Information Extraction

CIS, LMU München
Winter Semester 2022-2023

Prof. Dr. Alexander Fraser, CIS

Information Extraction – Administravia - I

- Vorlesung
  - Learn the basics of Information Extraction (IE)
  - Weekly lecture from me here in BU 101
  - Exactly the same scientific content as last year
    - "Administravia" is slightly different, see these slides (also uploaded after class)
    - 2020-2021 videos are available, use these to review
- Seminar
  - You will present an oral report and then submit a written version later
  - Probably some additional lectures (*not* part of the Vorlesung/Klausur) and some simple exercises

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!)
    - Please **ignore the Modulteilprüfung** entries
    - Make sure you are registered for the Seminar and the Vorlesung
  - There are a good number of people only in the Vorlesung
  - There are just a couple of people only in the Seminar

Information Extraction – Administravia - III

- Vorlesung and Seminar are two separate courses (in same module for CIS people)
  - However, there will be some shifting around of slots
- Vorlesung (Grade):
  - Klausur in February entirely determines the Vorlesung grade
- Seminar (Grade):
  - Referat
  - Hausarbeit (write-up of the Referat) (6 pages, due **3 weeks** after you hold your Referat)
- Grades of Vorlesung and Seminar are independent
- CIS-ler: No Notenverbesserung

Information Extraction – Administravia - IV

- Syllabus: see WS last year
  - Brief idea at end of this slide deck
- List of Referatsthemen for the Seminar
  - I will announce when the presentation of topics will take place soon, this will be in early November
- 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/)

Information Extraction – Administravia - V

- DISCUSSION: we have very few students in the seminar this year.
  - Should we have one seminar (on Wed at 10:00) or two (on Wed at 10:00 and Thurs at 12:00)?

- 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

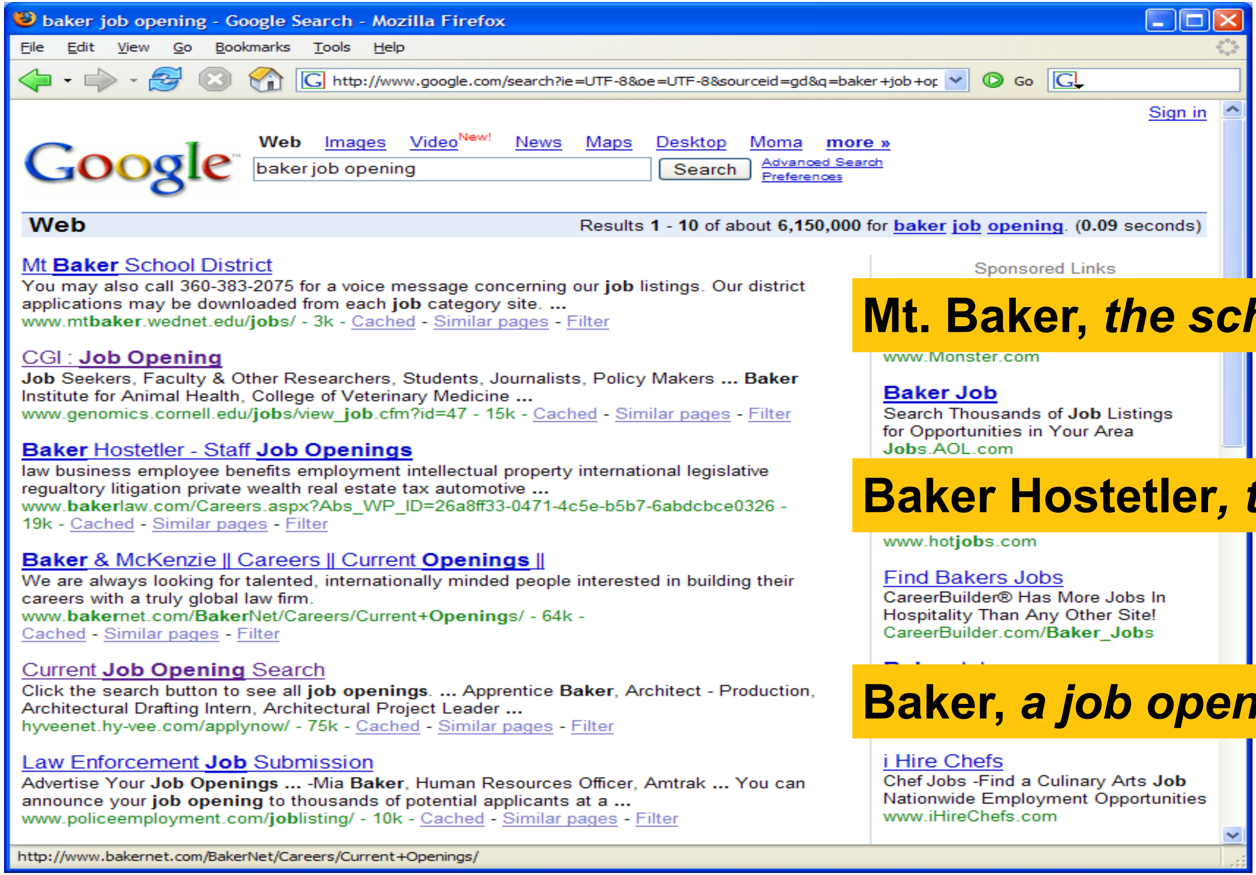
My Biases

- As you may have noticed by now: I am from the US (PhD in Computer Science from USC/ISI, Artifical Intelligence division)
- I am a professor here at CIS
- I do research in the broad area of **statistical NLP**
  - I mostly work on **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



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

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

CGI : **Job Opening**
Job Seekers, Faculty & Other Researchers, Students, Journalists, Policy Makers ... **Baker** Institute for Animal Health, College of Veterinary Medicine ...
www.genomics.cornell.edu/**jobs**/view_**job**.cfm?id=47 - 15k - Cached - Similar pages - Filter

Baker Hostetler - Staff **Job Openings**
law business employee benefits employment intellectual property international legislative regualtory litigation private wealth real estate tax automotive ...
www.**baker**law.com/Careers.aspx?Abs_WP_ID=26a8ff33-0471-4c5e-b5b7-6abdcbce0326 - 19k - Cached - Similar pages - Filter

Baker & McKenzie || Careers || Current **Openings** ||
We are always looking for talented, internationally minded people interested in building their careers with a truly global law firm.
www.**baker**net.com/**Baker**Net/Careers/Current+**Openings**/ - 64k - Cached - Similar pages - Filter

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

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

Sponsored Links

www.Monster.com

Baker Job
Search Thousands of **Job** Listings for Opportunities in Your Area
Jobs.AOL.com

www.hot**jobs**.com

Find Bakers Jobs
CareerBuilder® Has More Jobs In Hospitality Than Any Other Site!
CareerBuilder.com/**Baker**_**Jobs**

i Hire Chefs
Chef Jobs -Find a Culinary Arts **Job** Nationwide Employment Opportunities
www.iHireChefs.com

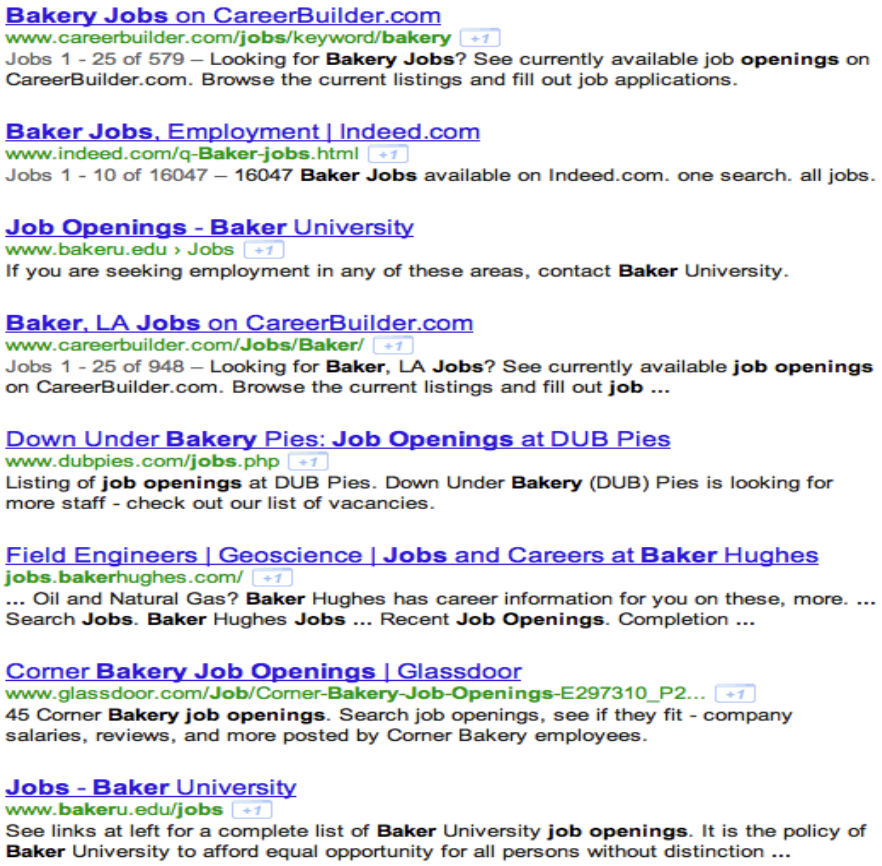
http://www.bakernet.com/BakerNet/Careers/Current+Openings/

Mt. Baker, *the school district*

Baker Hostetler, *the company*

Baker, *a job opening*

Slide from Cohen/Mccallum



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

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

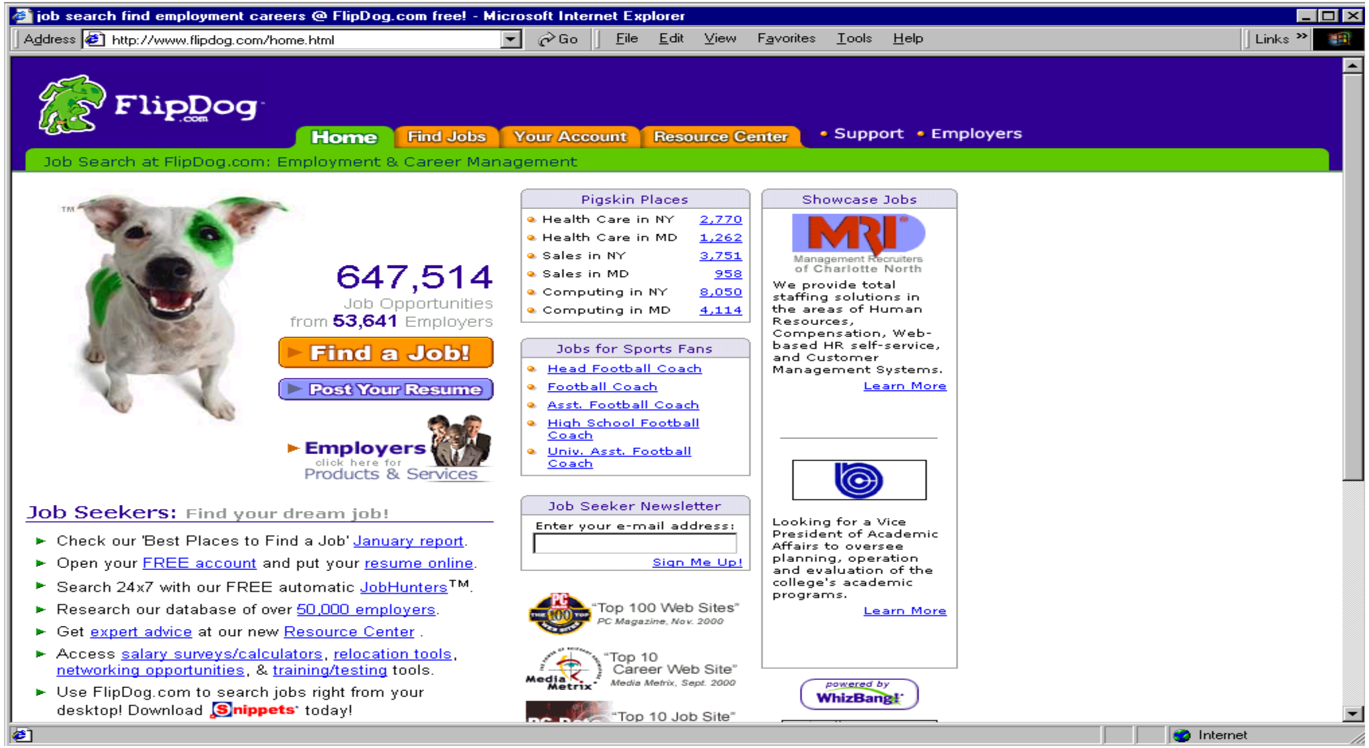
Field Engineers | Geoscience | **Jobs** and Careers at **Baker** Hughes
jobs.**baker**hughes.com/
... Oil and Natural Gas? **Baker** Hughes has career information for you on these, more. ... Search **Jobs**. **Baker** Hughes **Jobs** ... Recent **Job Openings**. Completion ...

