Managing Information Extraction
SIGMOD 2006 Tutorial

AnHai Doan

UIUC → UW-Madison

Raghu Ramakrishnan

UW-Madison → Yahoo! Research

Shiv Vaithyanathan

IBM Almaden

Tutorial Roadmap

- Introduction to managing IE [RR]
  - Motivation
  - What’s different about *managing* IE?
- Major research directions
  - Extracting mentions of entities and relationships [SV]
    - Uncertainty management
  - Disambiguating extracted mentions [AD]
    - Tracking mentions and entities over time
  - Understanding, correcting, and maintaining extracted data [AD]
    - Provenance and explanations
    - Incorporating user feedback

The Presenters

AnHai Doan

- Currently at Illinois
- Starts at UW-Madison in July
- Has worked extensively in semantic integration, data integration, at the intersection of databases, Web, and AI
- Leads the Cimple project and builds DBLife in collaboration with Raghu Ramakrishnan and a terrific team of students
- Search for “anhai” on the Web

Raghu Ramakrishnan

- Research Fellow at Yahoo! Research, where he moved from UW-Madison after finding out that AnHai was moving there
- Has worked on data mining and database systems, and is currently focused on Web data management and online communities
- Collaborates with AnHai and gang on the Cimple/DBlife project, and with Shiv on aspects of Avatar
- See www.cs.wisc.edu/~raghu

Shiv Vaithyanathan

- Shiv Vaithyanathan manages the Unstructured Information Mining group at IBM Almaden where he moved after stints in DEC and Altavista.
- Shiv leads the Avatar project at IBM and is considering moving out of California now that Raghu has moved in.

- See www.almaden.ibm.com/software/projects/avatar/

Introduction

Lots of Text, Many Applications!

- Free-text, semi-structured, streaming …
  - Web pages, email, news articles, call-center text records, business reports, annotations, spreadsheets, research papers, blogs, tags, instant messages (IM), …
- High-impact applications
  - Business intelligence, personal information management, Web communities, Web search and advertising, scientific data management, e-government, medical records management, …
- Growing rapidly
  - Your email inbox!

Exploiting Text ➔ Important Direction for Our Community

- Many other research communities are looking at how to exploit text
  - Most actively, Web, IR, AI, KDD
- Important direction for us as well!
  - We have lot to offer, and a lot to gain
- How is text exploited?

Two main directions: IR and IE

Exploiting Text via IR (Information Retrieval)

- Keyword search over data containing text (relational, XML)
  - What should the query language be? Ranking criteria?
  - How do we evaluate queries?
- Integrating IR systems with DB systems
  - Architecture?
  - See SIGMOD-04 panel; Baeza-Yates / Consens tutorial [SIGIR 05]

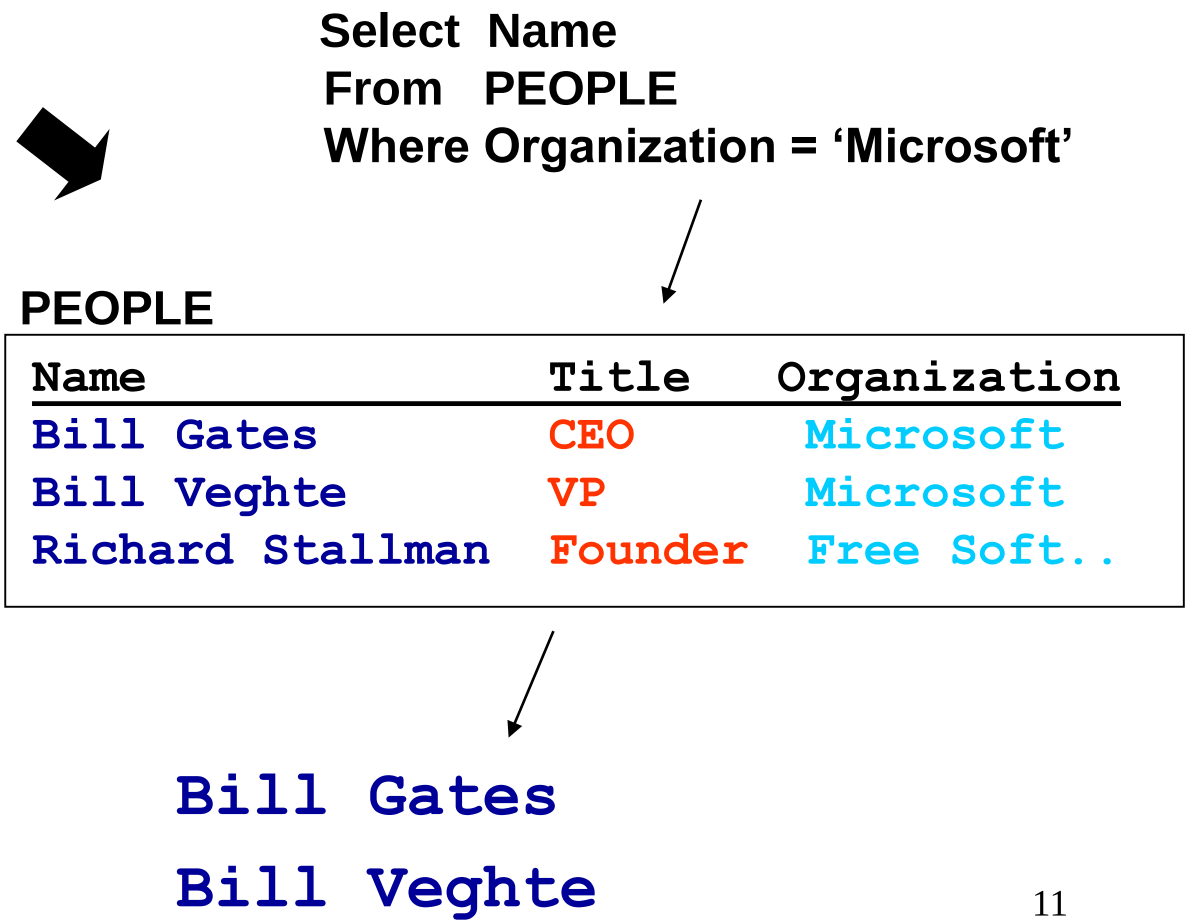
Not the focus of our tutorial

Exploiting Text via IE (Information Extraction)

- Extract, then exploit, structured data from raw text:

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…



Select Name
From PEOPLE
Where Organization = ‘Microsoft’

PEOPLE

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

Bill Gates

Bill Veghte

(from Cohen’s IE tutorial, 2003)

This Tutorial: Research at the Intersection of IE and DB Systems

- We can apply DB approaches to
  - Analyzing and using extracted information in the context of other related data, as well as
  - The process of extracting and maintaining structured data from text
- A “killer app” for database systems?
  - Lots of text, but until now, mostly outside DBMSs
  - Extracted information could make the difference!

Let’s use three concrete applications to illustrate what we can do with IE …

A Disclaimer

This tutorial touches upon a lot of areas, some with much prior work. Rather than attempt a comprehensive survey, we've tried to identify areas for further research by the DB community.

We've therefore drawn freely from our own experiences in creating specific examples and articulating problems.

We are creating an annotated bibliography site, and we hope you'll join us in maintaining it at

http://scratchpad.wikia.com/wiki/Dblife_bibs

Application 1: Enterprise Search

T.S. Jayram

Rajasekar Krishnamurthy

Sriram Raghavan

Huaiyu Zhu

Avatar Semantic Search @ IBM Almaden

http://www.almaden.ibm.com/software/projects/avatar/

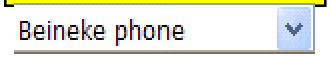
(and Shiv Vaithyanathan)

(SIGMOD Demo, 2006)

Overview of Avatar Semantic Search

Incorporate higher-level semantics into information retrieval to ascertain user-intent

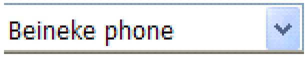
Conventional Search



Beineke phone

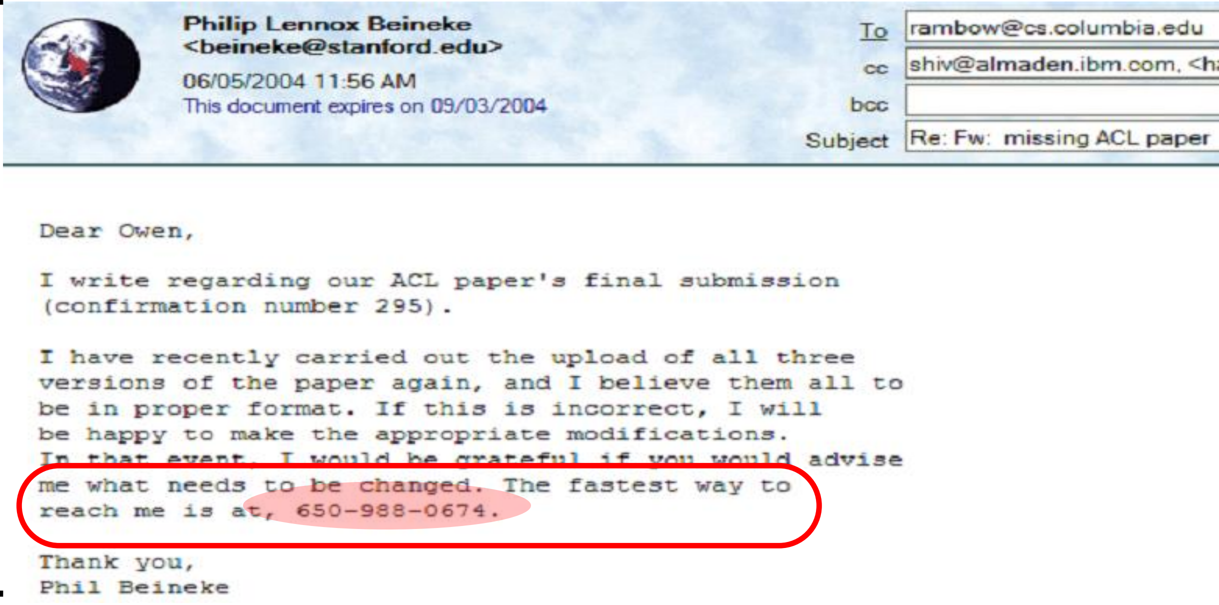
Interpreted as

Return emails that contain the keywords "Beineke" and phone



Beineke phone

It will miss



Philip Lennox Beineke
<beineke@stanford.edu>

06/05/2004 11:56 AM
This document expires on 09/03/2004

To: rambow@cs.columbia.edu
cc: shiv@almaden.ibm.com, <h
bcc:
Subject: Re: Fw: missing ACL paper

Dear Owen,

I write regarding our ACL paper's final submission (confirmation number 295).

I have recently carried out the upload of all three versions of the paper again, and I believe them all to be in proper format. If this is incorrect, I will be happy to make the appropriate modifications. In that event, I would be grateful if you would advise me what needs to be changed. The fastest way to reach me is at, 650-988-0674.

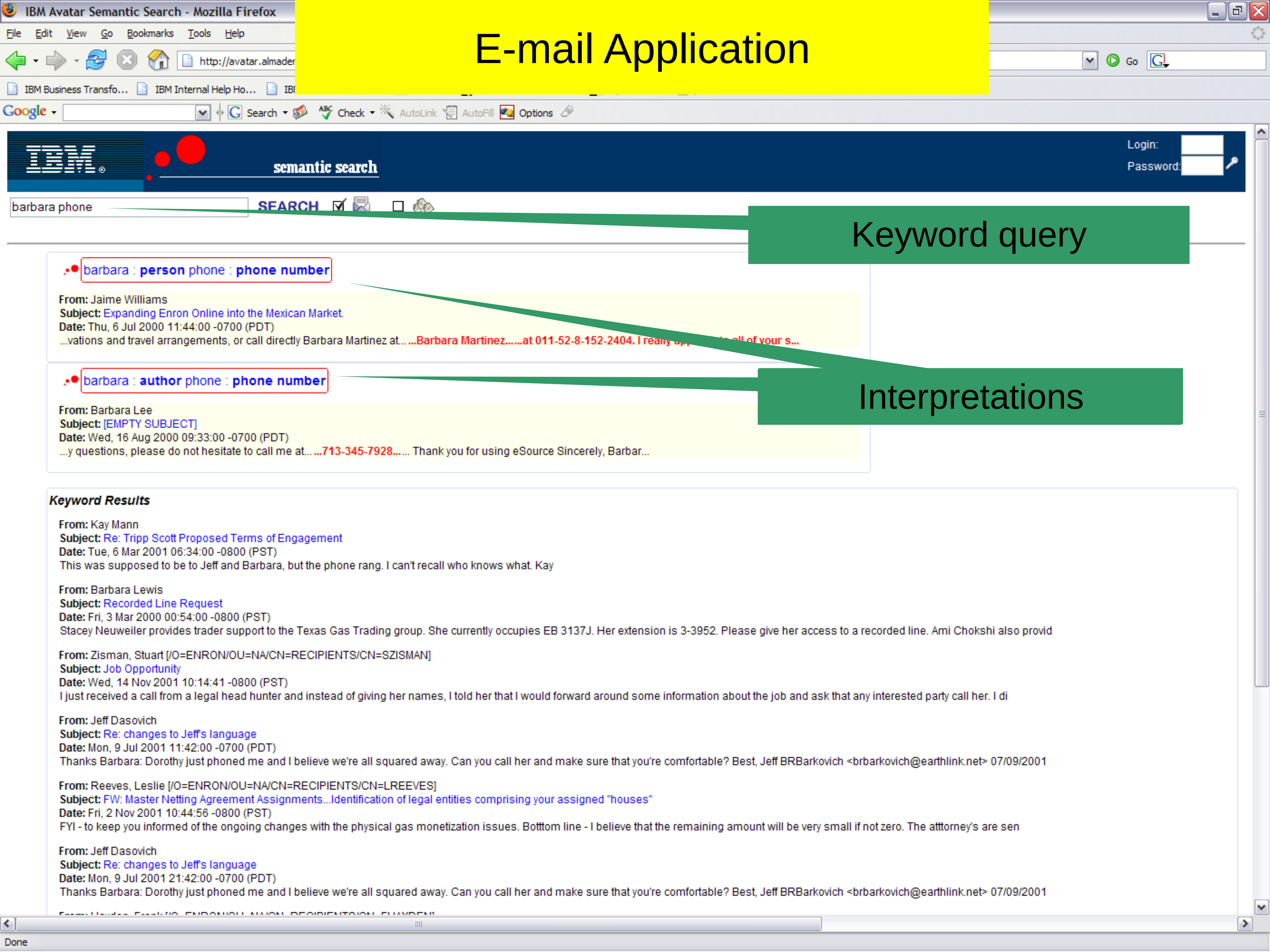
Thank you,
Phil Beineke

Avatar Semantic Search engages the user in a simple dialogue to ascertain user need

True user intent can be any of …

Query 1: return emails FROM Beineke that contain his contact telephone number
Query 2: return emails that contain Beineke's signature
Query 3: return emails FROM Beineke that contain a telephone number
More ………….

E-mail Application



barbara phone

Keyword query

Interpretations

barbara : **person** phone : **phone number**

From: Jaime Williams
Subject: Expanding Enron Online into the Mexican Market.
Date: Thu, 6 Jul 2000 11:44:00 -0700 (PDT)
...vations and travel arrangements, or call directly Barbara Martinez at... **...Barbara Martinez......at 011-52-8-152-2404. I really [illegible] all of your s...**

barbara : **author** phone : **phone number**

From: Barbara Lee
Subject: [EMPTY SUBJECT]
Date: Wed, 16 Aug 2000 09:33:00 -0700 (PDT)
...y questions, please do not hesitate to call me at... **...713-345-7928.....** Thank you for using eSource Sincerely, Barbar...

