (\*) (Subj <location>))”
  2. Machine Learning-based
    - Supervised: Learn relation classifier from examples
    - Partially-supervised: bootstrap rules/patterns from “seed” examples

# Relation Extraction: Disease Outbreaks

May 19 1995, Atlanta -- The Centers for Disease Control and Prevention, which is in the front line of the world's response to the deadly Ebola epidemic in Zaire, is finding itself hard pressed to cope with the crisis...

**Information  
Extraction System**

<i>Date</i>	<i>Disease Name</i>	<i>Location</i>
Jan. 1995	Malaria	Ethiopia
July 1995	Mad Cow Disease	U.K.
Feb. 1995	Pneumonia	U.S.

# Relation Extraction: Protein Interactions

“We show that CBF-A and CBF-C interact with each other to form a CBF-A-CBF-C complex and that CBF-B does not interact with CBF-A or CBF-C individually but that it associates with the CBF-A-CBF-C complex.”

CBF-A  $\xleftrightarrow[\text{complex}]{\text{interact}}$  CBF-C

CBF-B  $\xrightarrow{\text{associates}}$  CBF-A-CBF-C complex

# Binary Relation Association as Binary Classification

**Christos Faloutsos** conferred with **Ted Senator**, the **KDD 2003 General Chair**.

Person

Person

Role

Person-Role (**Christos Faloutsos**, **KDD 2003 General Chair**) → NO

Person-Role ( **Ted Senator**, **KDD 2003 General Chair**) → YES

# Resolving coreference (both within and across documents)

John Fitzgerald Kennedy was born at 83 Beals Street in Brookline, Massachusetts on Tue 29, 1917, at 3:00 pm,[7] the second son of Joseph P. Kennedy, Sr., and Rose Fitzgerald; R turn, was the eldest child of John "Honey Fitz" Fitzgerald, a prominent Boston political fi was the city's mayor and a three-term member of Congress. Kennedy lived in Brookline years and attended Edward Devotion School, Noble and Greenough Lower School, and the School, through 4th grade. In 1927, the family moved to 5040 Independence Avenue in l Bronx, New York City: two years later, they moved to 294 Pondfield Road in Bronxville, N where Kennedy was a member of Scout Troop 2 (and was the first Boy Scout to become President).[8] Kennedy spent summers with his family at their home in Hyannisport, Massachusetts, and Christmas and Easter holidays with his family at their winter home in Beach, Florida. For the 5th through 7th grade, Kennedy attended Riverdale Country School, a private school for boys. For 8th grade in September 1930, the 13-year old Kennedy attended Canterbury School in New Milford, Connecticut.



# Rough Accuracy of Information Extraction

Information type	Accuracy
Entities	90-98%
Attributes	80%
Relations	60-70%
Events	50-60%

- Errors cascade (error in entity tag → error in relation extraction)
- These are very rough, actually optimistic, numbers
  - Hold for well-established tasks, but lower for many specific/novel IE tasks

# What we will cover in this class (briefly)

- History of IE, Related Fields
- Source Selection (which text?)
- Tokenization and Normalization
- Named Entity Recognition
- Instance Extraction
- Fact/Event Extraction
- Ontological IE/Open IE
- Probably: multilingual extraction
- Some of your suggestions, which you will give in the practical session



# Seminar

- You attend EITHER Thursdays (starting tomorrow) or Wednesdays (starting next week)
- Survey: PUT YOUR NAME ON THIS
- Quiz/feedback: optionally \*anonymous\*
  
- Also, don't forget the reading for next week!
- Sarawagi: Information Extraction. Introduction

- Thank you for your attention!