Corner **Bakery Job Openings** | Glassdoor
www.glassdoor.com/**Job**/Corner-**Bakery**-**Job**-**Openings**-E297310_P2...
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.**baker**u.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 ...

Slide from Kauchak

A solution



Job Search at FlipDog.com: Employment & Career Management

647,514 Job Opportunities from 53,641 Employers

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:

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

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

"Top 10 Job Site"

Showcase Jobs

MRI Management Recruiters of Charlotte North

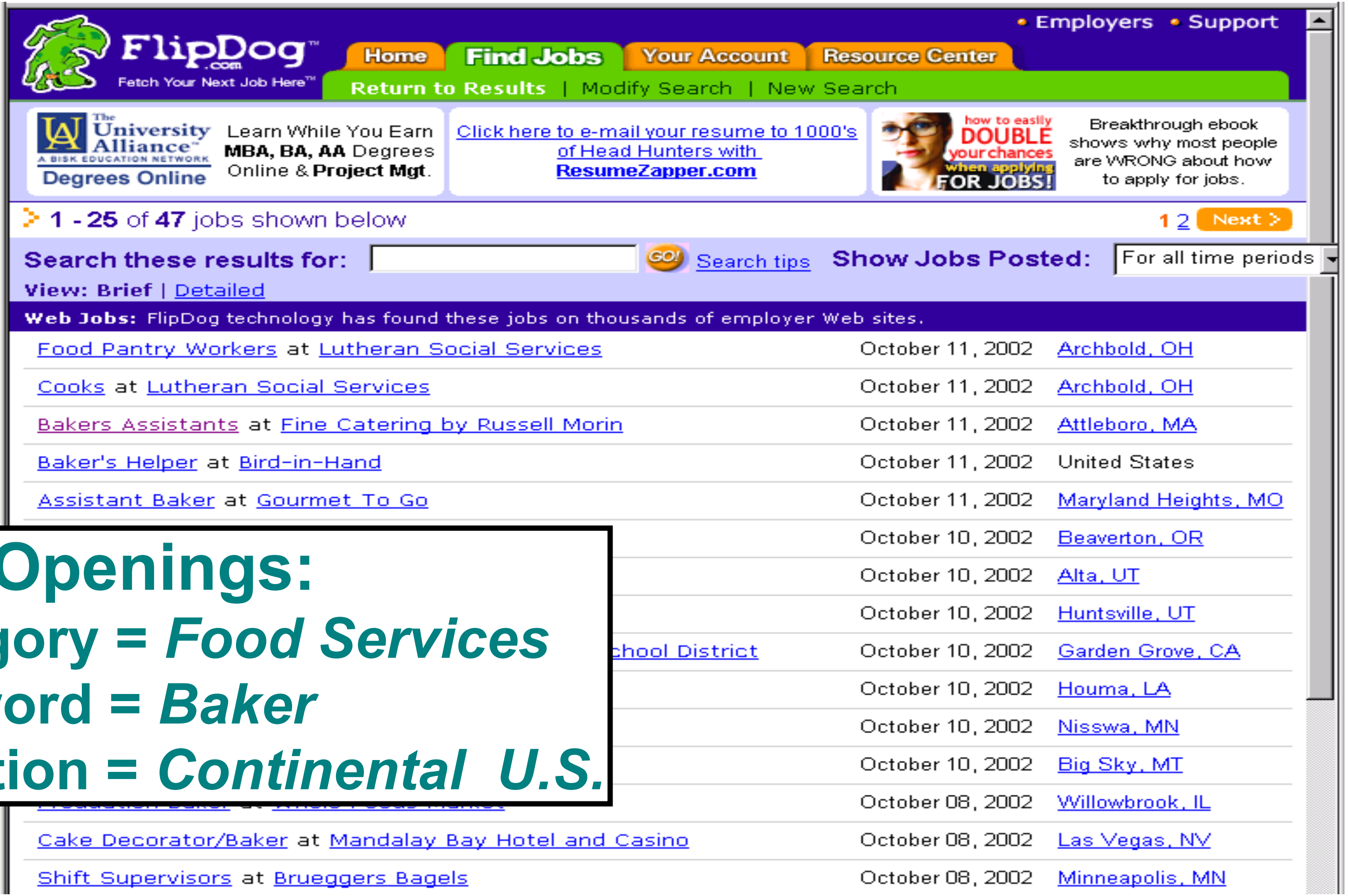
We provide total staffing solutions in the areas of Human Resources, Compensation, Web-based HR self-service, and Customer Management Systems. Learn More

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

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!

Slide from Cohen/McCallum



1 - 25 of **47** jobs shown below

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

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

Job Openings:

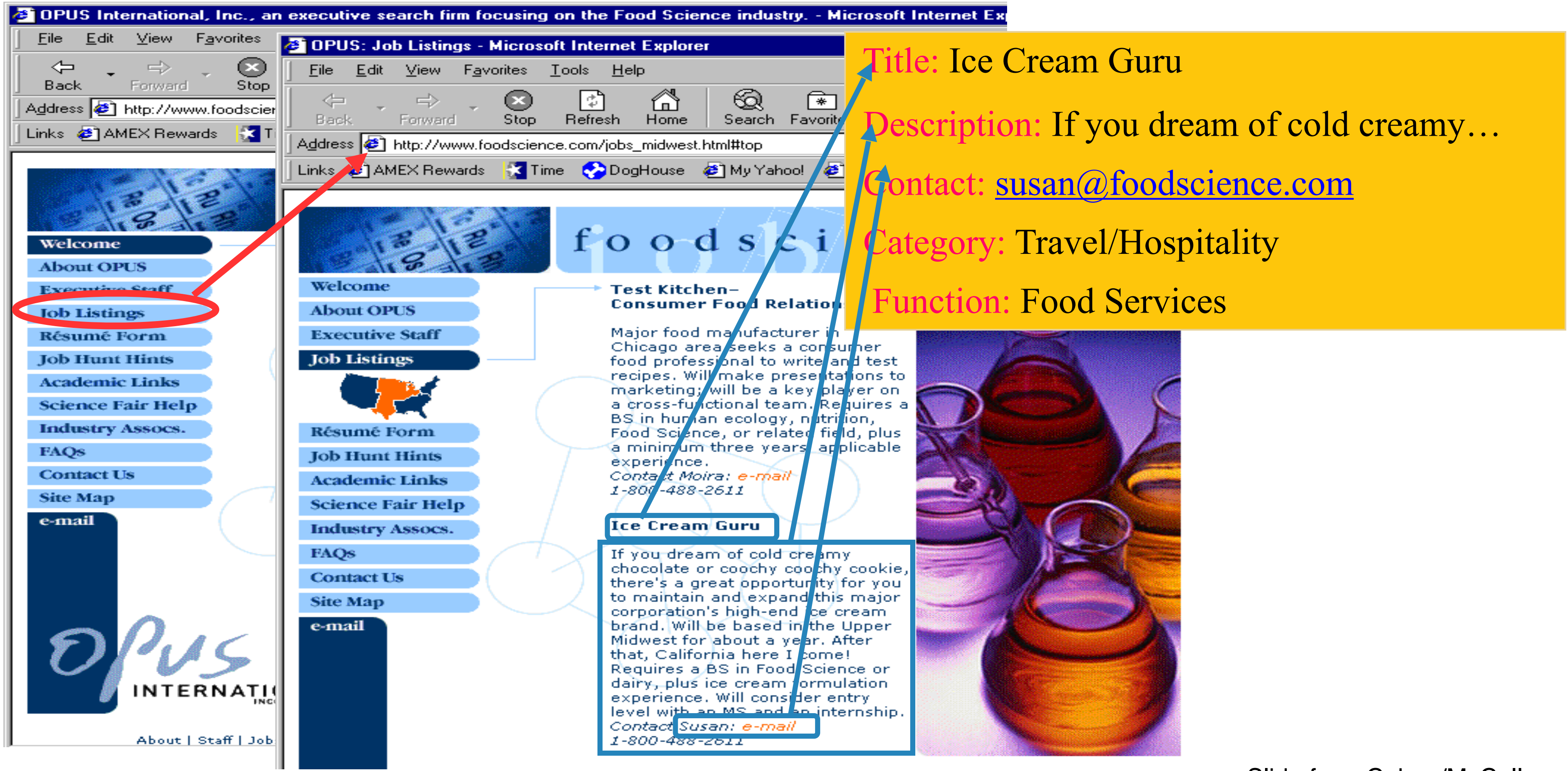
Category = ***Food Services***

Keyword = ***Baker***

Location = ***Continental U.S.***

Slide from Cohen/McCallum

Extracting Job Openings from the Web



Test Kitchen– Consumer Food Relation

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

Ice Cream Guru

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

Title: Ice Cream Guru

Description: If you dream of cold creamy…

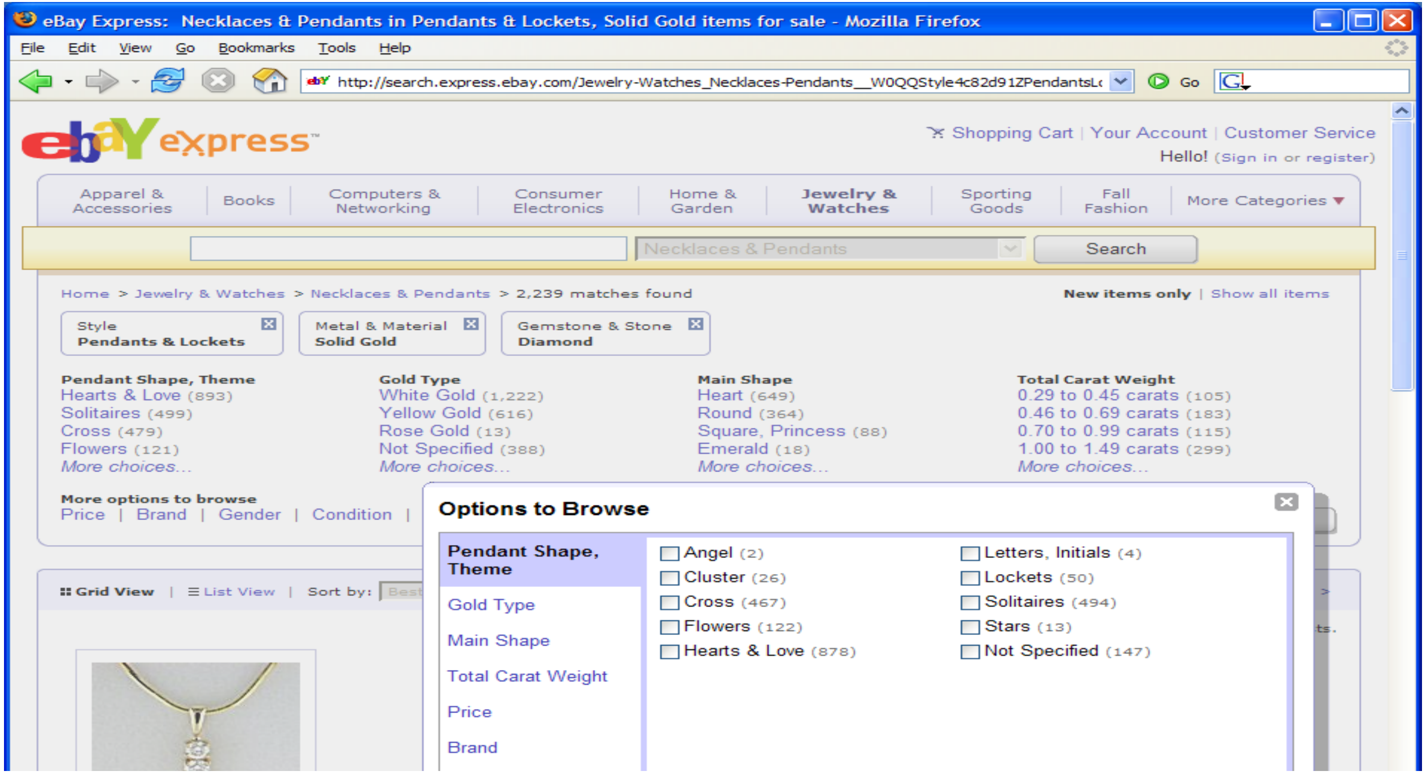
Contact: susan@foodscience.com

Category: Travel/Hospitality

Function: Food Services

Slide from Cohen/McCallum

Another Problem


eBay Express: Necklaces & Pendants in Pendants & Lockets, Solid Gold items for sale - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://search.express.ebay.com/Jewelry-Watches_Necklaces-Pendants__W0QQStyle4c82d91ZPendantsL
Go
ebY express
Shopping Cart | Your Account | Customer Service
Hello! (Sign in or register)
Apparel & Accessories
Books
Computers & Networking
Consumer Electronics
Home & Garden
Jewelry & Watches
Sporting Goods
Fall Fashion
More Categories
Necklaces & Pendants
Search