Keyword Results

From: Kay Mann
Subject: Re: Tripp Scott Proposed Terms of Engagement
Date: Tue, 6 Mar 2001 06:34:00 -0800 (PST)
This was supposed to be to Jeff and Barbara, but the phone rang. I can't recall who knows what. Kay

From: Barbara Lewis
Subject: Recorded Line Request
Date: Fri, 3 Mar 2000 00:54:00 -0800 (PST)
Stacey Neuweiler provides trader support to the Texas Gas Trading group. She currently occupies EB 3137J. Her extension is 3-3952. Please give her access to a recorded line. Ami Chokshi also provid

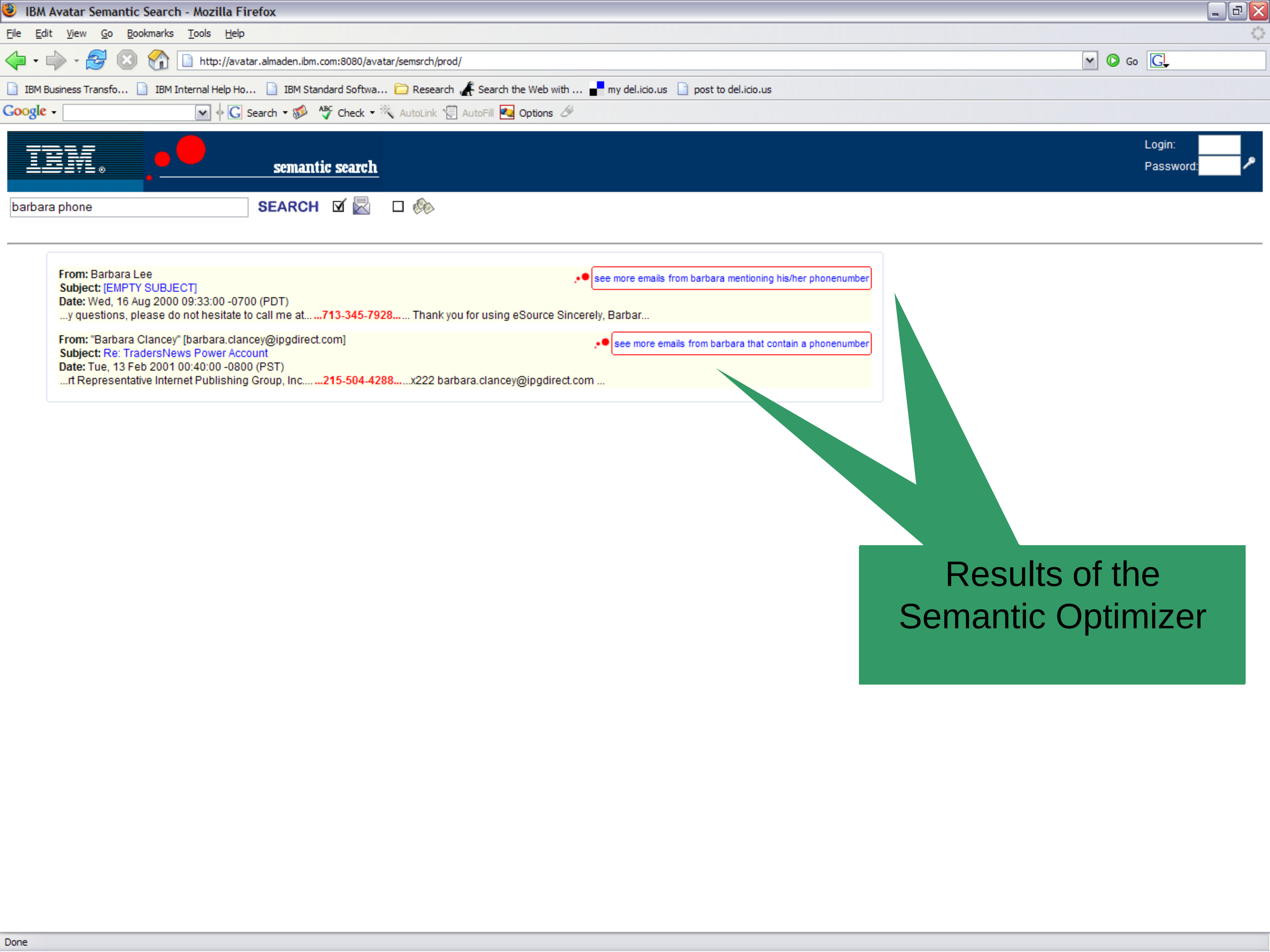
From: Zisman, Stuart [/O=ENRON/OU=NA/CN=RECIPIENTS/CN=SZISMAN]
Subject: Job Opportunity
Date: Wed, 14 Nov 2001 10:14:41 -0800 (PST)
I just received a call from a legal head hunter and instead of giving her names, I told her that I would forward around some information about the job and ask that any interested party call her. I di

From: Jeff Dasovich
Subject: Re: changes to Jeff's language
Date: Mon, 9 Jul 2001 11:42:00 -0700 (PDT)
Thanks Barbara: Dorothy just phoned me and I believe we're all squared away. Can you call her and make sure that you're comfortable? Best, Jeff BRBarkovich <brbarkovich@earthlink.net> 07/09/2001

From: Reeves, Leslie [/O=ENRON/OU=NA/CN=RECIPIENTS/CN=LREEVES]
Subject: FW: Master Netting Agreement Assignments...Identification of legal entities comprising your assigned "houses"
Date: Fri, 2 Nov 2001 10:44:56 -0800 (PST)
FYI - to keep you informed of the ongoing changes with the physical gas monetization issues. Botttom line - I believe that the remaining amount will be very small if not zero. The atttorney's are sen

From: Jeff Dasovich
Subject: Re: changes to Jeff's language
Date: Mon, 9 Jul 2001 21:42:00 -0700 (PDT)
Thanks Barbara: Dorothy just phoned me and I believe we're all squared away. Can you call her and make sure that you're comfortable? Best, Jeff BRBarkovich <brbarkovich@earthlink.net> 07/09/2001

From: Hayden, Frank [/O=ENRON/OU=NA/CN=RECIPIENTS/CN=FHAYDEN]


IBM Avatar Semantic Search - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://avatar.almaden.ibm.com:8080/avatar/semsrch/prod/
Go
IBM Business Transfo... IBM Internal Help Ho... IBM Standard Softwa... Research Search the Web with ... my del.icio.us post to del.icio.us
Google Search Check AutoLink AutoFill Options
IBM
semantic search
Login:
Password:
SEARCH
Done

barbara phone

SEARCH ☑ □

From: Barbara Lee
Subject: [EMPTY SUBJECT]
Date: Wed, 16 Aug 2000 09:33:00 -0700 (PDT)
...y questions, please do not hesitate to call me at... ...713-345-7928...... Thank you for using eSource Sincerely, Barbar...

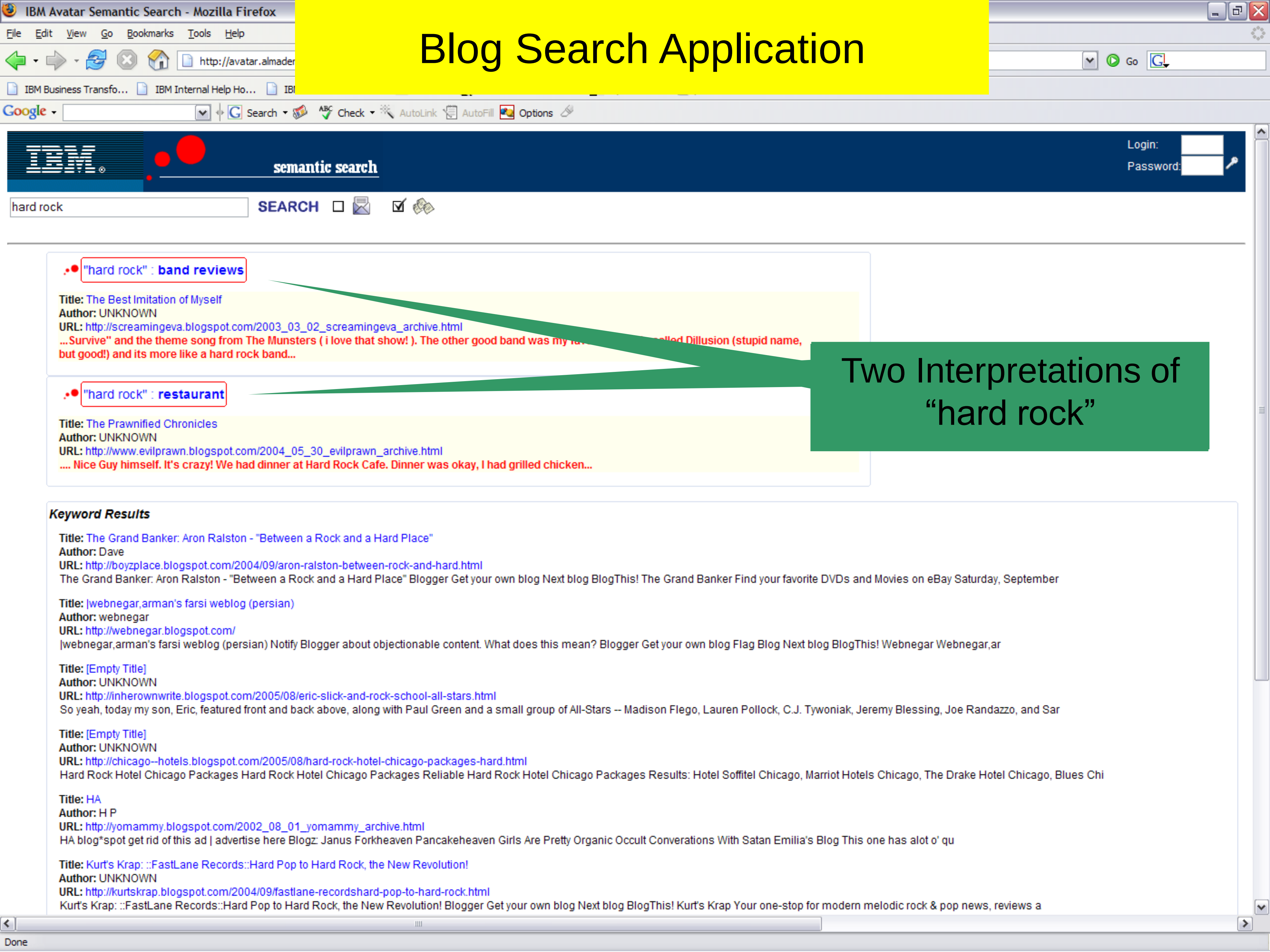
see more emails from barbara mentioning his/her phonenumber

From: "Barbara Clancey" [barbara.clancey@ipgdirect.com]
Subject: Re: TradersNews Power Account
Date: Tue, 13 Feb 2001 00:40:00 -0800 (PST)
...rt Representative Internet Publishing Group, Inc.... ...215-504-4288......x222 barbara.clancey@ipgdirect.com ...

see more emails from barbara that contain a phonenumber

Results of the Semantic Optimizer

Blog Search Application



hard rock

"hard rock" : **band reviews**

Title: The Best Imitation of Myself
Author: UNKNOWN
URL: http://screamingeva.blogspot.com/2003_03_02_screamingeva_archive.html
...Survive" and the theme song from The Munsters (i love that show!). The other good band was my fa alled Dillusion (stupid name, but good!) and its more like a hard rock band...

"hard rock" : **restaurant**

Title: The Prawnified Chronicles
Author: UNKNOWN
URL: http://www.evilprawn.blogspot.com/2004_05_30_evilprawn_archive.html
.... Nice Guy himself. It's crazy! We had dinner at Hard Rock Cafe. Dinner was okay, I had grilled chicken...

Keyword Results

Title: The Grand Banker: Aron Ralston - "Between a Rock and a Hard Place"
Author: Dave
URL: http://boyzplace.blogspot.com/2004/09/aron-ralston-between-rock-and-hard.html
The Grand Banker: Aron Ralston - "Between a Rock and a Hard Place" Blogger Get your own blog Next blog BlogThis! The Grand Banker Find your favorite DVDs and Movies on eBay Saturday, September

Title: |webnegar,arman's farsi weblog (persian)
Author: webnegar
URL: http://webnegar.blogspot.com/
|webnegar,arman's farsi weblog (persian) Notify Blogger about objectionable content. What does this mean? Blogger Get your own blog Flag Blog Next blog BlogThis! Webnegar Webnegar,ar

Title: [Empty Title]
Author: UNKNOWN
URL: http://inherownwrite.blogspot.com/2005/08/eric-slick-and-rock-school-all-stars.html
So yeah, today my son, Eric, featured front and back above, along with Paul Green and a small group of All-Stars -- Madison Flego, Lauren Pollock, C.J. Tywoniak, Jeremy Blessing, Joe Randazzo, and Sar

Title: [Empty Title]
Author: UNKNOWN
URL: http://chicago--hotels.blogspot.com/2005/08/hard-rock-hotel-chicago-packages-hard.html
Hard Rock Hotel Chicago Packages Hard Rock Hotel Chicago Packages Reliable Hard Rock Hotel Chicago Packages Results: Hotel Soffitel Chicago, Marriot Hotels Chicago, The Drake Hotel Chicago, Blues Chi

Title: HA
Author: H P
URL: http://yomammy.blogspot.com/2002_08_01_yomammy_archive.html
HA blog*spot get rid of this ad | advertise here Blogz: Janus Forkheaven Pancakeheaven Girls Are Pretty Organic Occult Converations With Satan Emilia's Blog This one has alot o' qu

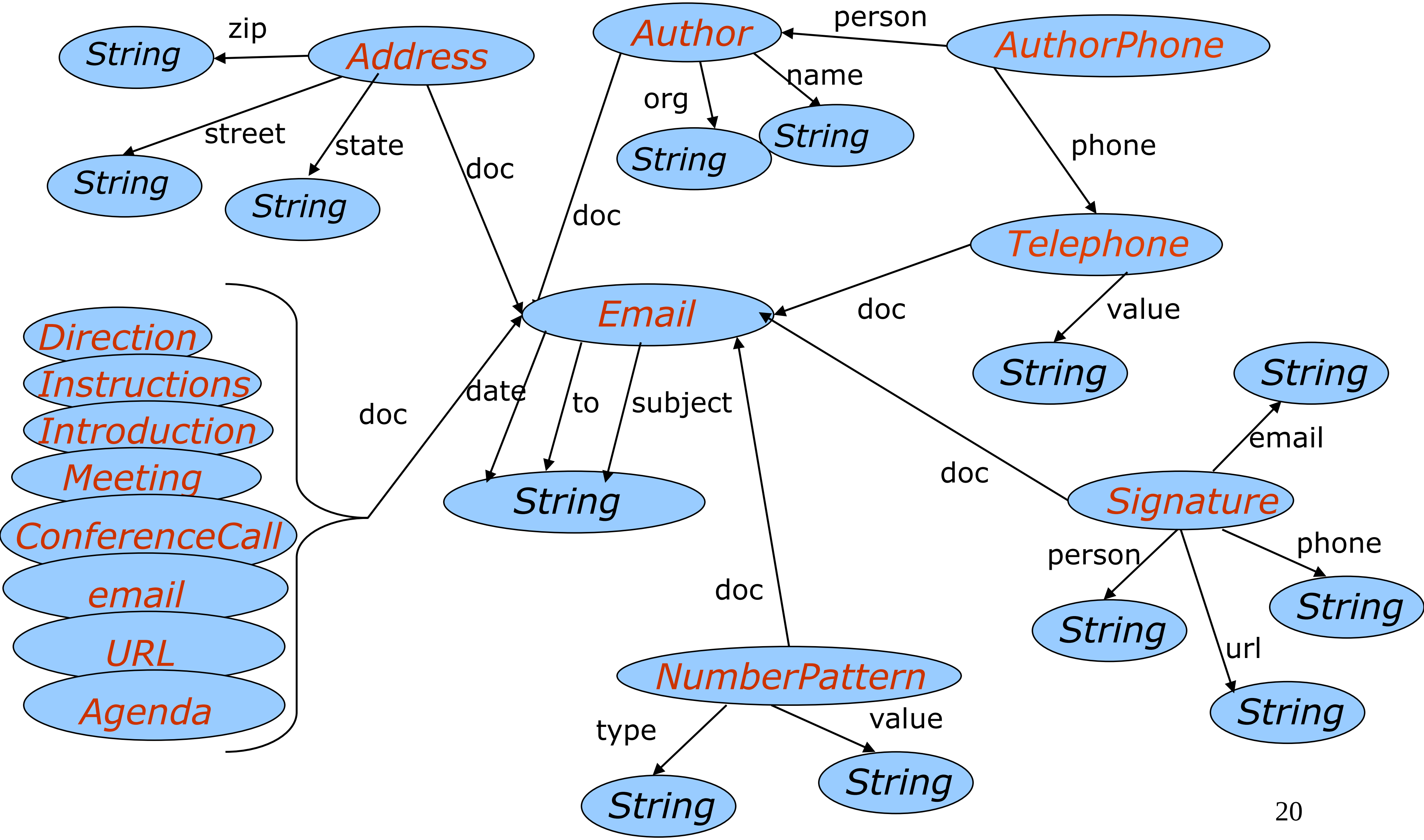
Title: Kurt's Krap: ::FastLane Records::Hard Pop to Hard Rock, the New Revolution!
Author: UNKNOWN
URL: http://kurtskrap.blogspot.com/2004/09/fastlane-recordshard-pop-to-hard-rock.html
Kurt's Krap: ::FastLane Records::Hard Pop to Hard Rock, the New Revolution! Blogger Get your own blog Next blog BlogThis! Kurt's Krap Your one-stop for modern melodic rock & pop news, reviews a

Two Interpretations of "hard rock"

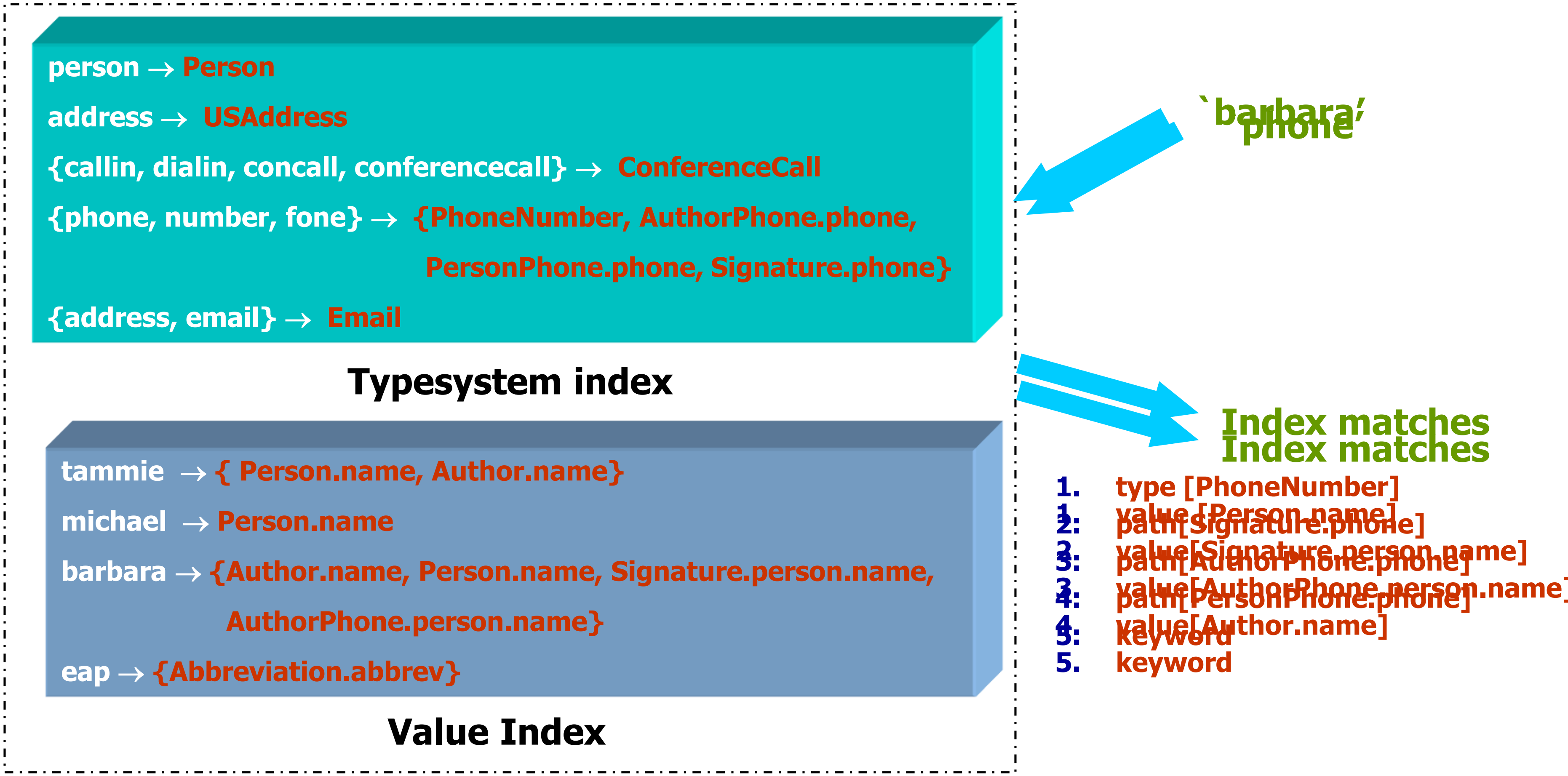
How Semantic Search Works

- Semantic Search is basically KIDO (Keywords In Documents Out) enhanced by text-analytics
- During offline processing, information extraction algorithms are used to extract specific facts from the raw text
- At runtime, a “semantic optimizer” disambiguates the keyword query in the context of the extracted information and selects the best interpretations to present to the user

Partial Type-System for Email



Translation Index



person → Person

address → USAddress

{callin, dialin, concall, conferencecall} → ConferenceCall

{phone, number, fone} → {PhoneNumber, AuthorPhone.phone, PersonPhone.phone, Signature.phone}

{address, email} → Email

Typesystem index

tammie → { Person.name, Author.name}

michael → Person.name

barbara → {Author.name, Person.name, Signature.person.name, AuthorPhone.person.name}

eap → {Abbreviation.abbrev}

Value Index

`barbara'

phone

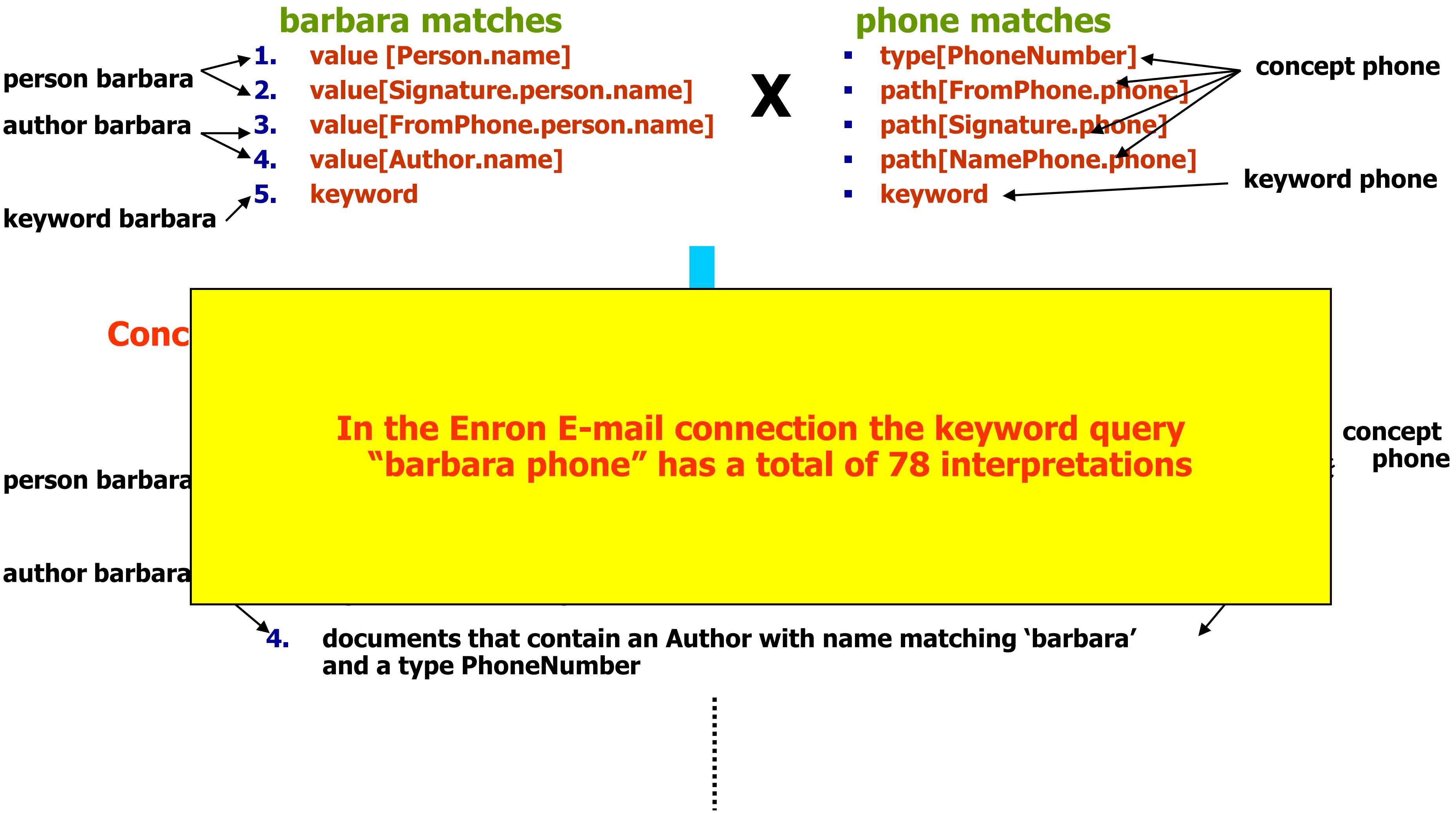
Index matches

1. type [PhoneNumber]
2. path[Signature.phone]
3. path[AuthorPhone.phone]
4. path[PersonPhone.phone]
5. keyword

Index matches

1. value [Person.name]
2. value[Signature.person.name]
3. value[AuthorPhone.person.name]
4. value[Author.name]
5. keyword

Concept tagged matches



barbara matches

person barbara

author barbara

keyword barbara

1. value [Person.name]
2. value[Signature.person.name]
3. value[FromPhone.person.name]
4. value[Author.name]
5. keyword

X

phone matches

- type[PhoneNumber]
- path[FromPhone.phone]
- path[Signature.phone]
- path[NamePhone.phone]
- keyword

concept phone

keyword phone

Conc

person barbara

author barbara

concept phone

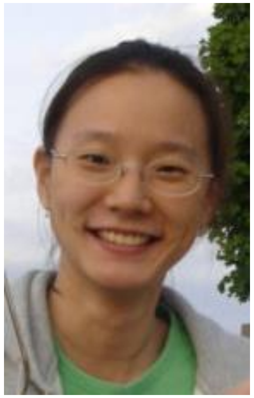
In the Enron E-mail connection the keyword query "barbara phone" has a total of 78 interpretations

4. documents that contain an Author with name matching 'barbara' and a type PhoneNumber

Application 2: Community Information Management (CIM)

Fei Chen

Pedro DeRose



Yoonkyong Lee

Warren Shen

The DBLife System @ Illinois / Wisconsin

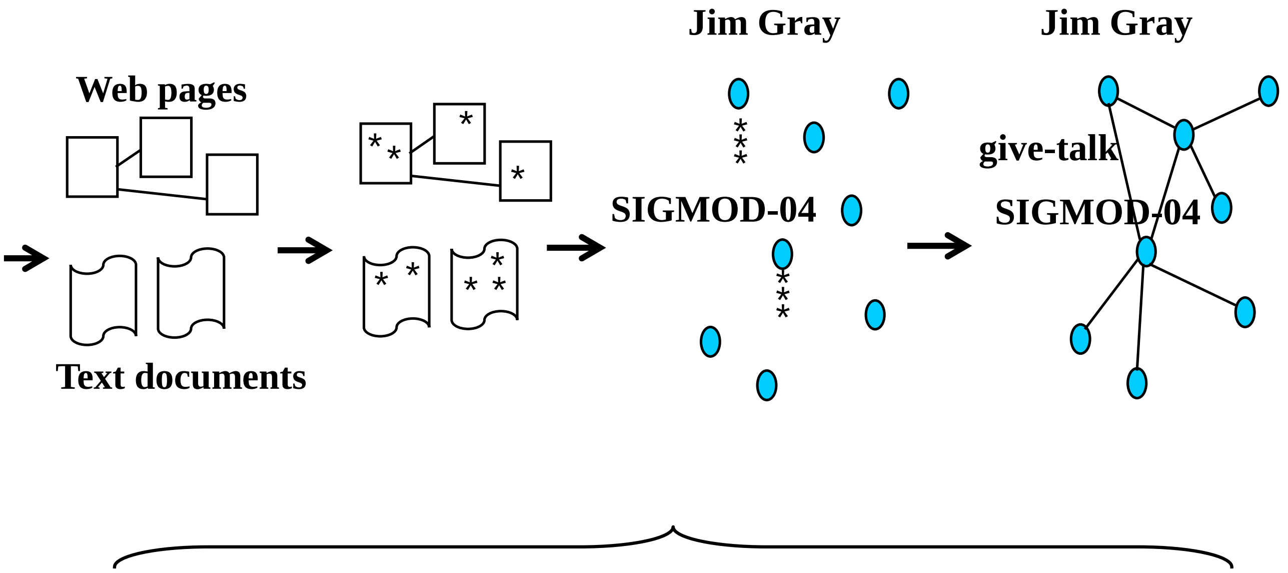
(and AnHai Doan, Raghu Ramakrishnan)

Best-Effort, Collaborative Data Integration for Web Communities

- There are many data-rich communities
  - Database researchers, movie fans, bioinformatics
  - Enterprise intranets, tech support groups
- Each community = many disparate data sources + many people
- By integrating relevant data, we can enable search, monitoring, and information discovery:
  - Any interesting connection between researchers X and Y?
  - Find all citations of this paper in the past one week on the Web
  - What is new in the past 24 hours in the database community?
  - Which faculty candidates are interviewing this year, where?
  - What are current hot topics? Who has moved where?