Home > Jewelry & Watches > Necklaces & Pendants > 2,239 matches found

New items only | Show all items

Style **Pendants & Lockets** ☒ | Metal & Material **Solid Gold** ☒ | Gemstone & Stone **Diamond** ☒

Pendant Shape, Theme	Gold Type	Main Shape	Total Carat Weight
Hearts & Love (893)	White Gold (1,222)	Heart (649)	0.29 to 0.45 carats (105)
Solitaires (499)	Yellow Gold (616)	Round (364)	0.46 to 0.69 carats (183)
Cross (479)	Rose Gold (13)	Square, Princess (88)	0.70 to 0.99 carats (115)
Flowers (121)	Not Specified (388)	Emerald (18)	1.00 to 1.49 carats (299)
More choices...	*More choices...*	*More choices...*	*More choices...*

More options to browse
Price | Brand | Gender | Condition |

Grid View | List View | Sort by: Best

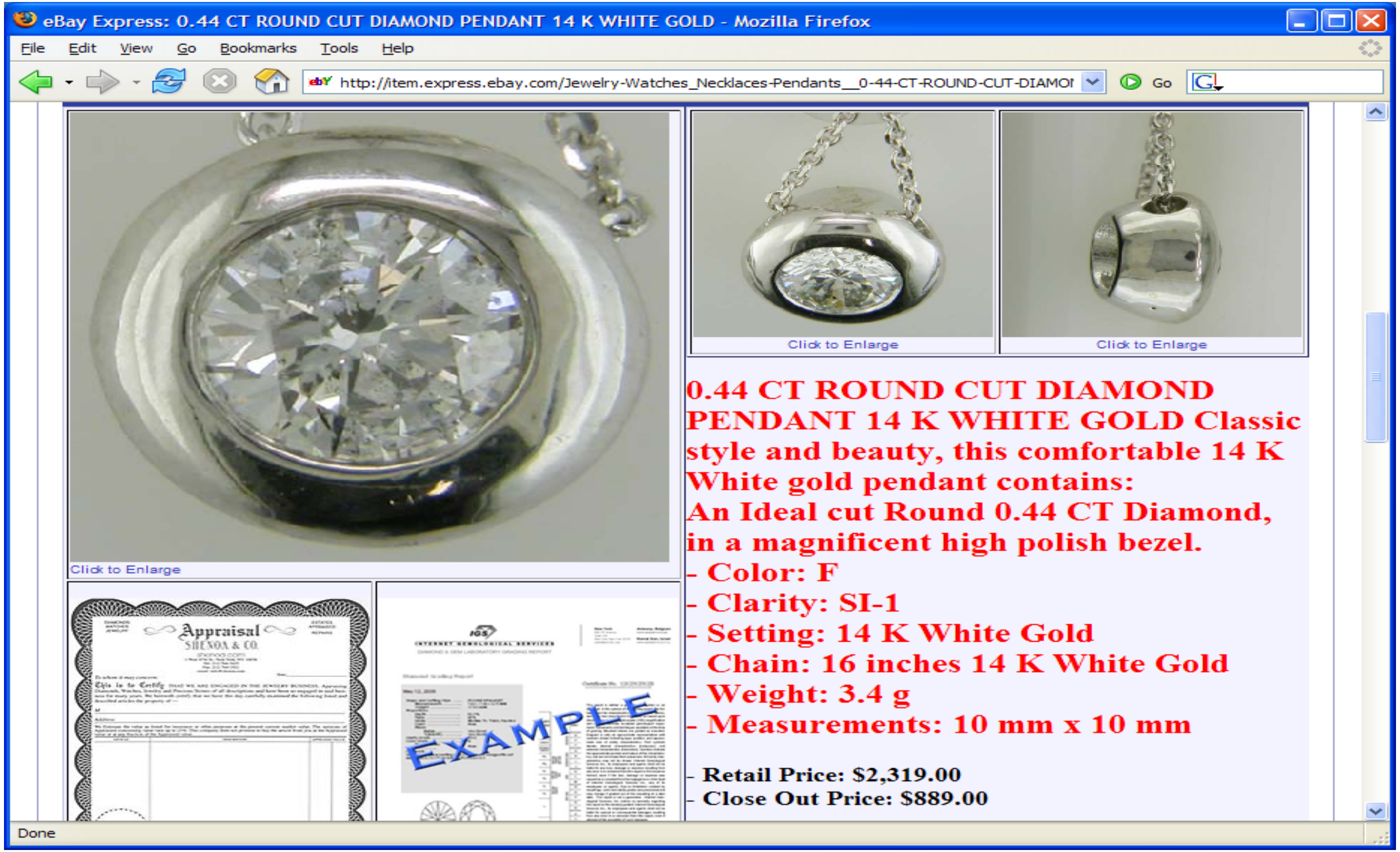
Options to Browse ☒

Pendant Shape, Theme
Gold Type
Main Shape
Total Carat Weight
Price
Brand

☐ Angel (2)
☐ Cluster (26)
☐ Cross (467)
☐ Flowers (122)
☐ Hearts & Love (878)
☐ Letters, Initials (4)
☐ Lockets (50)
☐ Solitaires (494)
☐ Stars (13)
☐ Not Specified (147)

Slide from Cohen/McCallum

Often structured information in text

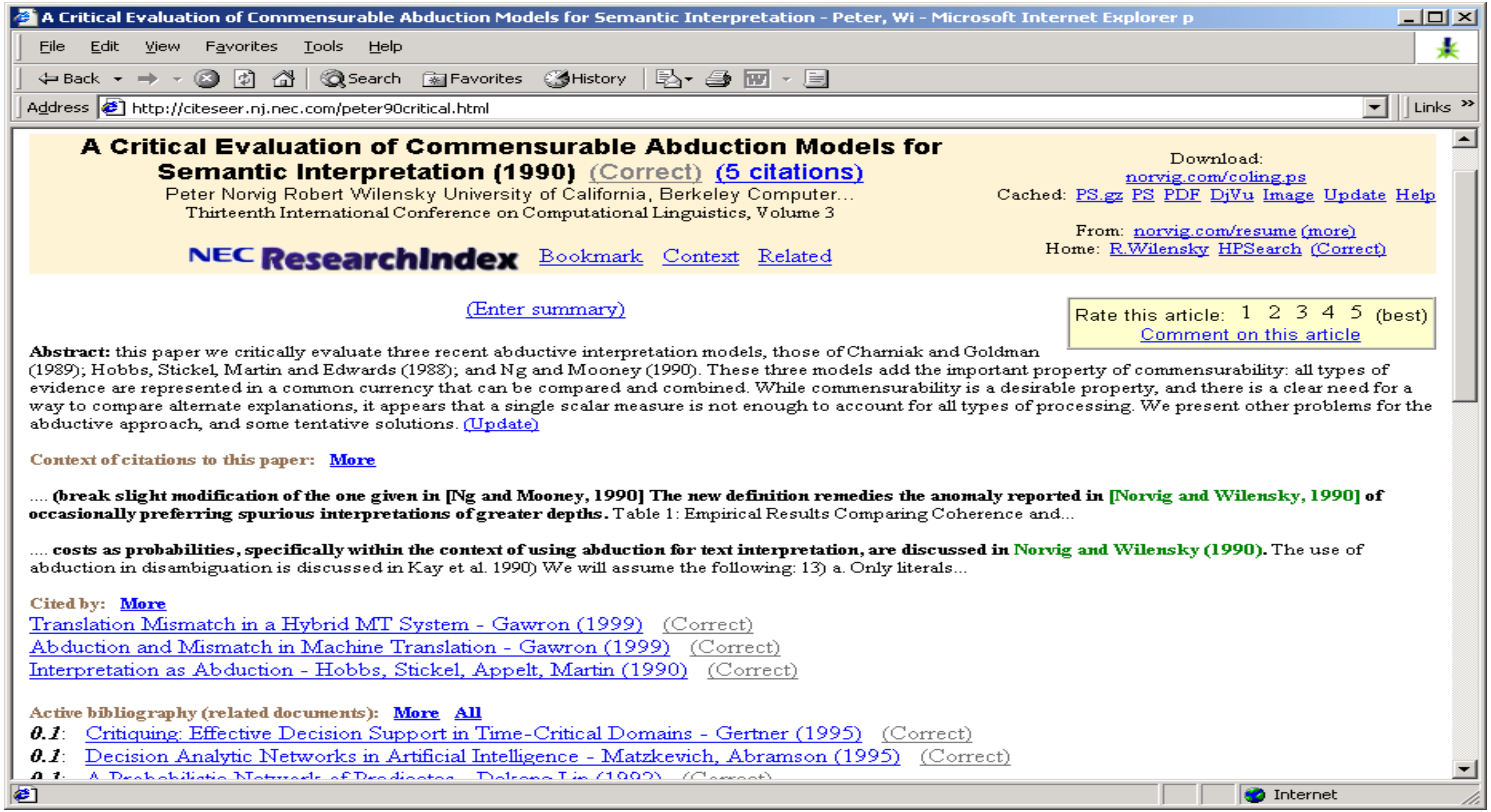

eBay Express: 0.44 CT ROUND CUT DIAMOND PENDANT 14 K WHITE GOLD - Mozilla Firefox
File
Edit
View
Go
Bookmarks
Tools
Help
http://item.express.ebay.com/Jewelry-Watches_Necklaces-Pendants__0-44-CT-ROUND-CUT-DIAMOI
Go
Click to Enlarge
Click to Enlarge
Click to Enlarge
Appraisal
SHENOA & CO.
EXAMPLE
Done

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

Slide from Cohen/McCallum

Another Problem


A Critical Evaluation of Commensurable Abduction Models for Semantic Interpretation - Peter, Wi - Microsoft Internet Explorer p
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://citeseer.nj.nec.com/peter90critical.html
Links
Internet

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

NEC ResearchIndex Bookmark Context Related

Download:
norvig.com/coling.ps
Cached: PS.gz PS PDF DjVu Image Update Help

From: norvig.com/resume (more)
Home: R.Wilensky HPSearch (Correct)

(Enter summary)

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

Abstract: this paper we critically evaluate three recent abductive interpretation models, those of Charniak and Goldman (1989); Hobbs, Stickel, Martin and Edwards (1988); and Ng and Mooney (1990). These three models add the important property of commensurability: all types of evidence are represented in a common currency that can be compared and combined. While commensurability is a desirable property, and there is a clear need for a way to compare alternate explanations, it appears that a single scalar measure is not enough to account for all types of processing. We present other problems for the abductive approach, and some tentative solutions. (Update)

Context of citations to this paper: **More**

.... **(break slight modification of the one given in [Ng and Mooney, 1990] The new definition remedies the anomaly reported in [Norvig and Wilensky, 1990] of occasionally preferring spurious interpretations of greater depths.** Table 1: Empirical Results Comparing Coherence and...