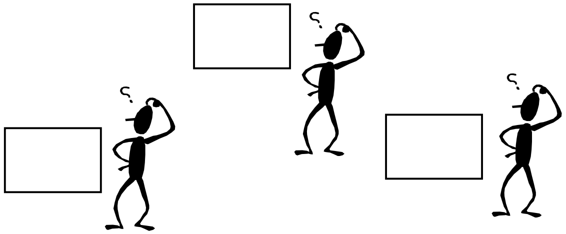
Cimple Project @ Illinois/Wisconsin

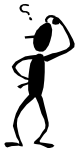
- Researcher Homepages
- Conference Pages
- Group pages
- DBworld mailing list
- DBLP



Import & personalize data

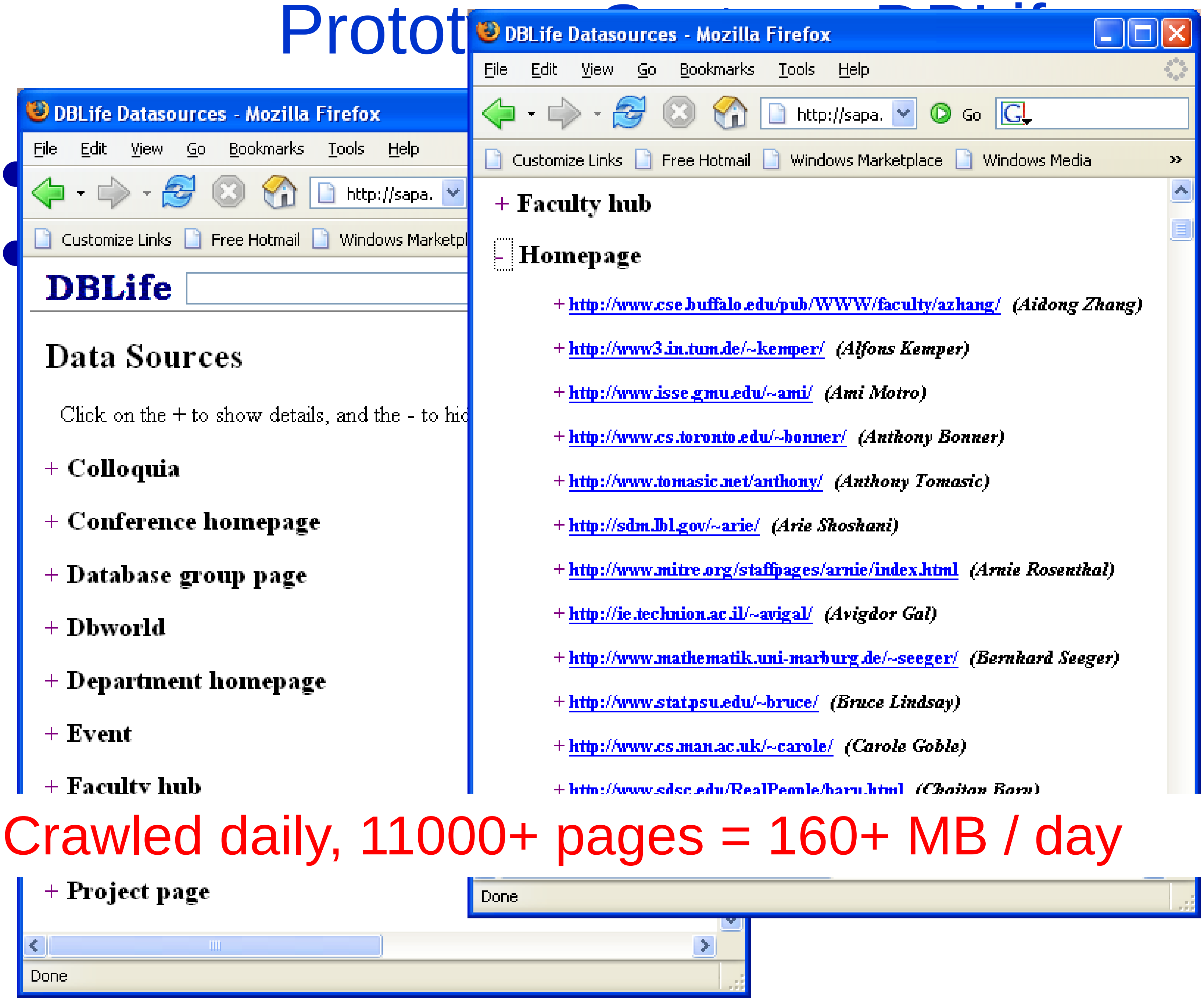
Modify data, provide feedback





- Keyword search
- SQL querying
- Question answering
- Browse
- Mining
- Alerts, tracking
- News summary

Protot



DBLife Datasources - Mozilla Firefox

DBLife

Data Sources

Click on the + to show details, and the - to hid

- + Colloquia
- + Conference homepage
- + Database group page
- + Dbworld
- + Department homepage
- + Event
- + Faculty hub
- + Project page

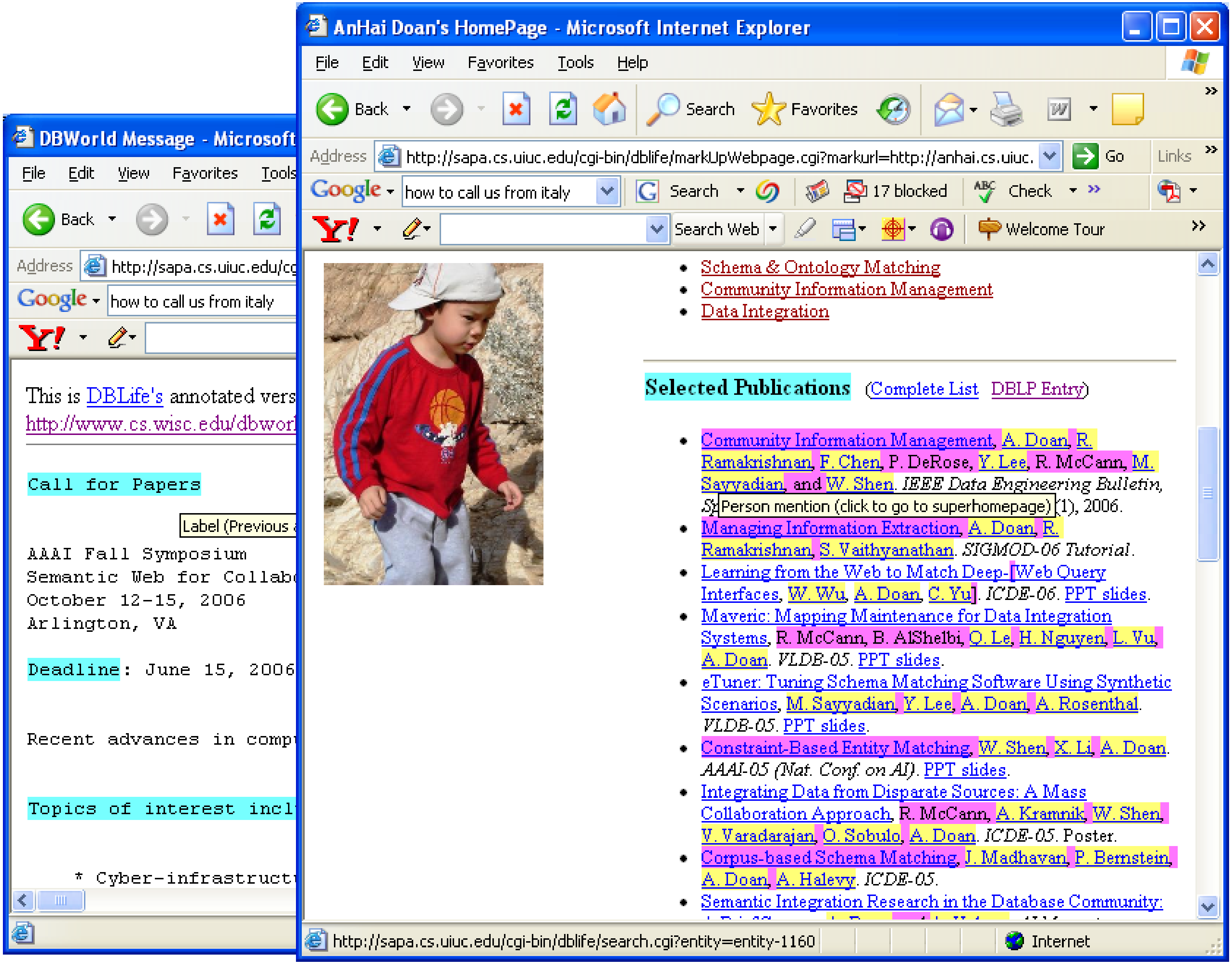
DBLife Datasources - Mozilla Firefox

+ Faculty hub

- Homepage

- + http://www.cse.buffalo.edu/pub/WWW/faculty/azhang/ *(Aidong Zhang)*
- + http://www3.in.tum.de/~kemper/ *(Alfons Kemper)*
- + http://www.isse.gmu.edu/~ami/ *(Ami Motro)*
- + http://www.cs.toronto.edu/~bonner/ *(Anthony Bonner)*
- + http://www.tomasic.net/anthony/ *(Anthony Tomasic)*
- + http://sdm.lbl.gov/~arie/ *(Arie Shoshani)*
- + http://www.mitre.org/staffpages/arnie/index.html *(Arnie Rosenthal)*
- + http://ie.technion.ac.il/~avigal/ *(Avigdor Gal)*
- + http://www.mathematik.uni-marburg.de/~seeger/ *(Bernhard Seeger)*
- + http://www.stat.psu.edu/~bruce/ *(Bruce Lindsay)*
- + http://www.cs.man.ac.uk/~carole/ *(Carole Goble)*
- + http://www.sdsc.edu/RealPeople/baru.html *(Chaitan Baru)*

Crawled daily, 11000+ pages = 160+ MB / day



http://sapa.cs.uiuc.edu/cg

This is DBLife's annotated vers
http://www.cs.wisc.edu/dbworl

```
Call for Papers

Label (Previous a

AAAI Fall Symposium
Semantic Web for Collabo
October 12-15, 2006
Arlington, VA

Deadline: June 15, 2006


Recent advances in comp


Topics of interest incl


    * Cyber-infrastructu
```

http://sapa.cs.uiuc.edu/cgi-bin/dblife/markUpWebpage.cgi?markurl=http://anhai.cs.uiuc.

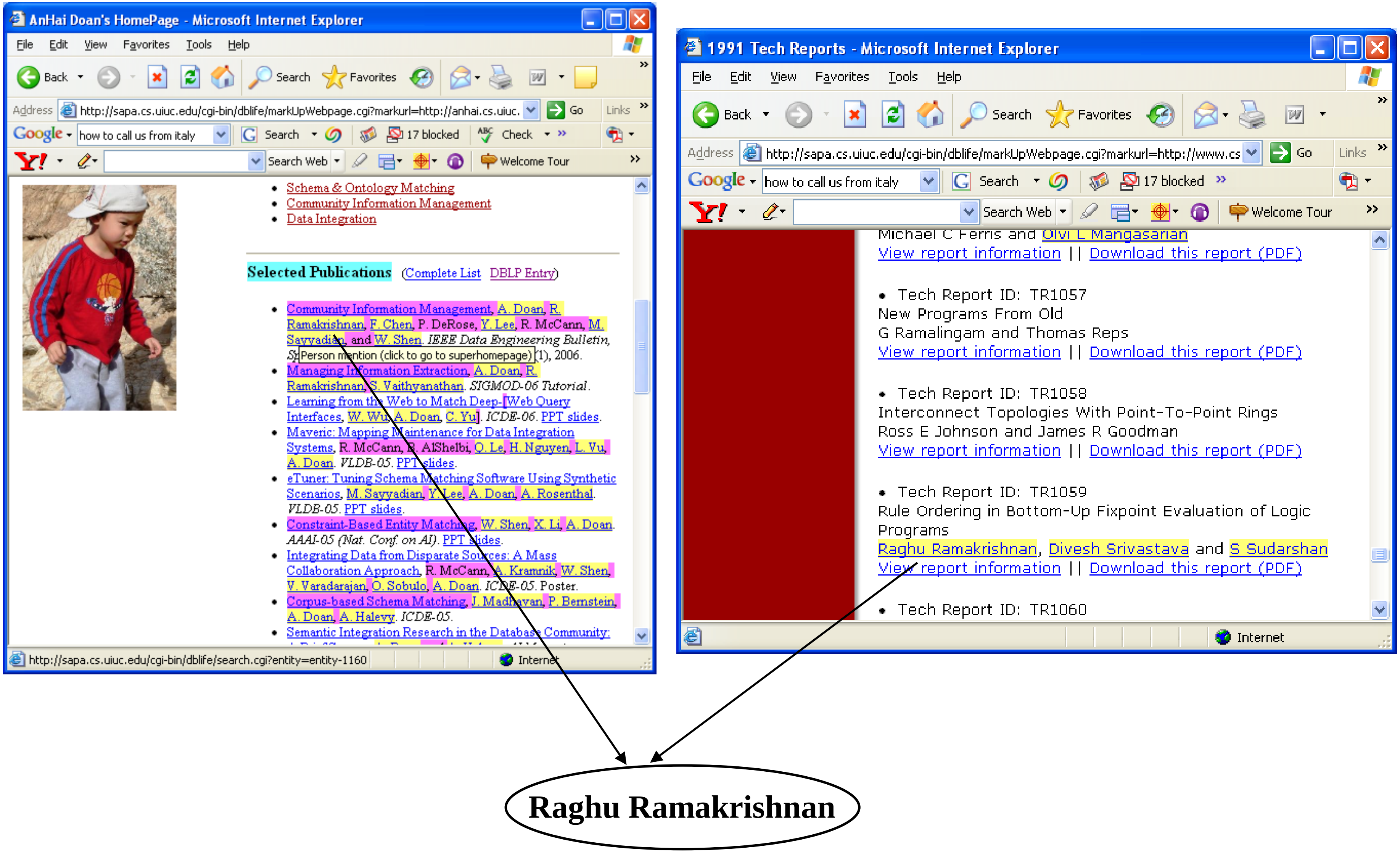
- Schema & Ontology Matching
- Community Information Management
- Data Integration

Selected Publications (Complete List DBLP Entry)

- Community Information Management, A. Doan, R. Ramakrishnan, F. Chen, P. DeRose, Y. Lee, R. McCann, M. Sayyadian, and W. Shen. *IEEE Data Engineering Bulletin, Sp* Person mention (click to go to superhomepage) (1), 2006.
- Managing Information Extraction, A. Doan, R. Ramakrishnan, S. Vaithyanathan. *SIGMOD-06 Tutorial*.
- Learning from the Web to Match Deep-[Web Query Interfaces, W. Wu, A. Doan, C. Yu]. *ICDE-06*. PPT slides.
- Maveric: Mapping Maintenance for Data Integration Systems, R. McCann, B. AlShelbi, Q. Le, H. Nguyen, L. Vu, A. Doan. *VLDB-05*. PPT slides.
- eTuner: Tuning Schema Matching Software Using Synthetic Scenarios, M. Sayyadian, Y. Lee, A. Doan, A. Rosenthal. *VLDB-05*. PPT slides.
- Constraint-Based Entity Matching, W. Shen, X. Li, A. Doan. *AAAI-05 (Nat. Conf. on AI)*. PPT slides.
- Integrating Data from Disparate Sources: A Mass Collaboration Approach, R. McCann, A. Kramnik, W. Shen, V. Varadarajan, O. Sobulo, A. Doan. *ICDE-05*. Poster.
- Corpus-based Schema Matching, J. Madhavan, P. Bernstein, A. Doan, A. Halevy. *ICDE-05*.
- Semantic Integration Research in the Database Community:

http://sapa.cs.uiuc.edu/cgi-bin/dblife/search.cgi?entity=entity-1160

Data Cleaning, Matching, Fusion



AnHai Doan's HomePage - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address Go Links
Google how to call us from italy Search 17 blocked Check
Y! Search Web Welcome Tour
Internet

http://sapa.cs.uiuc.edu/cgi-bin/dblife/markUpWebpage.cgi?markurl=http://anhai.cs.uiuc.