.... **costs as probabilities, specifically within the context of using abduction for text interpretation, are discussed in Norvig and Wilensky (1990).** The use of abduction in disambiguation is discussed in Kay et al. 1990) We will assume the following: 13) a. Only literals...

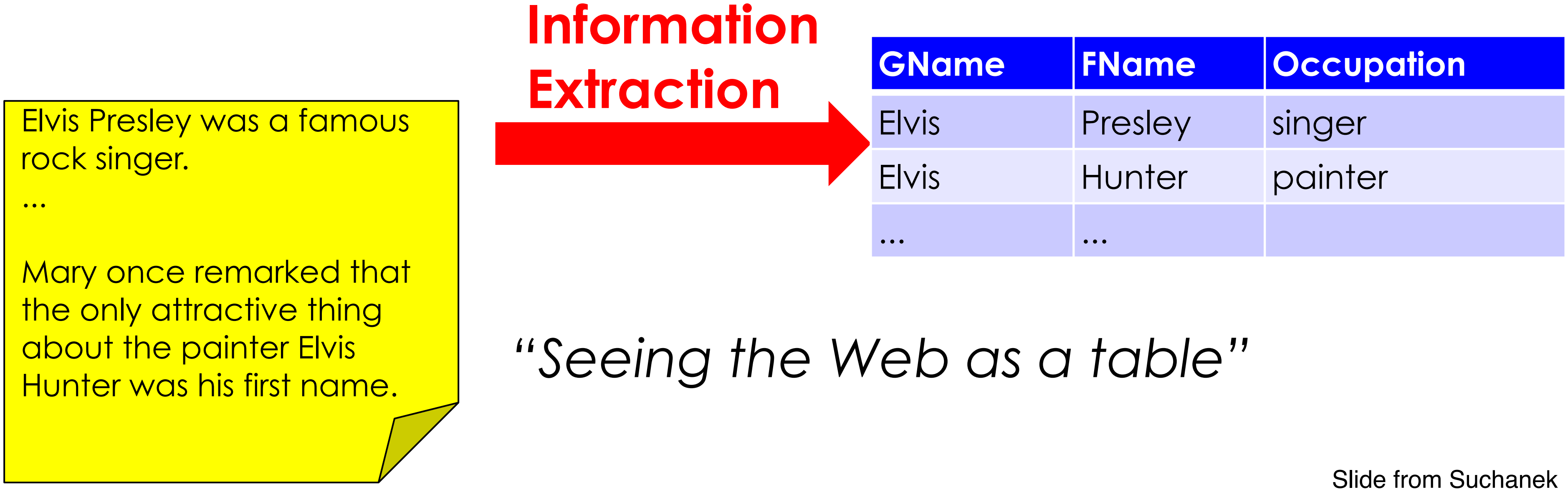
Cited by: **More**
Translation Mismatch in a Hybrid MT System - Gawron (1999) (Correct)
Abduction and Mismatch in Machine Translation - Gawron (1999) (Correct)
Interpretation as Abduction - Hobbs, Stickel, Appelt, Martin (1990) (Correct)

Active bibliography (related documents): **More** **All**
0.1: Critiquing: Effective Decision Support in Time-Critical Domains - Gertner (1995) (Correct)
0.1: Decision Analytic Networks in Artificial Intelligence - Matzkevich, Abramson (1995) (Correct)
0.1: A Probabilistic Network of Predicates - Dekang Lin (1992) (Correct)

Slide from Cohen/McCallum

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.

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

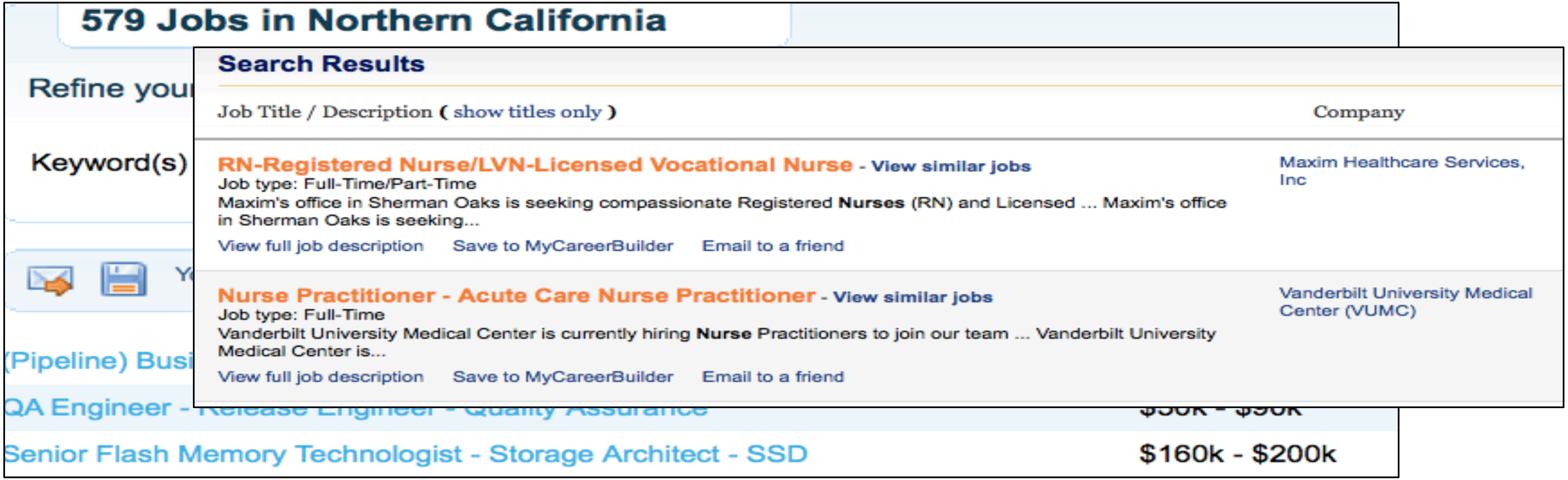
"Seeing the Web as a table"

Slide from Suchanek

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 you

Keyword(s)

Yo

(Pipeline) Busi

QA Engineer - Release Engineer - Quality Assurance $50k - $90k

Senior Flash Memory Technologist - Storage Architect - SSD $160k - $200k

Search Results

Job Title / Description (show titles only) Company

RN-Registered Nurse/LVN-Licensed Vocational Nurse - View similar jobs — Maxim Healthcare Services, Inc

Job type: Full-Time/Part-Time

Maxim's office in Sherman Oaks is seeking compassionate Registered **Nurses** (RN) and Licensed ... Maxim's office in Sherman Oaks is seeking...

View full job description   Save to MyCareerBuilder   Email to a friend

Nurse Practitioner - Acute Care Nurse Practitioner - View similar jobs — Vanderbilt University Medical Center (VUMC)

Job type: Full-Time

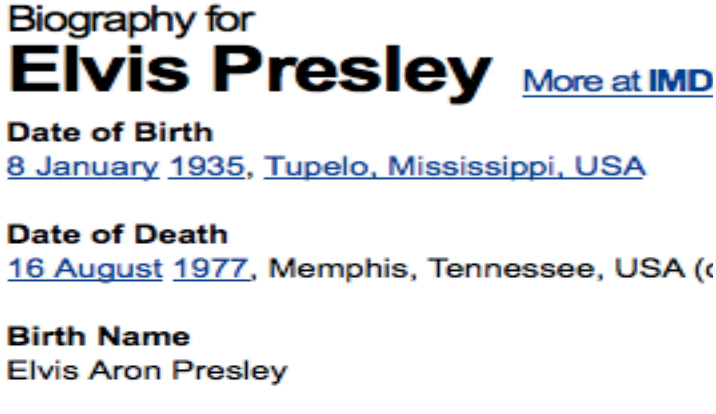
Vanderbilt University Medical Center is currently hiring **Nurse** Practitioners to join our team ... Vanderbilt University Medical Center is...

View full job description   Save to MyCareerBuilder   Email to a friend

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

Slide from Suchanek

Motivating Examples



Biography for

Elvis Presley More at IMD

Date of Birth

8 January 1935, Tupelo, Mississippi, USA

Date of Death

16 August 1977, Memphis, Tennessee, USA (c

Birth Name

Elvis Aron Presley

Nickname

The Pelvis
The King
The King Of Rock 'n

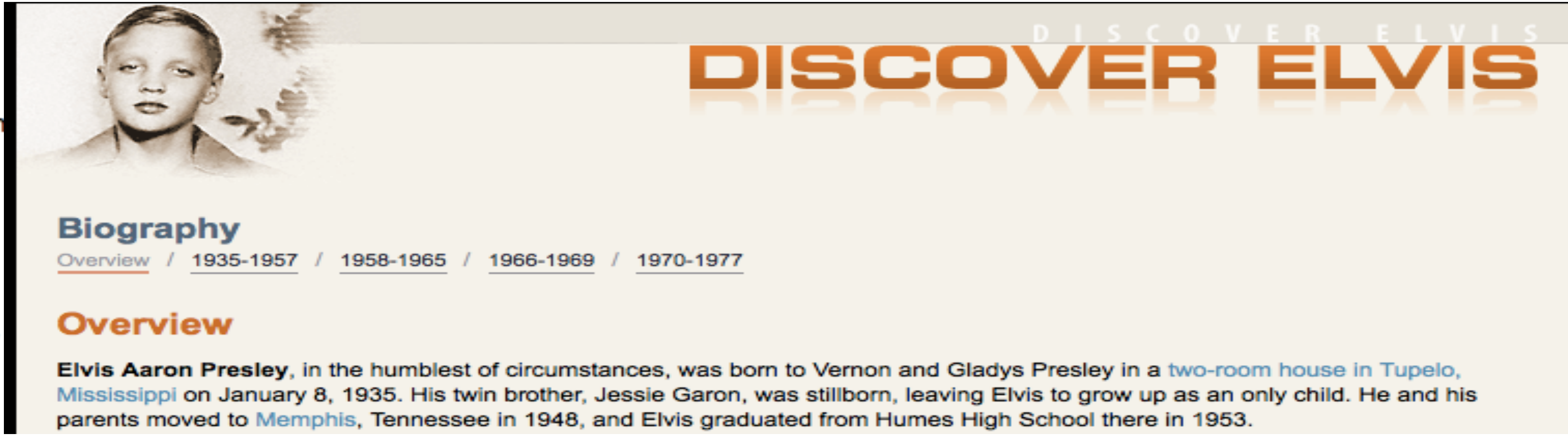
Height

6' (1.83 m)