- Schema & Ontology Matching
- Community Information Management
- Data Integration

Selected Publications (Complete List DBLP Entry)

- Community Information Management, A. Doan, R. Ramakrishnan, F. Chen, P. DeRose, Y. Lee, R. McCann, M. Sayyadian, and W. Shen. *IEEE Data Engineering Bulletin, Sp* Person mention (click to go to superhomepage) (1), 2006.
- Managing Information Extraction, A. Doan, R. Ramakrishnan, S. Vaithyanathan. *SIGMOD-06 Tutorial*.
- Learning from the Web to Match Deep-[Web Query Interfaces, W. Wu, A. Doan, C. Yu]. *ICDE-06*. PPT slides.
- Maveric: Mapping Maintenance for Data Integration Systems, R. McCann, B. AlShelbi, Q. Le, H. Nguyen, L. Vu, A. Doan. *VLDB-05*. PPT slides.
- eTuner: Tuning Schema Matching Software Using Synthetic Scenarios, M. Sayyadian, Y. Lee, A. Doan, A. Rosenthal. *VLDB-05*. PPT slides.
- Constraint-Based Entity Matching, W. Shen, X. Li, A. Doan. *AAAI-05 (Nat. Conf. on AI)*. PPT slides.
- Integrating Data from Disparate Sources: A Mass Collaboration Approach, R. McCann, A. Kramnik, W. Shen, V. Varadarajan, O. Sobulo, A. Doan. *ICDE-05*. Poster.
- Corpus-based Schema Matching, J. Madhavan, P. Bernstein, A. Doan, A. Halevy. *ICDE-05*.
- Semantic Integration Research in the Database Community:

http://sapa.cs.uiuc.edu/cgi-bin/dblife/search.cgi?entity=entity-1160

1991 Tech Reports - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address Go Links
Google how to call us from italy Search 17 blocked
Y! Search Web Welcome Tour
Internet

http://sapa.cs.uiuc.edu/cgi-bin/dblife/markUpWebpage.cgi?markurl=http://www.cs

Michael C Ferris and Olvi L Mangasarian
View report information || Download this report (PDF)

- Tech Report ID: TR1057
New Programs From Old
G Ramalingam and Thomas Reps
View report information || Download this report (PDF)

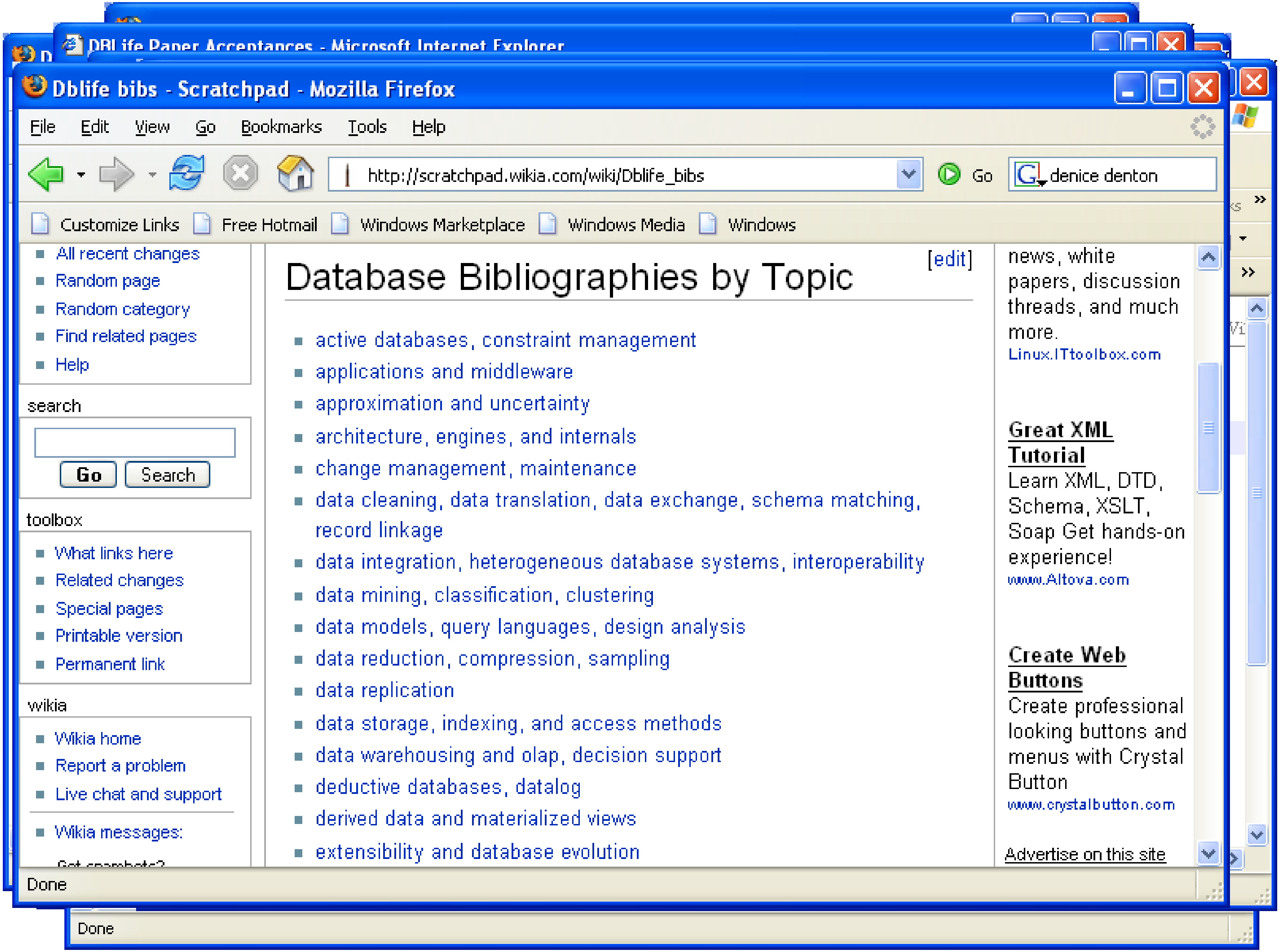
- Tech Report ID: TR1058
Interconnect Topologies With Point-To-Point Rings
Ross E Johnson and James R Goodman
View report information || Download this report (PDF)

- Tech Report ID: TR1059
Rule Ordering in Bottom-Up Fixpoint Evaluation of Logic Programs
Raghu Ramakrishnan, Divesh Srivastava and S Sudarshan
View report information || Download this report (PDF)

- Tech Report ID: TR1060

Raghu Ramakrishnan

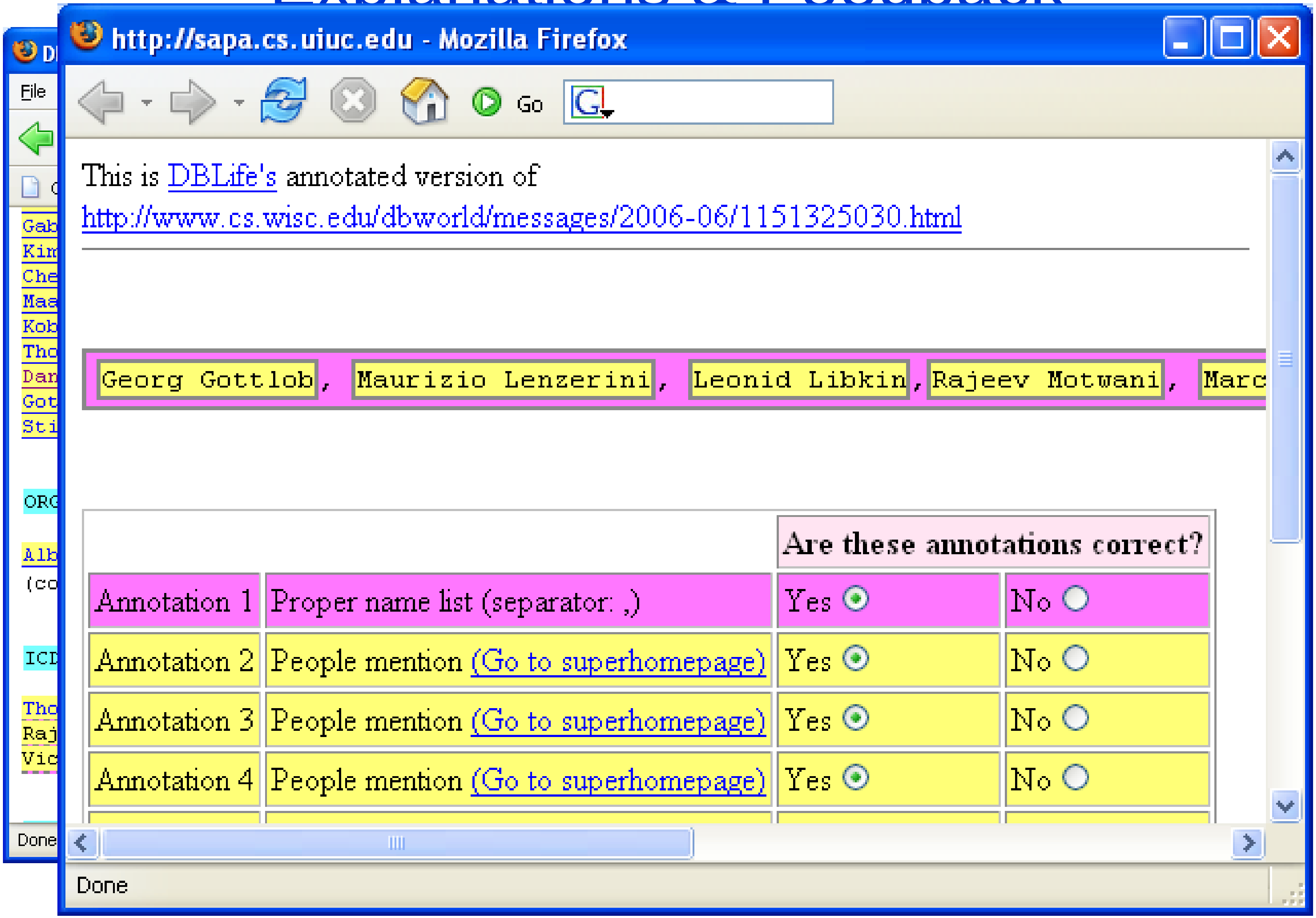
co-authors = A. Doan, Divesh Srivastava, ...



Database Bibliographies by Topic

- active databases, constraint management
- applications and middleware
- approximation and uncertainty
- architecture, engines, and internals
- change management, maintenance
- data cleaning, data translation, data exchange, schema matching, record linkage
- data integration, heterogeneous database systems, interoperability
- data mining, classification, clustering
- data models, query languages, design analysis
- data reduction, compression, sampling
- data replication
- data storage, indexing, and access methods
- data warehousing and olap, decision support
- deductive databases, datalog
- derived data and materialized views
- extensibility and database evolution

Explanations & Feedback



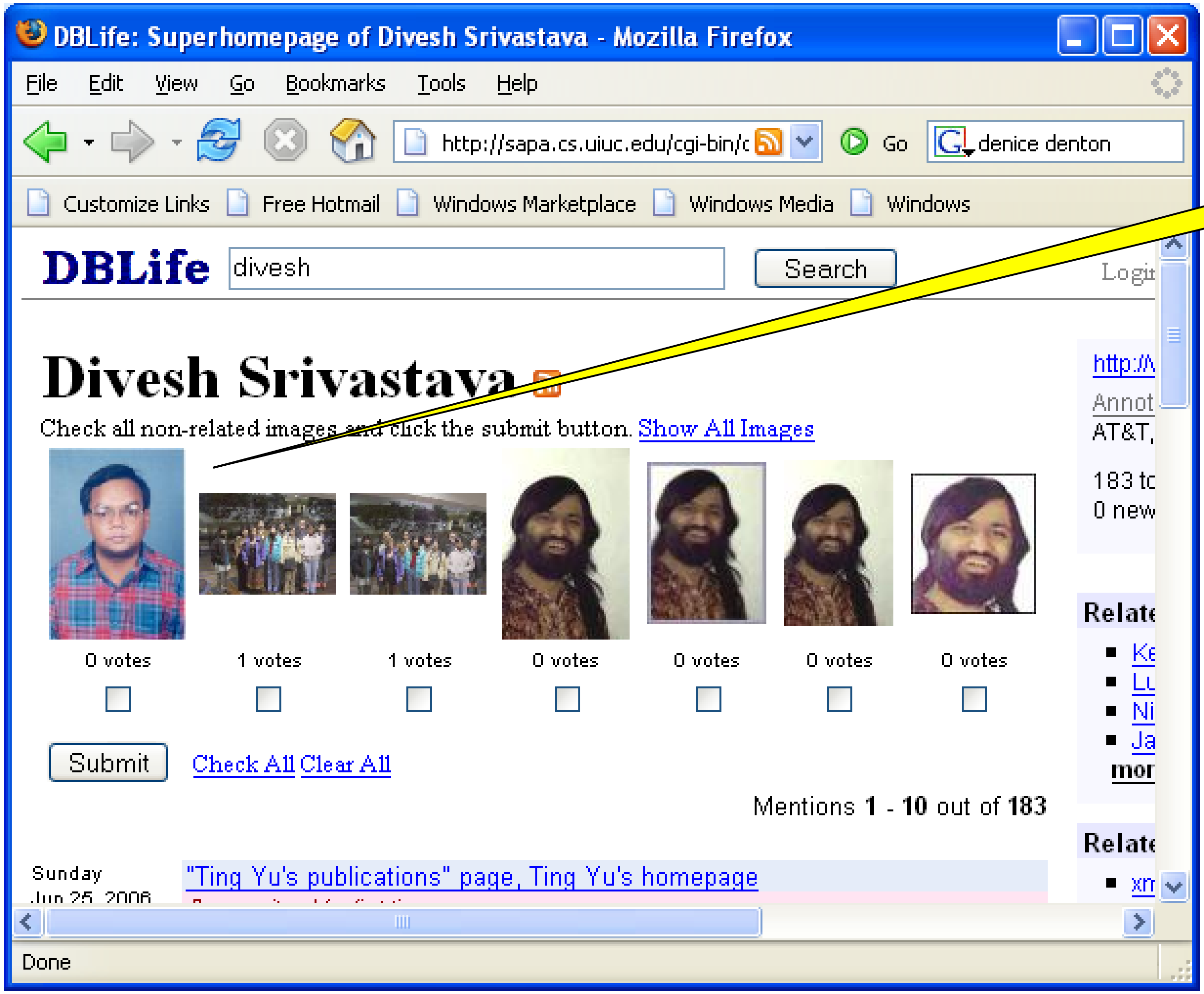
This is DBLife's annotated version of http://www.cs.wisc.edu/dbworld/messages/2006-06/1151325030.html

Georg Gottlob, Maurizio Lenzerini, Leonid Libkin, Rajeev Motwani, Marc

		Are these annotations correct?	
Annotation 1	Proper name list (separator: ,)	Yes ●	No ○
Annotation 2	People mention (Go to superhomepage)	Yes ●	No ○
Annotation 3	People mention (Go to superhomepage)	Yes ●	No ○
Annotation 4	People mention (Go to superhomepage)	Yes ●	No ○

Done

Mass Collaboration



Divesh Srivastava

Check all non-related images and click the submit button. Show All Images

0 votes	1 votes	1 votes	0 votes	0 votes	0 votes	0 votes
□	□	□	□	□	□	□

Check All Clear All

Mentions **1** - **10** out of **183**

Sunday
Jun 25, 2006

"Ting Yu's publications" page, Ting Yu's homepage

http://

Annot

AT&T,

183 to

0 new

Relate

- Ke
- Lu
- Ni
- Ja

mor

Relate

- xn

Not Divesh!

If enough users vote "not Divesh" on this picture, it is removed.

Current State of the Art

- Numerous domain-specific, hand-crafted solutions
  - imdb.com for movie domain
  - citeseer.com, dblp, rexa, Google scholar etc. for publication
  - techspec for engineering domain
- Very difficult to build and maintain,
  very hard to port solutions across domains
- The CIM Platform Challenge:
  - Develop a software platform that can be rapidly deployed and customized to manage data-rich Web communities
    - Creating an integrated, sustainable online community for, say, Chemical Engineering, or Finance, should be much easier, and should focus on leveraging domain knowledge, rather than on engineering details

Application 3: Scientific Data Management

AliBaba
@ Humboldt Univ. of Berlin

Summarizing PubMed Search Results

- PubMed/Medline
  - Database of paper abstracts in bioinformatics
  - 16 million abstracts, grows by 400K per year
- AliBaba: Summarizes results of keyword queries
  - User issues keyword query Q
  - AliBaba takes top 100 (say) abstracts returned by PubMed/Medline
  - Performs online entity and relationship extraction from abstracts
  - Shows ER graph to user
- For more detail
  - Contact Ulf Leser
  - System is online at http://wbi.informatik.hu-berlin.de:8080/

Examples of Entity-Relationship Extraction

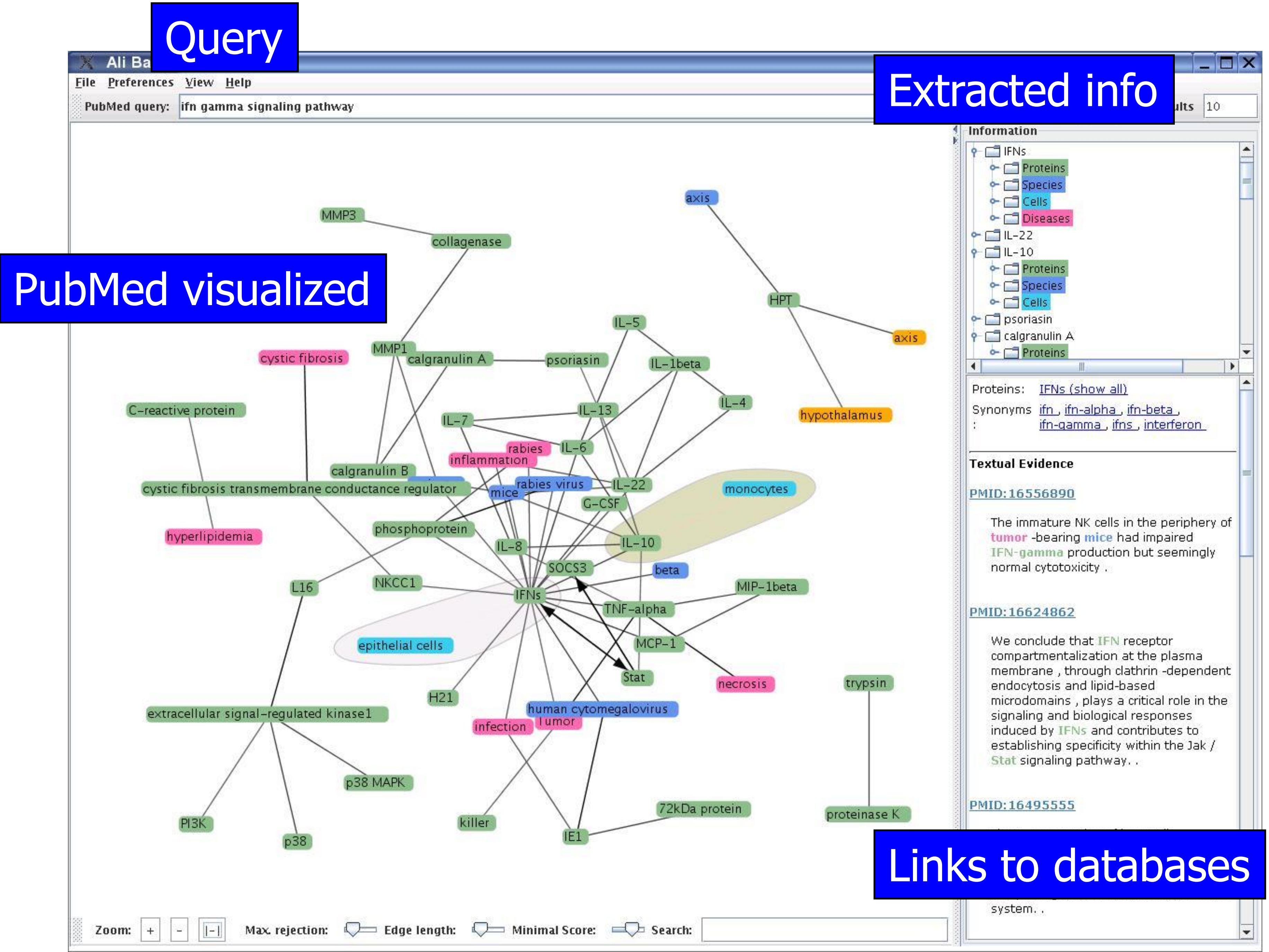
„We show that CBF-A and CBF-C interact with each other to form a CBF-A-CBF-C complex and that CBF-B does not interact with CBF-A or CBF-C individually but that it associates with the CBF-A-CBF-C complex.“

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

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

Another Example

Z-100 is an ***arabinomannan*** extracted from Mycobacterium tuberculosis that has various immunomodulatory activities, such as the induction of **interleukin 12**, **interferon gamma** (**IFN-gamma**) and beta-chemokines. The effects of ***Z-100*** on human immunodeficiency virus type 1 (HIV-1) replication in human **monocyte-derived macrophages** (**MDMs**) are investigated in this paper. In **MDMs**, ***Z-100*** markedly suppressed the replication of not only macrophage-tropic (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed amphotropic Moloney murine leukemia virus or vesicular stomatitis virus G envelopes. ***Z-100*** was found to inhibit HIV-1 expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv vector (in which the ***env*** gene is defective and the ***nef*** gene is replaced with the ***firefly luciferase*** gene) when this vector was transfected directly into **MDMs**. These findings suggest that ***Z-100*** inhibits virus replication, mainly at HIV-1 transcription. However, ***Z-100*** also downregulated expression of the cell surface receptors **CD4** and **CCR5** in **MDMs**, suggesting some inhibitory effect on HIV-1 entry. Further experiments revealed that ***Z-100*** induced **IFN-beta** production in these cells, resulting in induction of the 16-kDa **CCAAT/enhancer binding protein** (**C/EBP**) **beta transcription factor** that represses HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of **p38 mitogen-activated protein kinases** (**MAPK**), indicating that the **p38 MAPK** signalling pathway was involved in ***Z-100***-induced repression of HIV-1 replication in **MDMs**. These findings suggest that ***Z-100*** might be a useful immunomodulator for control of HIV-1 infection.


Query

Extracted info

PubMed visualized

Links to databases

PubMed query: ifn gamma signaling pathway

10

Proteins: IFNs (show all)

Synonyms : ifn , ifn-alpha , ifn-beta , ifn-gamma , ifns , interferon

Textual Evidence

PMID:16556890

The immature NK cells in the periphery of tumor -bearing mice had impaired IFN-gamma production but seemingly normal cytotoxicity .

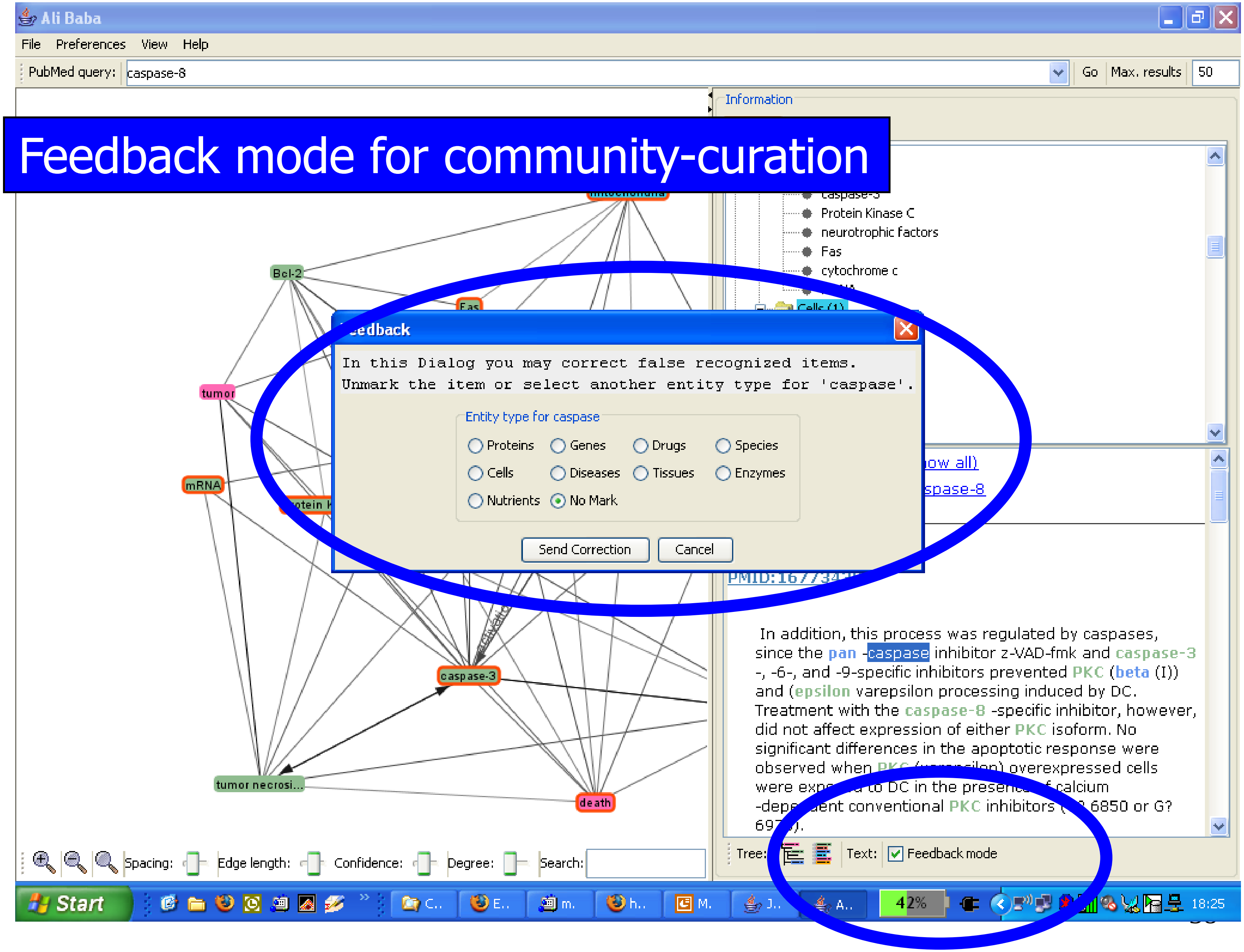
PMID:16624862

We conclude that IFN receptor compartmentalization at the plasma membrane , through clathrin -dependent endocytosis and lipid-based microdomains , plays a critical role in the signaling and biological responses induced by IFNs and contributes to establishing specificity within the Jak / Stat signaling pathway. .

PMID:16495555

system. .

Feedback mode for community-curation



PubMed query: caspase-8

Max. results: 50

Feedback

```
In this Dialog you may correct false recognized items.
Unmark the item or select another entity type for 'caspase'.
```

Entity type for caspase

○ Proteins ○ Genes ○ Drugs ○ Species

○ Cells ○ Diseases ○ Tissues ○ Enzymes

○ Nutrients ● No Mark

Send Correction | Cancel

In addition, this process was regulated by caspases, since the pan -caspase inhibitor z-VAD-fmk and caspase-3 -, -6-, and -9-specific inhibitors prevented PKC (beta (I)) and (epsilon varepsilon processing induced by DC. Treatment with the caspase-8 -specific inhibitor, however, did not affect expression of either PKC isoform. No significant differences in the apoptotic response were observed when PKC (varepsilon) overexpressed cells were exp d to DC in the presen f calcium -dep ent conventional PKC inhibitors (6850 or G? 69).

☑ Feedback mode

So we can do interesting and useful things with IE. And indeed there are many current IE efforts, and many with DB researchers involved

- AT&T Research, Boeing, CMU, Columbia, Google, IBM Almaden, IBM Yorktown, IIT-Mumbai, Lockheed-Martin, MIT, MSR, Stanford, UIUC, U. Mass, U. Washington, U. Wisconsin, Yahoo!

Still, these efforts have been carried out largely in isolation. In general, what does it take to build such an IE-based application?

Can we build a “System R” for IE-based applications?

To build a "System R" for IE applications, it turns out that

(1) It takes far more than what classical IE technologies offer

(2) Thus raising many open and important problems

(3) Several of which the DB community can address

The tutorial is about these three points

Tutorial Roadmap

- Introduction to managing IE [RR]
  - Motivation
  - What's different about *managing* IE?
- Major research directions
  - Extracting mentions of entities and relationships [SV]
    - Uncertainty management
  - Disambiguating extracted mentions [AD]
    - Tracking mentions and entities over time
  - Understanding, correcting, and maintaining extracted data [AD]
    - Provenance and explanations
    - Incorporating user feedback

Managing Information Extraction

Challenges in Real-Life IE, and Some Problems that the Database Community Can Address

Let’s Recap Classical IE

- Entity and relationship (link) extraction
  - Typically, these are done at the document level
- Entity resolution/matching
  - Done at the collection-level
- Efforts have focused mostly on
  - Improving the accuracy of IE algorithms for extracting entities/links
  - Scaling up IE algorithms to large corpora

Real-world IE applications need more!

- Complex IE tasks: Although not the focus of this tutorial, there is much work on extracting more complex concepts
  - Events
  - Opinions
  - Sentiments

Classical IE: Entity/Link Extraction

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…

```
Select  Name
From   PEOPLE
Where Organization = ‘Microsoft’
```

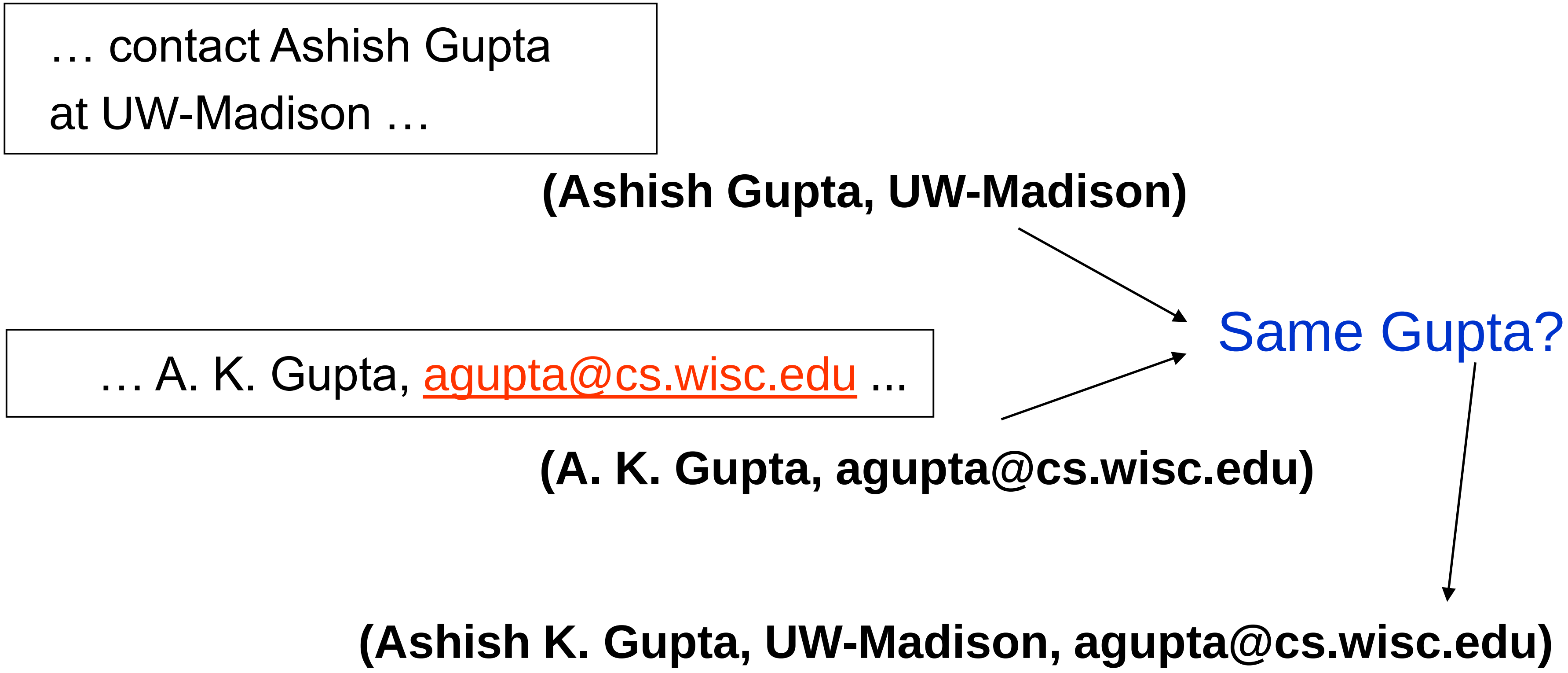
PEOPLE

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

Bill Gates

Bill Veghte

Classical IE: Entity Resolution (Mention Disambiguation / Matching)



… contact Ashish Gupta at UW-Madison …

(Ashish Gupta, UW-Madison)

Same Gupta?