Mini Biography

Elvis Aaron Presley

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



Biography

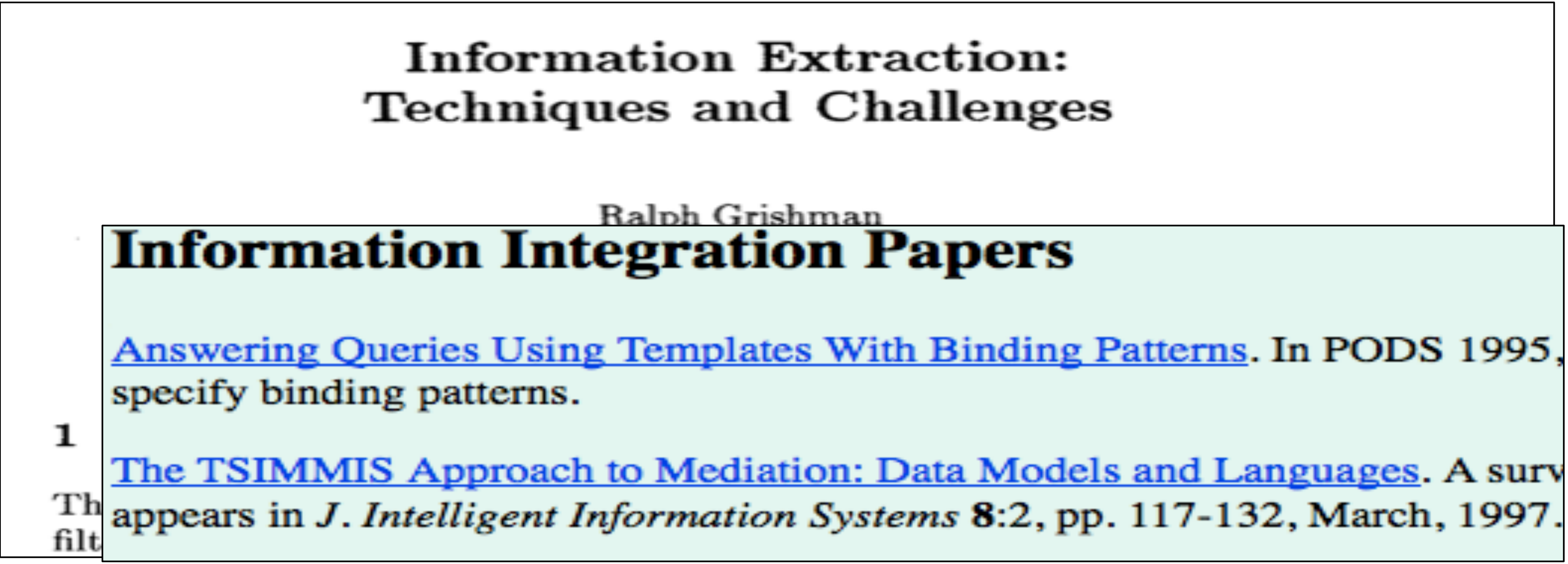
Overview / 1935-1957 / 1958-1965 / 1966-1969 / 1970-1977

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.

Slide from Suchanek

Motivating Examples



Information Extraction:
Techniques and Challenges

Ralph Grishman

1

Th

filt

Information Integration Papers

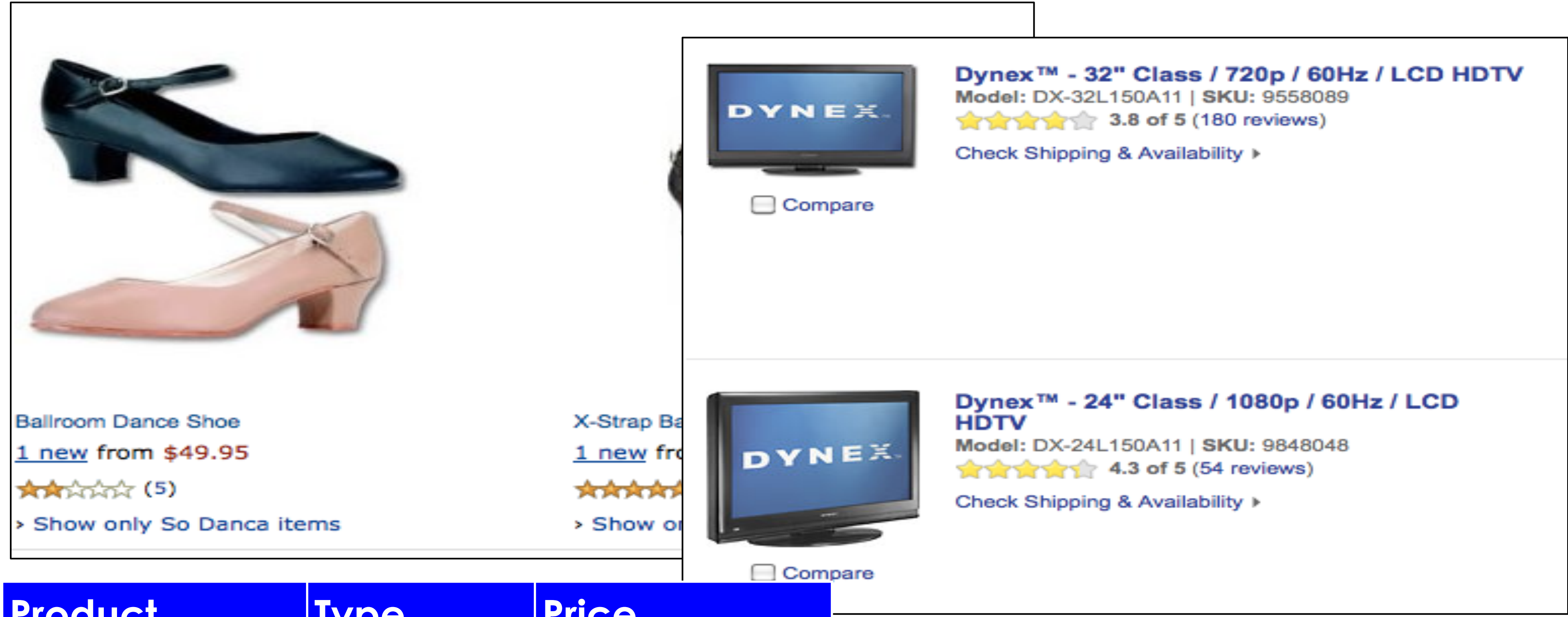
Answering Queries Using Templates With Binding Patterns. In PODS 1995,
specify binding patterns.

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

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

Slide from Suchanek

Motivating Examples



Ballroom Dance Shoe

1 new from $49.95

★★☆☆☆ (5)

› Show only So Danca items

X-Strap Ba

1 new fr

› Show o

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 ›

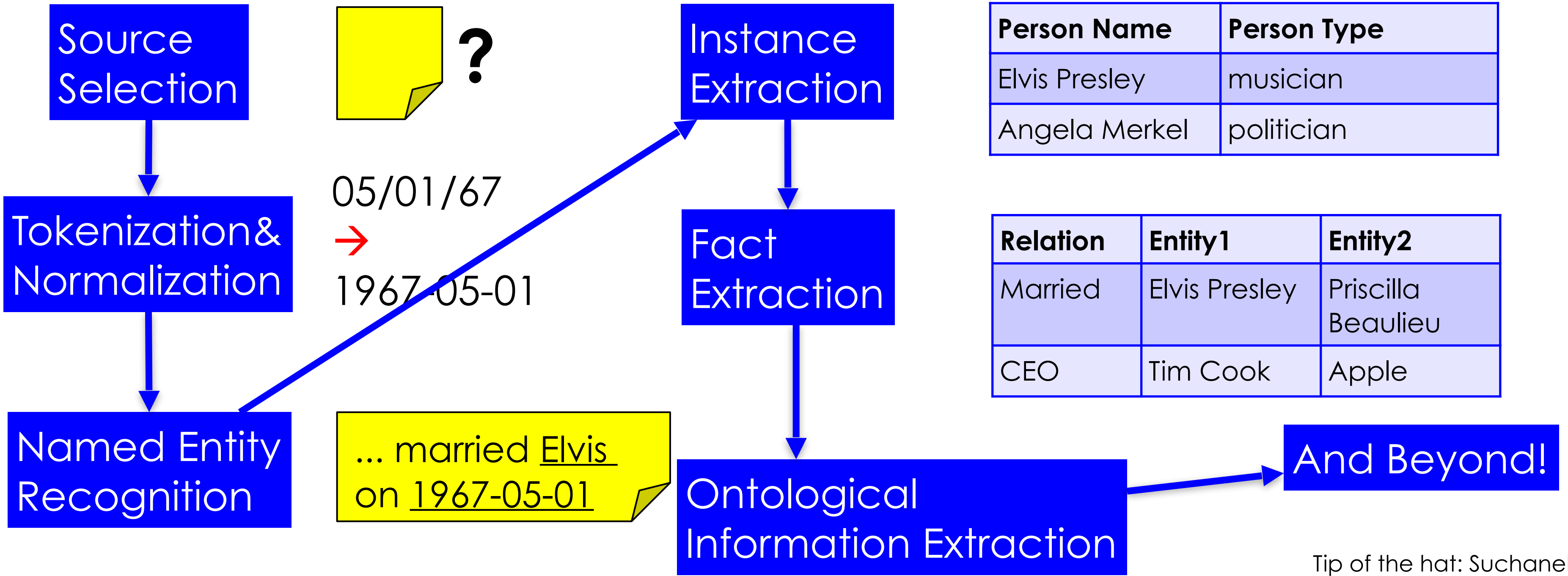
□ Compare

Product	Type	Price
Dynex 32"	LCD TV	$1000
…	…	

Slide from Suchanek

Information Extraction

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



Person Name	**Person Type**
Elvis Presley	musician
Angela Merkel	politician

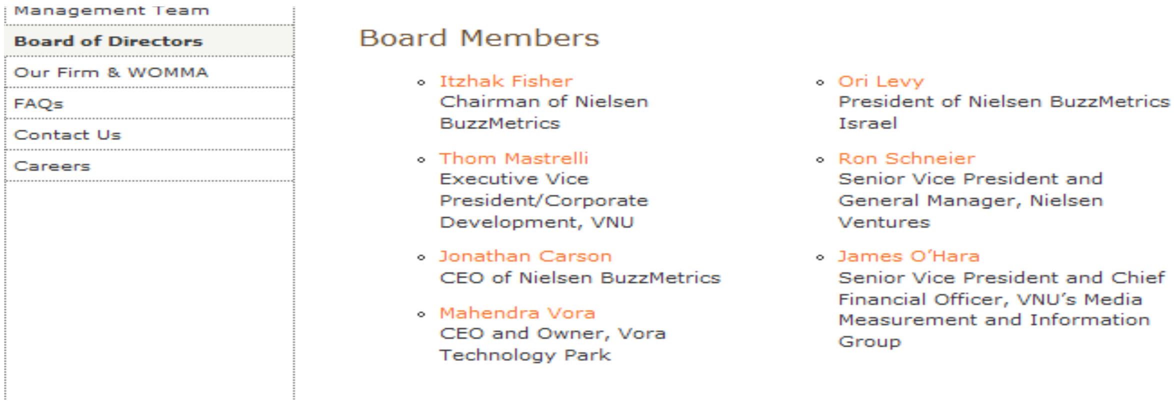
Relation	**Entity1**	**Entity2**
Married	Elvis Presley	Priscilla Beaulieu
CEO	Tim Cook	Apple

Tip of the hat: Suchanek

Information Extraction

Traditional definition: Recovering structured data from text

What are some of the sub-problems/challenges?



- Management Team
- **Board of Directors**
- Our Firm & WOMMA
- FAQs
- Contact Us
- Careers

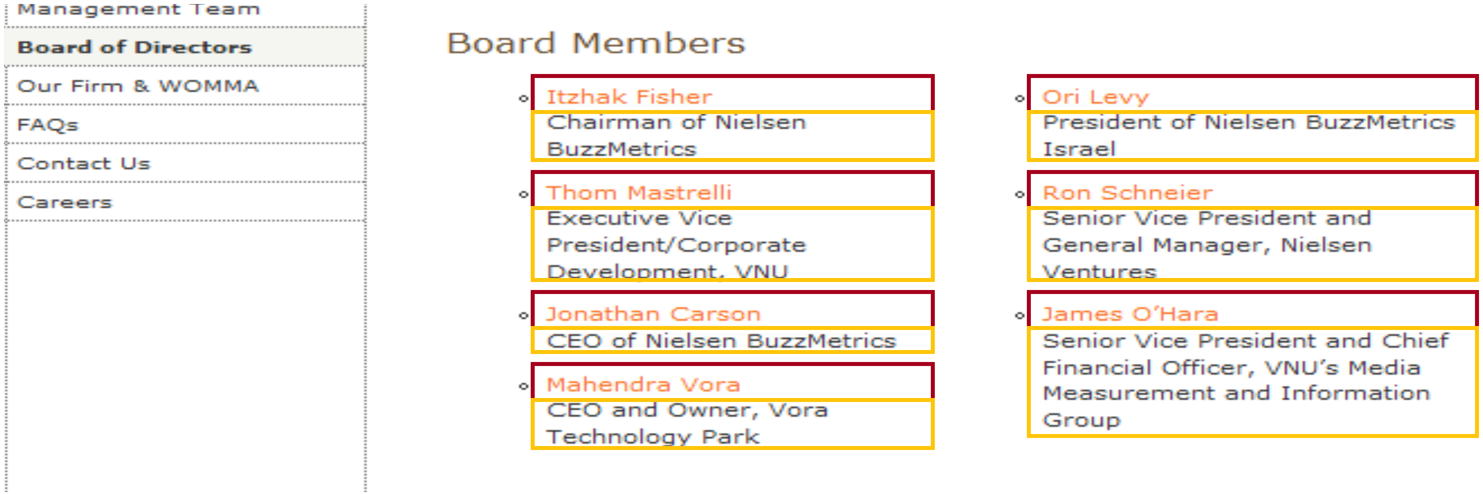
Board Members

- 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

Slide from Nigam/Cohen/McCallum

Information Extraction?