… A. K. Gupta, agupta@cs.wisc.edu ...

(A. K. Gupta, agupta@cs.wisc.edu)

(Ashish K. Gupta, UW-Madison, agupta@cs.wisc.edu)

- Common, because text is inherently ambiguous; must disambiguate and merge extracted data

IE Meets Reality (Scratching the Surface)

1) Complications in Extraction and Disambiguation
   - Multi-step, user-guided workflows
     - In practice, developed iteratively
     - Each step must deal with uncertainty / errors of previous steps
   - Integrating multiple data sources
     - Extractors and workflows tuned for one source may not work well for another source
     - Cannot tune extraction manually for a large number of data sources
   - Incorporating background knowledge (e.g., dictionaries, properties of data sources, such as reliability/structure/patterns of change)
   - Continuous extraction, i.e., monitoring
     - Challenges: Reconciling prior results, avoiding repeated work, tracking real-world changes by analyzing changes in extracted data

IE Meets Reality (Scratching the Surface)

2) Complications in Understanding and Using Extracted Data

- Answering queries over extracted data, adjusting for extraction uncertainty and errors in a principled way
- Maintaining provenance of extracted data and generating understandable user-level explanations
- Incorporating user feedback to refine extraction/disambiguation
  - Want to correct specific mistake a user points out, and ensure that this is not “lost” in future passes of continuous monitoring scenarios
  - Want to generalize source of mistake and catch other similar errors (e.g., if Amer-Yahia pointed out error in extracted version of last name, and we recognize it is because of incorrect handling of hyphenation, we want to automatically apply the fix to all hyphenated last names)

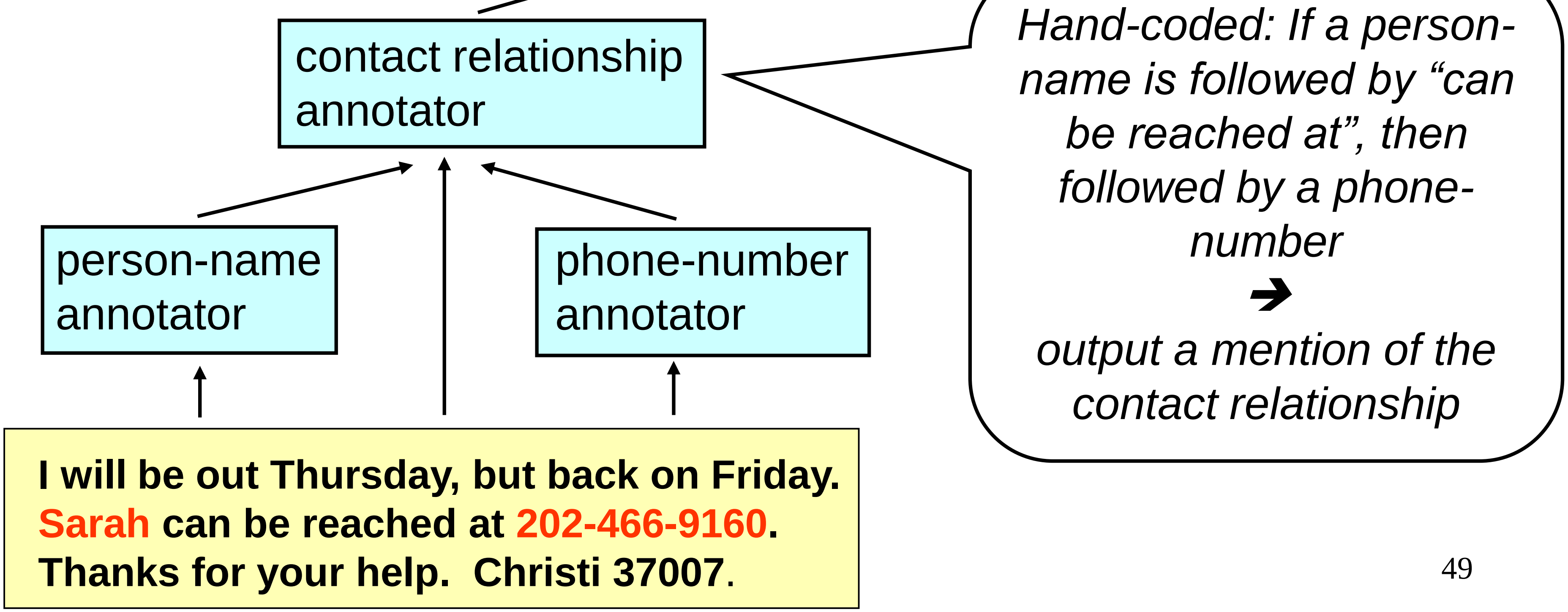
Workflows in Extraction Phase

- Example: extract Person's contact PhoneNumber

I will be out Thursday, but back on Friday.
Sarah can be reached at 202-466-9160.
Thanks for your help. Christi 37007.

⇨ Sarah's number is 202-466-9160

- A possible workflow



contact relationship annotator

person-name annotator

phone-number annotator

Hand-coded: If a person-name is followed by "can be reached at", then followed by a phone-number
➔
output a mention of the contact relationship

I will be out Thursday, but back on Friday.
Sarah can be reached at 202-466-9160.
Thanks for your help. Christi 37007.

Workflows in Entity Resolution

- Workflows also arise in the matching phase
- As an example, we will consider two different matching strategies used to resolve entities extracted from collections of user home pages and from the DBLP citation website
  - The key idea in this example is that a more liberal matcher can be used in a simple setting (user home pages) and the extracted information can then guide a more conservative matcher in a more confusing setting (DBLP pages)

Example: Entity Resolution Workflow

d_1: Gravano's Homepage

L. Gravano, K. Ross.
Text Databases. SIGMOD 03

L. Gravano, J. Sanz.
Packet Routing. SPAA 91

d_2: Columbia DB Group Page

Members
L. Gravano K. Ross J. Zhou

L. Gravano, J. Zhou.
Text Retrieval. VLDB 04

d_3: DBLP

Luis Gravano, Kenneth Ross.
Digital Libraries. SIGMOD 04

Luis Gravano, Jingren Zhou.
Fuzzy Matching. VLDB 01

Luis Gravano, Jorge Sanz.
Packet Routing. SPAA 91

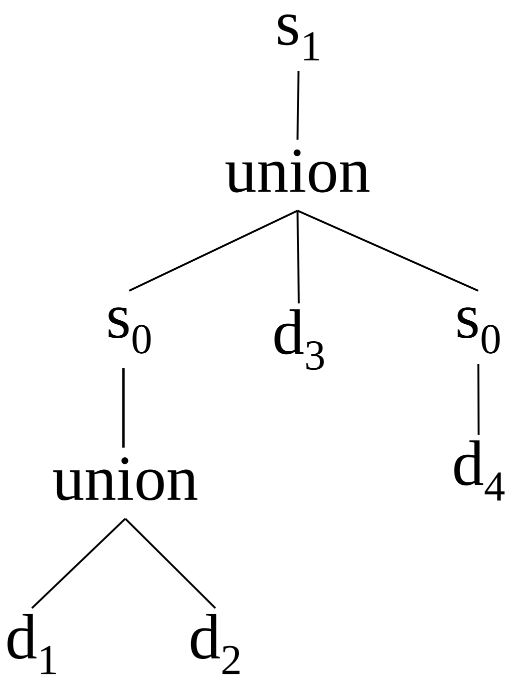
Chen Li, Anthony Tung.
Entity Matching. KDD 03

Chen Li, Chris Brown.
Interfaces. HCI 99

d_4: Chen Li's Homepage

C. Li.
Machine Learning. AAAI 04

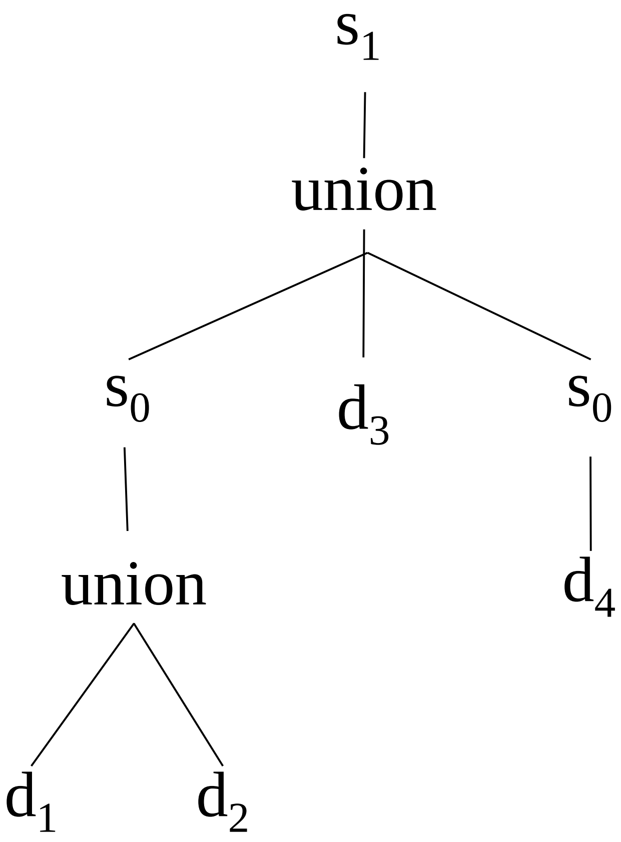
C. Li, A. Tung.
Entity Matching. KDD 03



s_0 matcher: Two mentions match if they share the same name.

s_1 matcher: Two mentions match if they share the same name and at least one co-author name.

Intuition Behind This Workflow



Since homepages are often unambiguous, we first match homepages using the simple matcher s0. This allows us to collect co-authors for Luis Gravano and Chen Li.

So when we finally match with tuples in DBLP, which is more ambiguous, we
(a) already have more evidence in the form
(b) of co-authors, and (b) can use the more conservative matcher s1.

Entity Resolution With Background Knowledge

… contact Ashish Gupta at UW-Madison …

(Ashish Gupta, UW-Madison)

Same Gupta?

(A. K. Gupta, agupta@cs.wisc.edu)

Entity/Link DB

A. K. Gupta	agupta@cs.wisc.edu
D. Koch	koch@cs.uiuc.edu

cs.wisc.edu	UW-Madison
cs.uiuc.edu	U. of Illinois

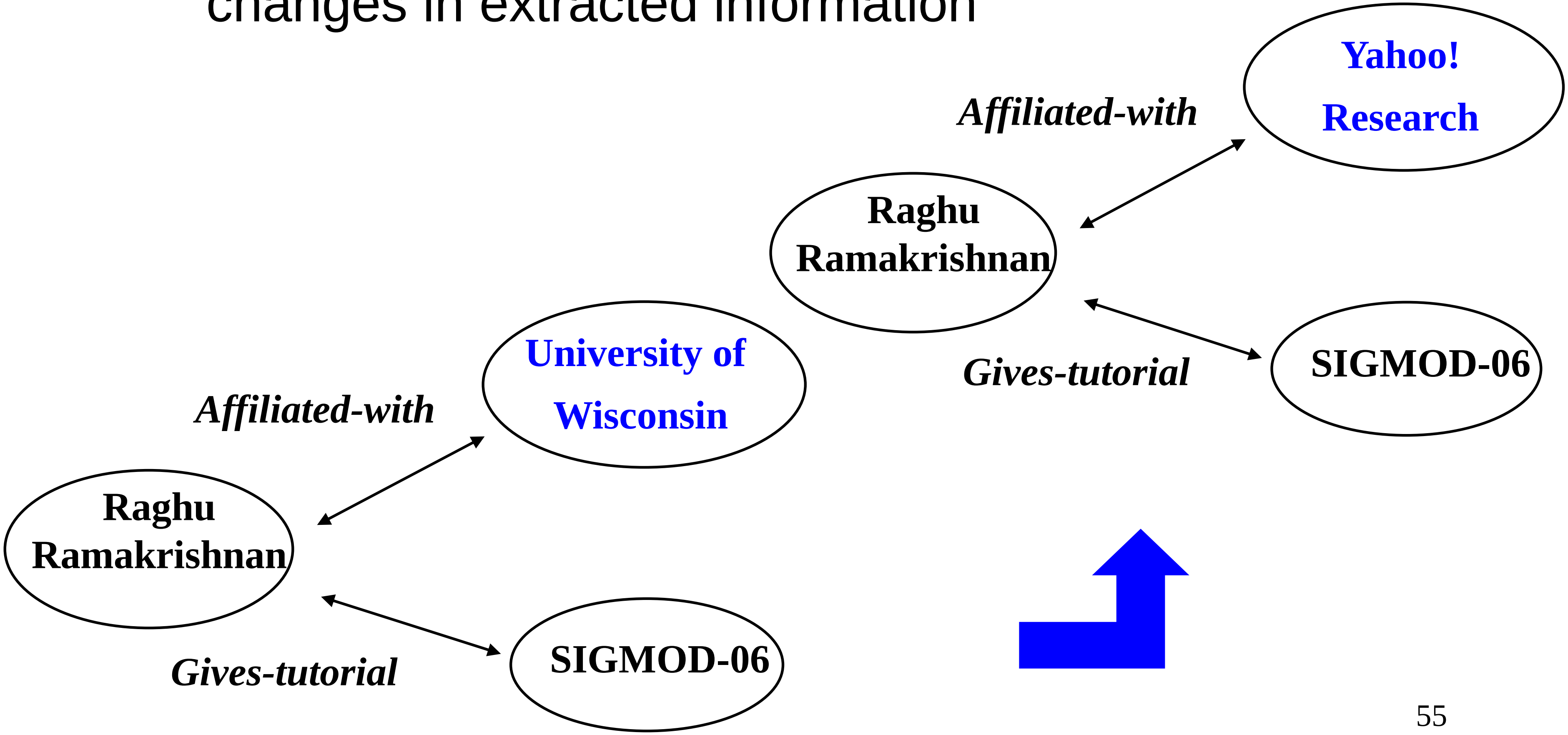
- Database of previously resolved entities/links
- Some other kinds of background knowledge:
  - “Trusted” sources (e.g., DBLP, DBworld) with known characteristics (e.g., format, update frequency)

Continuous Entity Resolution

- What if Entity/Link database is continuously updated to reflect changes in the real world? (E.g., Web crawls of user home pages)
- Can use the fact that few pages are new (or have changed) between updates. Challenges:
  - How much belief in *existing* entities and links?
  - Efficient organization and indexing
    - Where there is no meaningful change, recognize this and minimize repeated work

Continuous ER and Event Detection

- The real world might have changed!
  - And we need to detect this by analyzing changes in extracted information



Real-life IE: What Makes Extracted Information Hard to Use/Understand

- The extraction process is riddled with errors
  - How should these errors be represented?
  - Individual annotators are black-boxes with an internal probability model and typically output only the probabilities. While composing annotators how should their combined uncertainty be modeled?
- Semantics for queries over extracted data must handle the inherent ambiguity
- Lots of work
  - Classics: Fuhr-Rollecke; Imielinski-Lipski; ProbView; Halpern; …
  - Recent: See March 2006 Data Engineering bulletin for special issue on probabilistic data management (includes Green-Tannen survey/discussion of several proposals)
  - Dalvi-Suciu tutorial in Sigmod 2005, Halpern tutorial in PODS 2006

Some Recent Work on Uncertainty

- Many representations proposed, e.g.,
  - Confidence scores; Or-sets; Hierarchical imprecision
- Lots of recent work on querying uncertain data
  - E.g., Dalvi-Suciu identified classes of easy (PTIME) and hard (P#) queries and gave PTIME processing algorithms for easy ones
  - E.g., Burdick et al. (VLDB 05) considered single-table aggregations and showed how to assign confidence scores to hierarchically imprecise data in an intuitive way
  - E.g., Trio project (ICDE 06) considering how lineage can constrain the values taken by an imprecisely known object
  - E.g., Deshpande et al. (VLDB 04) consider data acquisition
  - E.g., Fagin et al. (ICDT 03) consider data exchange

Real-life IE: What Makes Extracted Information Hard to Use/Understand

- Users want to “drill down” on extracted data
  - We need to be able to explain the basis for an extracted piece of information when users “drill down”.
  - Many proof-tree based explanation systems built in deductive DB / LP /AI communities (Coral, LDL, EKS-V1, XSB, McGuinness, …)
  - Studied in context of provenance of integrated data (Buneman et al.; Stanford warehouse lineage, and more recently Trio)
- Concisely explaining complex extractions (e.g., using statistical models, workflows, and reflecting uncertainty) is hard
  - And especially useful because users are likely to drill down when they are surprised or confused by extracted data (e.g., due to errors, uncertainty).