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



Management Team

Board of Directors

Our Firm & WOMMA

FAQs

Contact Us

Careers

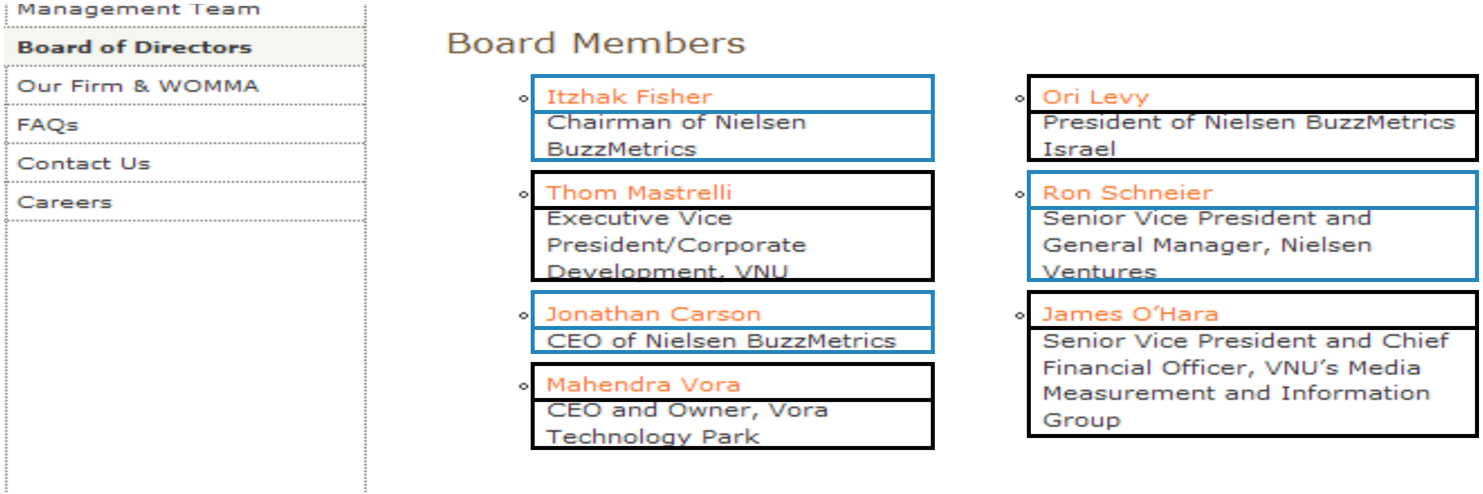
Board Members

- 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

Slide from Nigam/Cohen/McCallum

Information Extraction?

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



Management Team

Board of Directors

Our Firm & WOMMA

FAQs

Contact Us

Careers

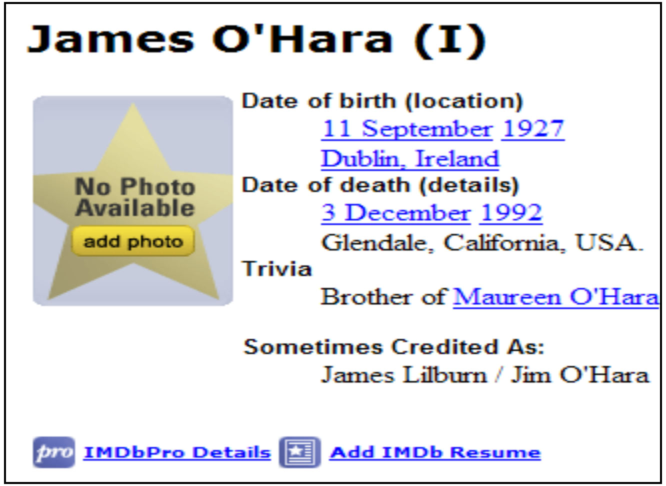
Board Members

- 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

Slide from Nigam/Cohen/McCallum

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)

No Photo Available

add photo

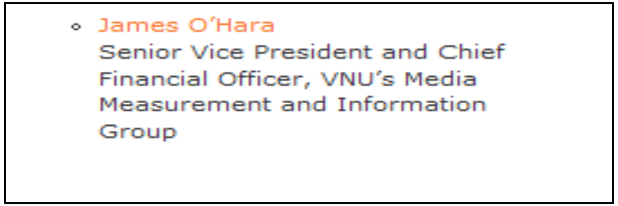
Date of birth (location)
11 September 1927
Dublin, Ireland

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

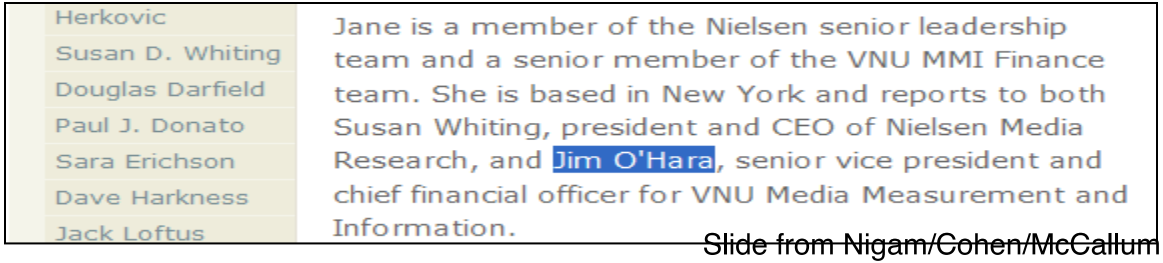
Trivia
Brother of Maureen O'Hara

Sometimes Credited As:
James Lilburn / Jim O'Hara

IMDbPro Details Add IMDb Resume



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



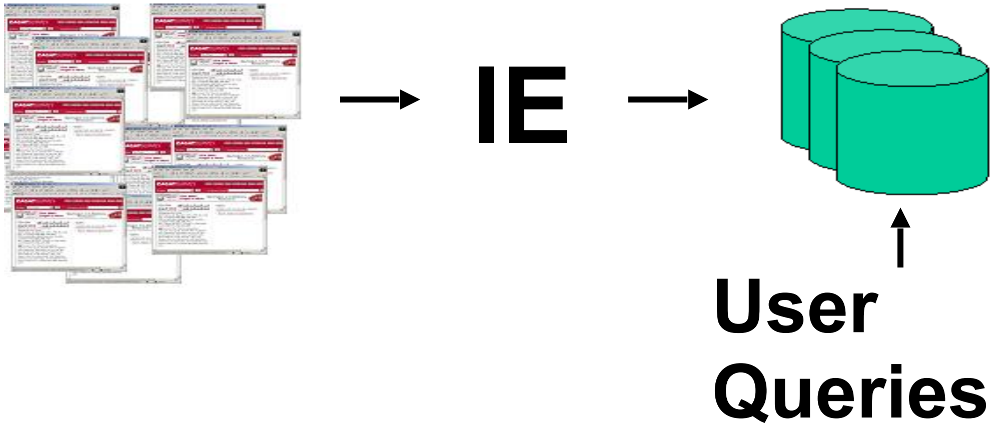
Herkovic
Susan D. Whiting
Douglas Darfield
Paul J. Donato
Sara Erichson
Dave Harkness
Jack Loftus

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

Slide from Nigam/Cohen/McCallum

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



IE

User Queries

Slide from McCallum

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

Slide from McCallum

Natural Text: MEDLINE Journal Abstracts

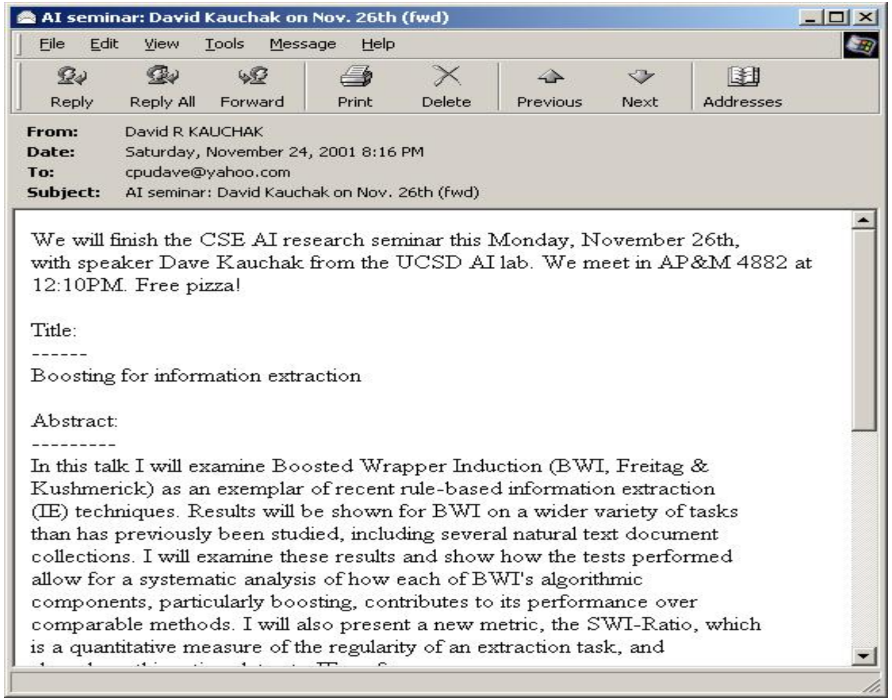
Extract number of subjects, type of study, conditions, etc.

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

Slide from Kauchak

Partially Structured: Seminar Announcements

Extract time, location, speaker, etc.



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

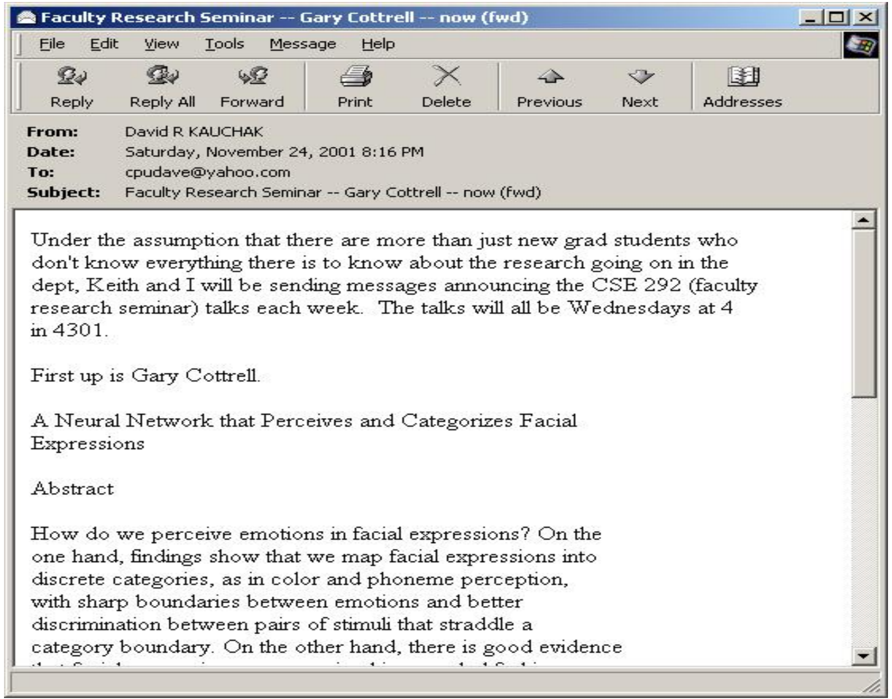
We will finish the CSE AI research seminar this Monday, November 26th, with speaker Dave Kauchak from the UCSD AI lab. We meet in AP&M 4882 at 12:10PM. Free pizza!

Title:

Boosting for information extraction

Abstract:

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



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

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

First up is Gary Cottrell.

A Neural Network that Perceives and Categorizes Facial Expressions

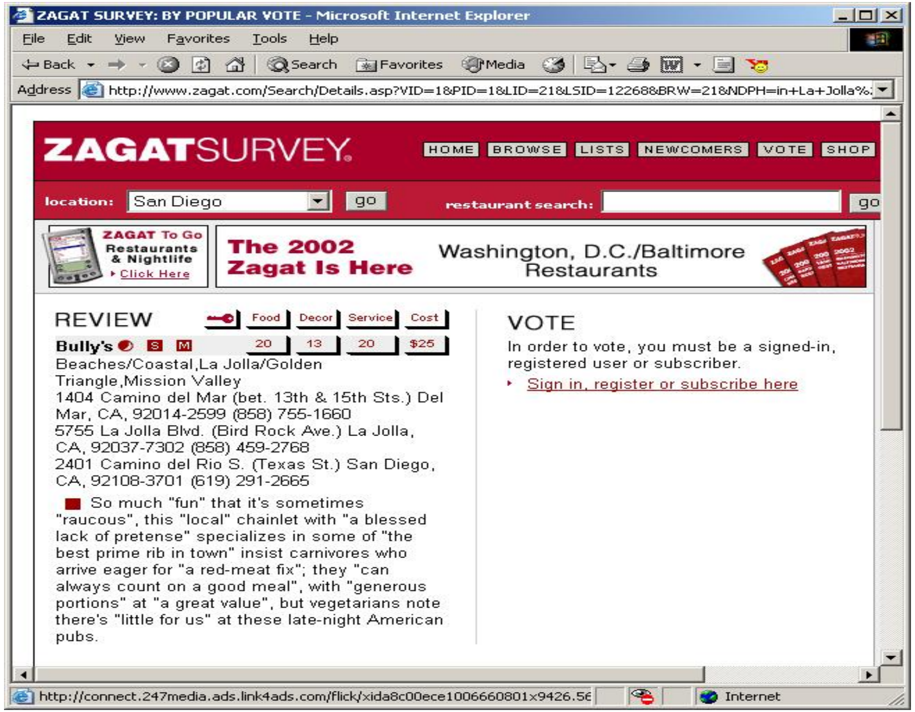
Abstract

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

Slide from Kauchak

Highly Structured:
Zagat's Reviews

Extract restaurant, location, cost, etc.


ZAGAT SURVEY: BY POPULAR VOTE - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
ZAGATSURVEY
HOME BROWSE LISTS NEWCOMERS VOTE SHOP
location:
go
restaurant search:
go
ZAGAT To Go Restaurants & Nightlife
Click Here
The 2002 Zagat Is Here
Washington, D.C./Baltimore Restaurants
Internet

http://www.zagat.com/Search/Details.asp?VID=1&PID=1&LID=21&LSID=12268&BRW=21&NDPH=in+La+Jolla%

location: San Diego

REVIEW

	Food	Decor	Service	Cost
Bully's ● S M	20	13	20	$25

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

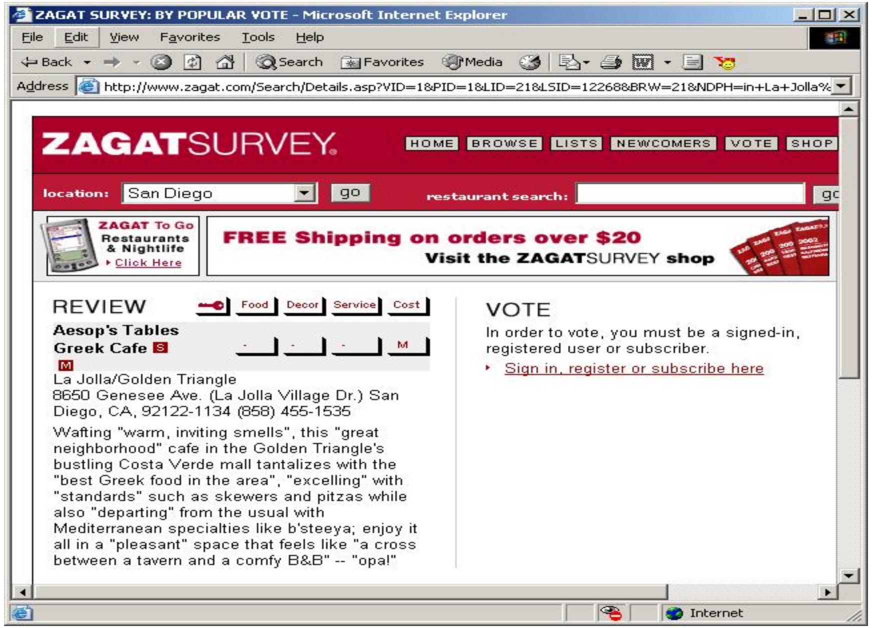
■ So much "fun" that it's sometimes "raucous", this "local" chainlet with "a blessed lack of pretense" specializes in some of "the best prime rib in town" insist carnivores who arrive eager for "a red-meat fix"; they "can always count on a good meal", with "generous portions" at "a great value", but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

VOTE

In order to vote, you must be a signed-in, registered user or subscriber.

- Sign in, register or subscribe here

http://connect.247media.ads.link4ads.com/flick/xida8c00ece1006660801x9426.56


ZAGAT SURVEY: BY POPULAR VOTE - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
ZAGATSURVEY
HOME BROWSE LISTS NEWCOMERS VOTE SHOP
location:
go
restaurant search:
go
ZAGAT To Go Restaurants & Nightlife
Click Here
FREE Shipping on orders over $20
Visit the ZAGATSURVEY shop
Internet

http://www.zagat.com/Search/Details.asp?VID=1&PID=1&LID=21&LSID=12268&BRW=21&NDPH=in+La+Jolla%

location: San Diego

REVIEW

	Food	Decor	Service	Cost
Aesop's Tables Greek Cafe S M	-	-	-	M

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

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

VOTE

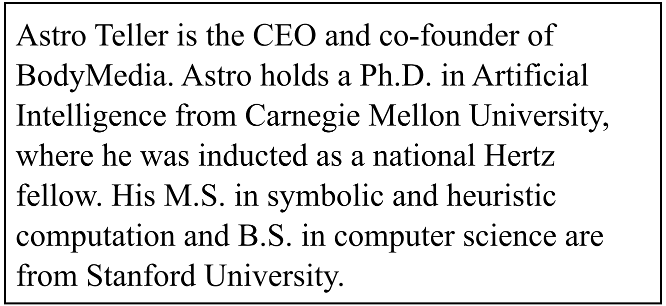
In order to vote, you must be a signed-in, registered user or subscriber.

- Sign in, register or subscribe here

Slide from Kauchak

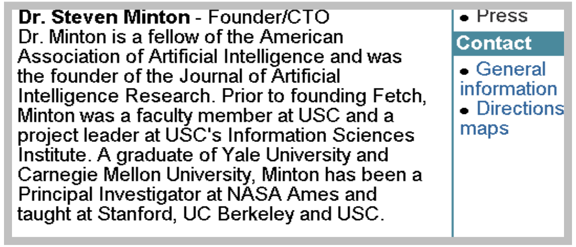
Landscape of IE Tasks:
Document Formatting

Text paragraphs without formatting



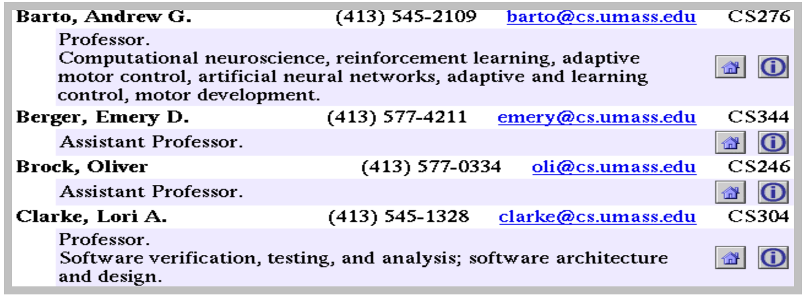
Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University.

Grammatical sentences and some formatting & links



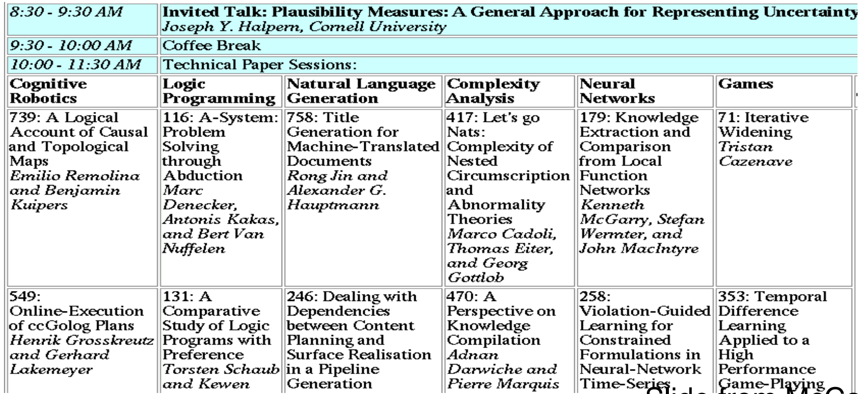
Dr. Steven Minton - Founder/CTO
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC's Information Sciences Institute. 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.

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.			⌂ ⓘ

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 and Gerhard Lakemeyer*	131: A Comparative Study of Logic Programs with Preference *Torsten Schaub and Kewen*	246: Dealing with Dependencies between Content Planning and Surface Realisation in a Pipeline Generation	470: A Perspective on Knowledge Compilation *Adnan Darwiche and Pierre Marquis*	258: Violation-Guided Learning for Constrained Formulations in Neural-Network Time-Series	353: Temporal Difference Learning Applied to a High Performance Game-Playing

Slide from McCallum

Landscape of IE Tasks
Intended Breadth of Coverage

Web site specific

Formatting

Amazon.com Book Pages

Genre specific

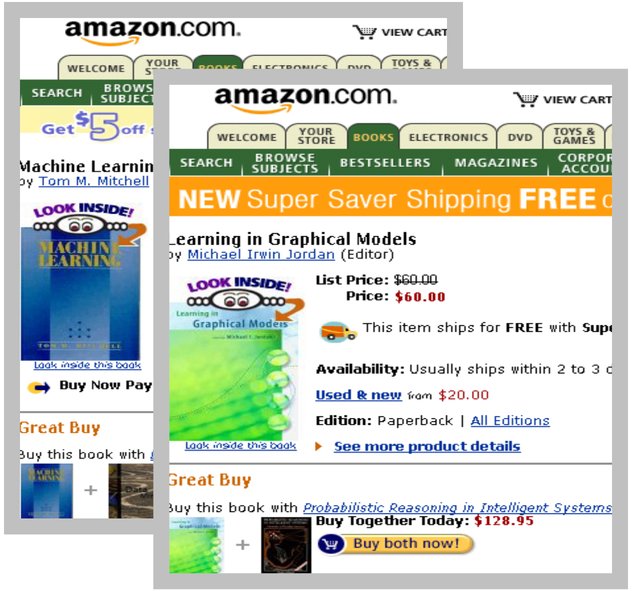
Layout

Resumes

Wide, non-specific

Language

University Names



Machine Learnin
by Tom M. Mitchell

Buy Now Pay

Great Buy

Buy this book with

NEW Super Saver Shipping FREE o

Learning in Graphical Models
by Michael Irwin Jordan (Editor)

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

This item ships for FREE with Supe

Availability: Usually ships within 2 to 3 d

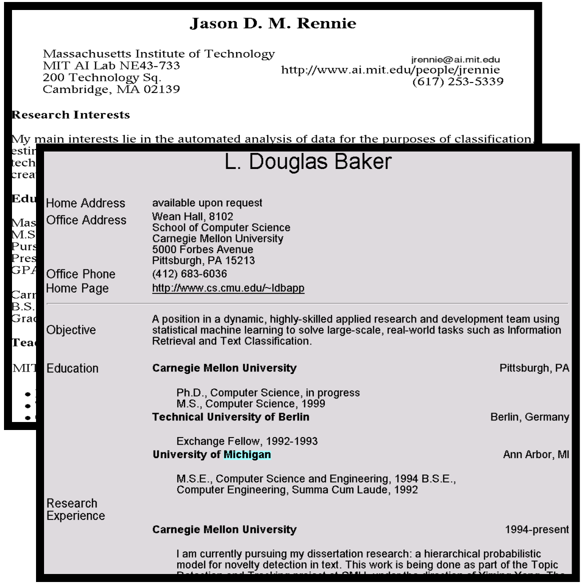
Used & new from $20.00

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,

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

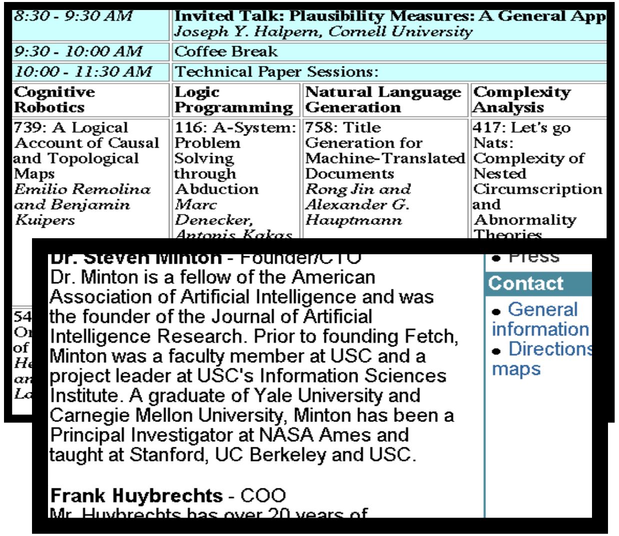
University of Michigan Ann Arbor, MI

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 App** *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*	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

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

Slide from McCallum

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, AR

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.

Slide from McCallum

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 Welsh**
In: **Jeffrey Immelt**

"Named entity" extraction

Slide from McCallum

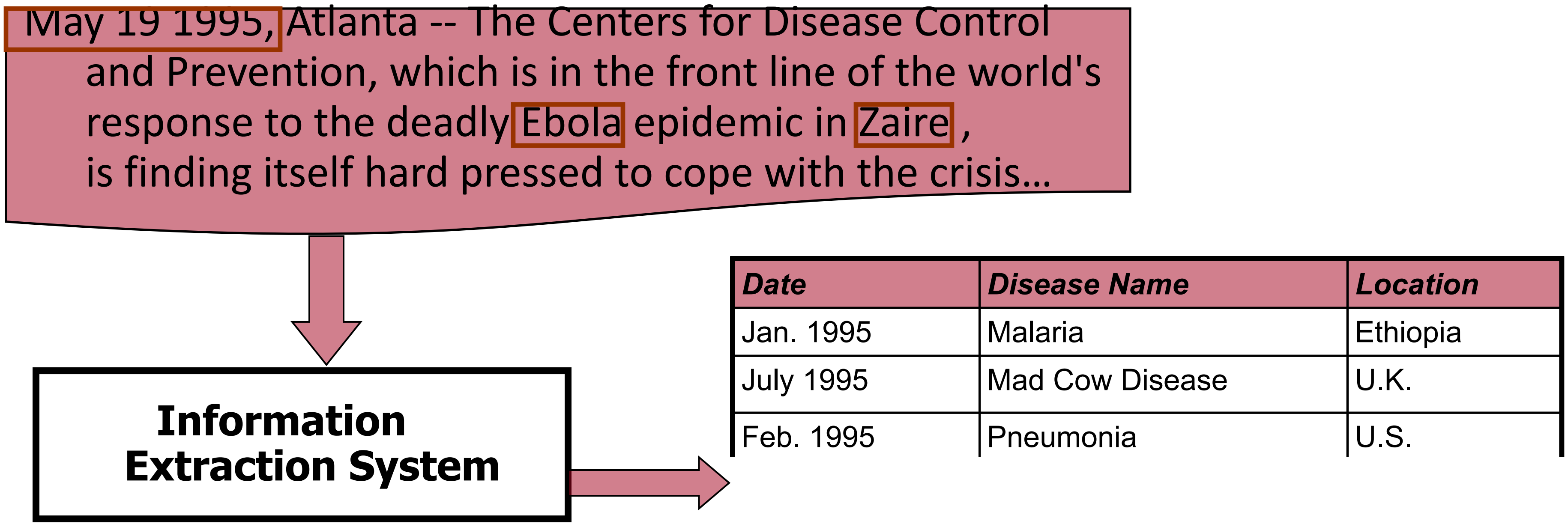
Association task = Relation Extraction

- 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:
        - “((Subj<company>) (Verb located) (*) (Obj <location>))”
2. Machine Learning-based
    - Supervised: Learn relation classifier from examples
    - Partially-supervised: bootstrap rules/patterns from “seed” examples

Slide modified from Manning

Relation Extraction: Disease Outbreaks



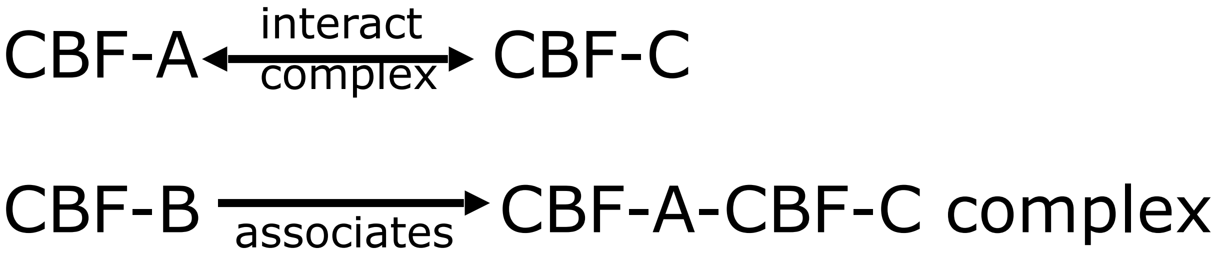
May 19 1995, Atlanta -- The Centers for Disease Control and Prevention, which is in the front line of the world's response to the deadly Ebola epidemic in Zaire , is finding itself hard pressed to cope with the crisis…

Date	***Disease Name***	***Location***
Jan. 1995	Malaria	Ethiopia
July 1995	Mad Cow Disease	U.K.
Feb. 1995	Pneumonia	U.S.

Slide from Manning

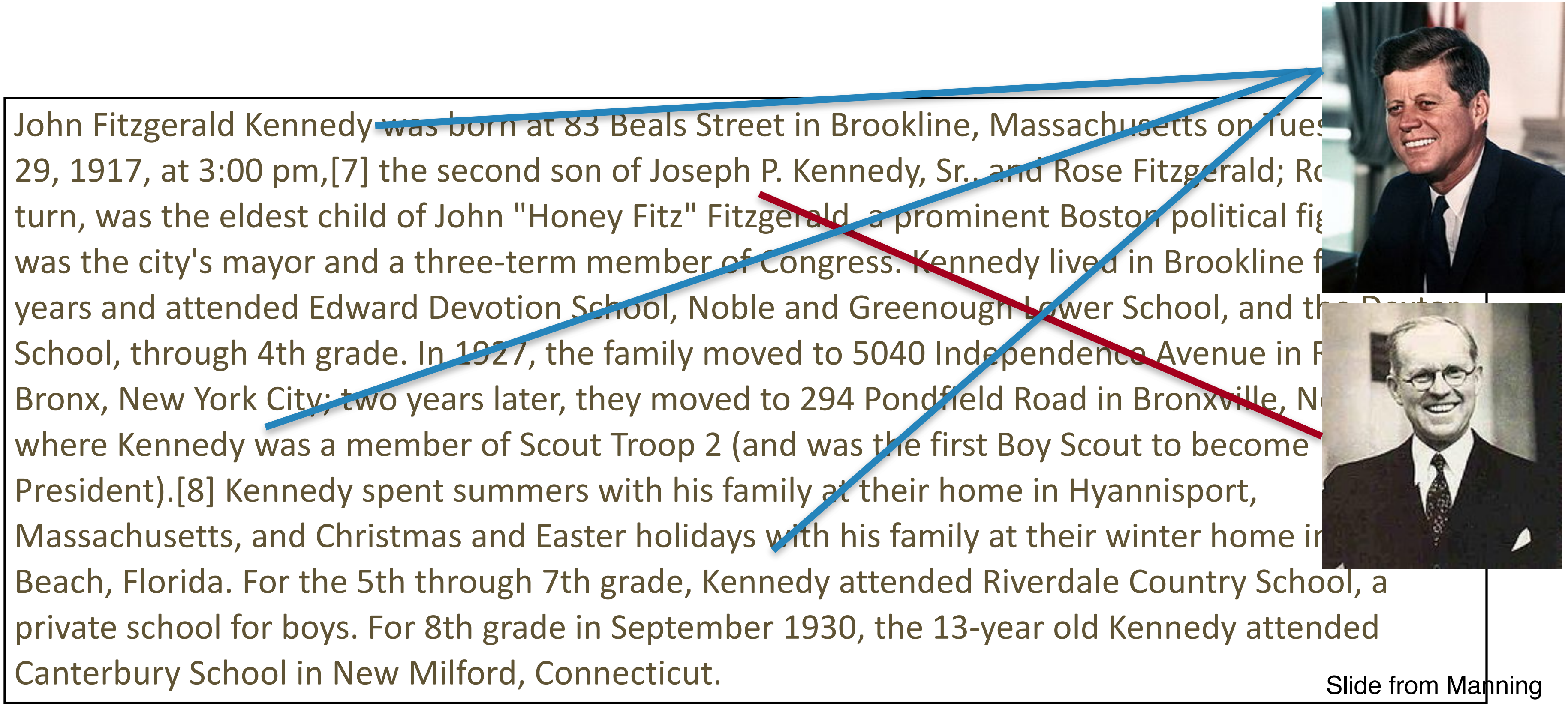
Relation Extraction: Protein Interactions

“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.“



Slide from Manning

Resolving coreference
(both within and across documents)



John Fitzgerald Kennedy was born at 83 Beals Street in Brookline, Massachusetts on Tues
29, 1917, at 3:00 pm,[7] the second son of Joseph P. Kennedy, Sr., and Rose Fitzgerald; Ro
turn, was the eldest child of John "Honey Fitz" Fitzgerald, a prominent Boston political fig
was the city's mayor and a three-term member of Congress. Kennedy lived in Brookline f
years and attended Edward Devotion School, Noble and Greenough Lower School, and the Dexter
School, through 4th grade. In 1927, the family moved to 5040 Independence Avenue in R
Bronx, New York City; two years later, they moved to 294 Pondfield Road in Bronxville, Ne
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.

Slide from Manning

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

Slide from Manning

What we will cover in this class (briefly)

- PART I: basic information extraction (through Named Entity Recognition)
  - History of IE, Related Fields
  - Source Selection
  - Tokenization and Normalization
  - Named Entity Recognition (NER)

What we will cover in this class (briefly)

- PART II: machine learning in depth (mostly tagging models used for named entities)
  - Decision Trees and Overfitting
  - Linear Models
  - Feature Engineering
  - Word Embeddings
  - Deep Learning (Non-Linear Models)

What we will cover in this class (briefly)

- PART III: advanced information extraction
  - Instance Extraction
  - Fact/Event Extraction
  - Ontological IE/Open IE

Last words

- The seminar tomorrow is cancelled
- The topics will be presented in early November
- Also, don't forget the reading for next week!
- **Sarawagi: Information Extraction** (available from web page)
  Read the introduction!
- These slides will be uploaded as well
  - The video of this lecture from WS 2020-2021 is available, note that the "Administravia" is different

- Thank you for your attention!