Provenance, Explanations

A. Gupta, D. Smith, Text mining, SIGMOD-06 → **System extracted "Gupta, D" as a person name**

Incorrect. But why?

System extracted "Gupta, D" using these rules:

(R1) David Gupta is a person name
(R2) If "first-name last-name" is a person name, then "last-name, f" is also a person name.

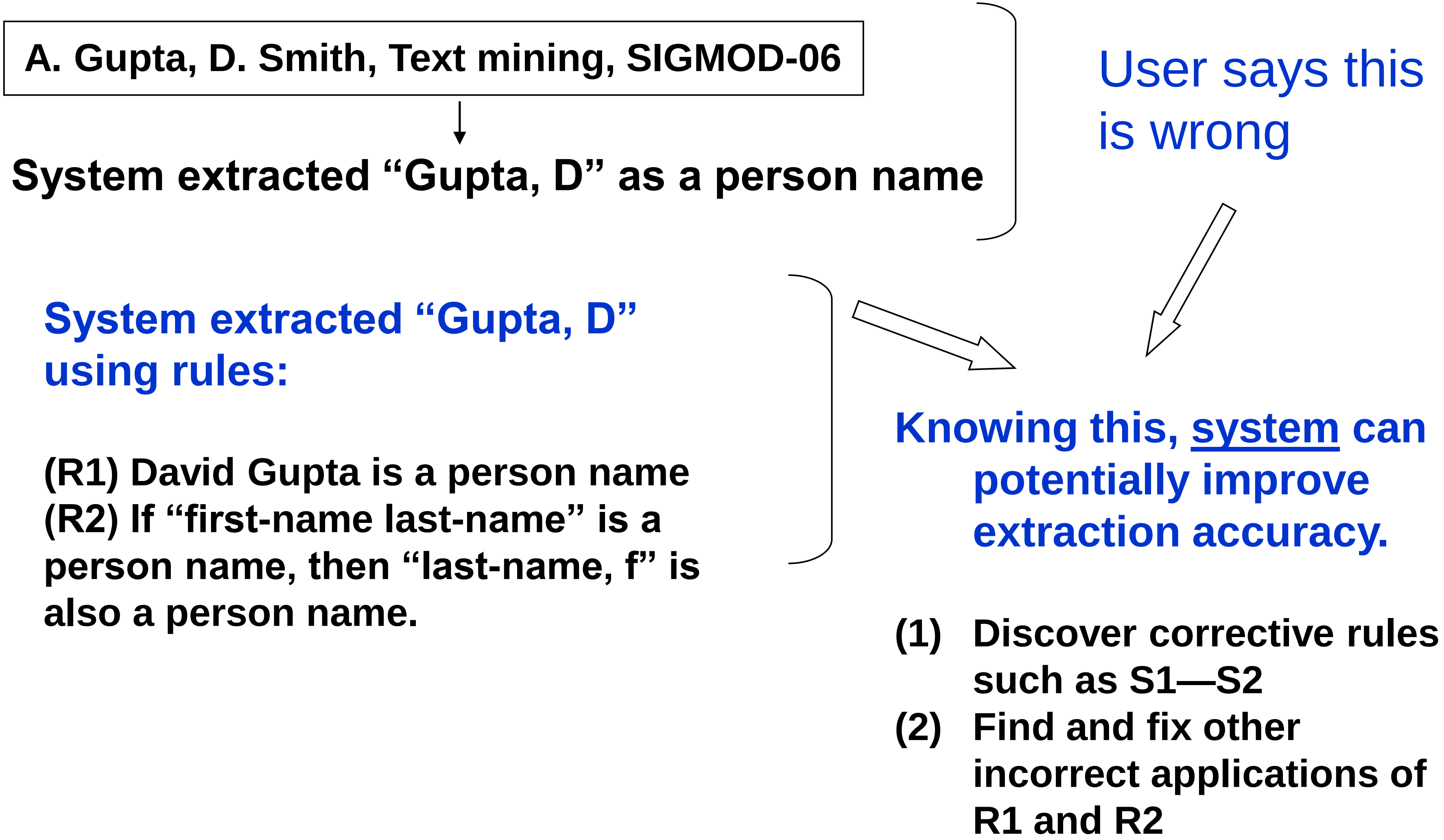
Knowing this, system builder can potentially improve extraction accuracy.

One way to do that:
(S1) Detect a list of items
(S2) If A straddles two items in a list
➔ A is not a person name

Real-life IE: What Makes Extracted Information Hard to Use/Understand

- Provenance becomes even more important if we want to leverage user feedback to improve the quality of extraction over time.
  - Maintaining an extracted “view” on a collection of documents over time is very costly; getting feedback from users can help
  - In fact, distributing the maintenance task across a large group of users may be the best approach
    - E.g., CIM

Incorporating Feedback



A. Gupta, D. Smith, Text mining, SIGMOD-06

System extracted "Gupta, D" as a person name

User says this is wrong

System extracted "Gupta, D" using rules:

(R1) David Gupta is a person name
(R2) If "first-name last-name" is a person name, then "last-name, f" is also a person name.

Knowing this, <u>system</u> can potentially improve extraction accuracy.

(1) Discover corrective rules such as S1—S2
(2) Find and fix other incorrect applications of R1 and R2

A general framework for incorporating feedback?

IE-Management Systems?

- In fact, everything about IE in practice is hard.
- Can we build a "System R for IE-in-practice"? *That's* the grand challenge of "Managing IE"
  - Key point: Such a platform must provide support for the range of tasks we've described, yet be readily customizable to new domains and applications

System Challenges

- Customizability to new applications
- Scalability
- Detecting broken extractors
- Efficient handling of previously extracted information when components (e.g., annotators, matchers) are upgraded
- …

Customizable Extraction

- Cannot afford to implement extraction, and extraction management, from scratch for each application.
- What tasks can we abstract into a platform that can be customized for different applications? What needs to be customizable?
  - “Schema” level definition of entity and link concepts
  - Extraction libraries
  - Choices in how to handle uncertainty
  - Choices in how to provide / incorporate feedback
  - Choices in entity resolution and integration decisions
  - Choices in frequency of updates, etc.

Scaling Up: Size is Just One Dimension!

- Corpus size
- Number of corpora
- Rate of change
- Size of extraction library
- Complexity of concepts to extract
- Complexity of background knowledge
- Complexity of guaranteeing uncertainty semantics when querying or updating extracted data

OK. But Why Now is the Right Time?

1. Emerging Attempts to Go Beyond Improving Accuracy of Single IE Algorithm

- Researchers are starting to examine
  - How to make blackboxes run efficiently [Sarawagi et al.]
  - How to integrate blackboxes
    - Combine IE and entity matching [McCallum etc.]
    - Combine multiple IE systems [Alpa et. al.]
- Attempts to standardize API of blackboxes, to ensure plug and play
  - GATE, UIMA, etc.
- Growing awareness of previously mentioned issues
  - Uncertainty management / provenance
  - Scalability
  - Exploiting user knowledge / user interaction
  - Exploit extracted data effectively

2. Multiple Efforts to Build IE Applications, in Industry and Academia

- However, each in isolation
  - Citeseer, Cora, Rexa, Dblife, what else?
  - Numerous systems in industry
    - Web search engines use IE to add some semantics to search (e.g., recognize place names), and to do better ad placement
    - Enterprise search, business intelligence
- We should share knowledge now

Summary

- Lots of text, and growing …
- IE can help us to better leverage text
- Managing the entire IE process is important
- Lot of opportunities for the DB community

Tutorial Roadmap

- Introduction to managing IE [RR]
  - Motivation
  - What’s different about *managing* IE?
- Major research directions
  - Extracting mentions of entities and relationships [SV]
    - Uncertainty management
  - Disambiguating extracted mentions [AD]
    - Tracking mentions and entities over time
  - Understanding, correcting, and maintaining extracted data [AD]
    - Provenance and explanations
    - Incorporating user feedback

Extracting Mentions of Entities and Relationships

Popular IE Tasks

- Named-entity extraction
  - Identify named-entities such as Persons, Organizations etc.
- Relationship extraction
  - Identify relationships between individual entities, e.g., Citizen-of, Employed-by etc.
  - e.g., Yahoo! acquired startup Flickr
- Event detection
  - Identifying incident occurrences between potentially multiple entities such Company-mergers, transfer-ownership, meetings, conferences, seminars etc.

But IE is Much, Much More ..

- Lesser known entities
  - Identifying rock-n-roll bands, restaurants, fashion designers, directions, passwords etc.
- Opinion / review extraction
  - Detect and extract informal reviews of bands, restaurants etc. from weblogs
  - Determine whether the opinions can be positive or negative

Email Example: Identify emails that contain directions

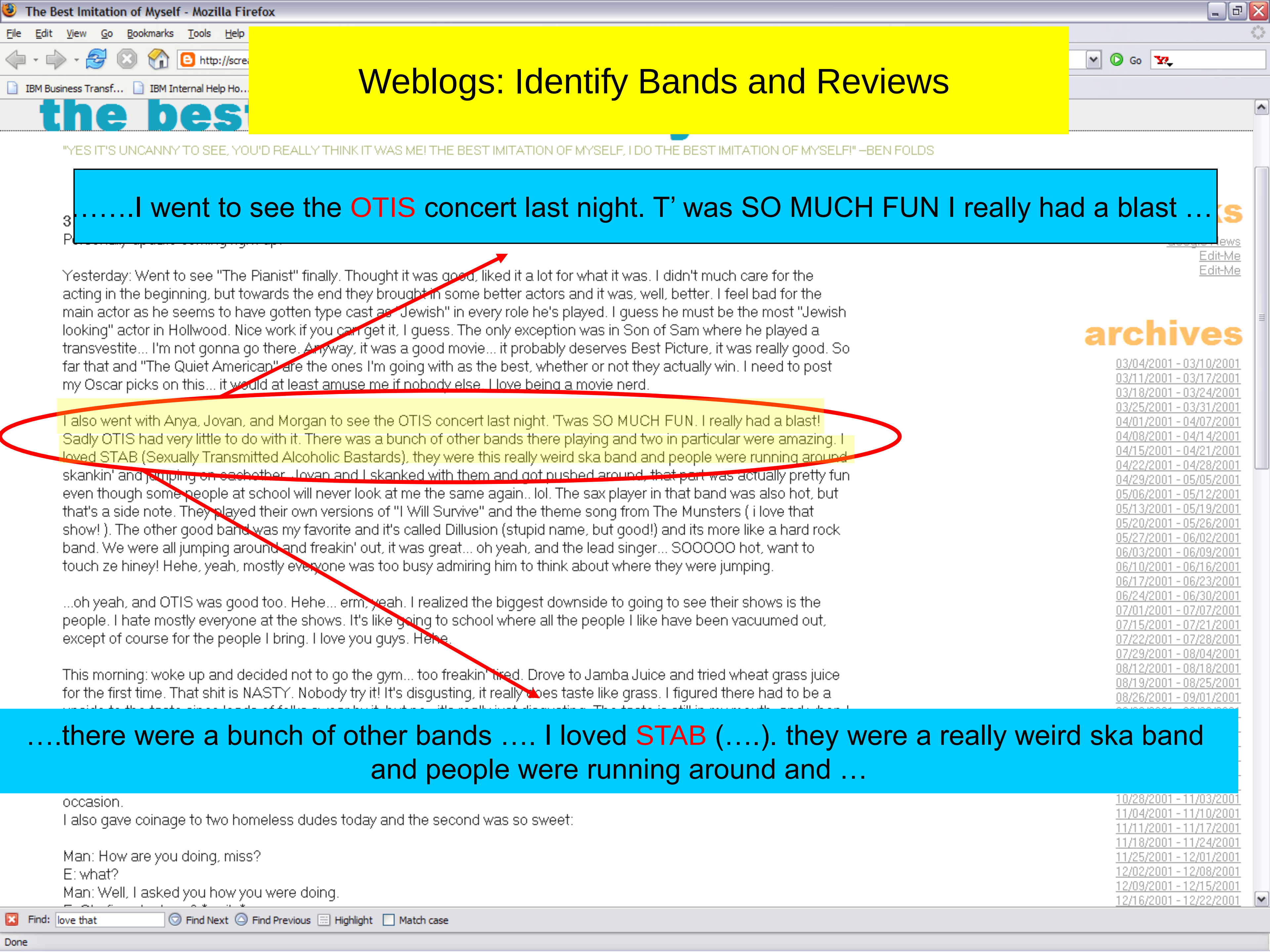
From: Shively, Hunter S.

Date: Tue, 26 Jun 2001 13:45:01 -0700 (PDT)

I-10W to exit 730 Peachridge RD (1 exit past Brookshire). Turn left on Peachridge RD. 2 miles down on the right--turquois 'horses for sale' sign

From the Enron email collection

Weblogs: Identify Bands and Reviews



"YES IT'S UNCANNY TO SEE, YOU'D REALLY THINK IT WAS ME! THE BEST IMITATION OF MYSELF, I DO THE BEST IMITATION OF MYSELF!" –BEN FOLDS

……I went to see the OTIS concert last night. T' was SO MUCH FUN I really had a blast …

Yesterday: Went to see "The Pianist" finally. Thought it was good, liked it a lot for what it was. I didn't much care for the acting in the beginning, but towards the end they brought in some better actors and it was, well, better. I feel bad for the main actor as he seems to have gotten type cast as "Jewish" in every role he's played. I guess he must be the most "Jewish looking" actor in Hollwood. Nice work if you can get it, I guess. The only exception was in Son of Sam where he played a transvestite... I'm not gonna go there. Anyway, it was a good movie... it probably deserves Best Picture, it was really good. So far that and "The Quiet American" are the ones I'm going with as the best, whether or not they actually win. I need to post my Oscar picks on this... it would at least amuse me if nobody else. I love being a movie nerd.

I also went with Anya, Jovan, and Morgan to see the OTIS concert last night. 'Twas SO MUCH FUN. I really had a blast! Sadly OTIS had very little to do with it. There was a bunch of other bands there playing and two in particular were amazing. I loved STAB (Sexually Transmitted Alcoholic Bastards), they were this really weird ska band and people were running around skankin' and jumping on eachother. Jovan and I skanked with them and got pushed around, that part was actually pretty fun even though some people at school will never look at me the same again.. lol. The sax player in that band was also hot, but that's a side note. They played their own versions of "I Will Survive" and the theme song from The Munsters (i love that show!). The other good band was my favorite and it's called Dillusion (stupid name, but good!) and its more like a hard rock band. We were all jumping around and freakin' out, it was great... oh yeah, and the lead singer... SOOOOO hot, want to touch ze hiney! Hehe, yeah, mostly everyone was too busy admiring him to think about where they were jumping.

...oh yeah, and OTIS was good too. Hehe... erm, yeah. I realized the biggest downside to going to see their shows is the people. I hate mostly everyone at the shows. It's like going to school where all the people I like have been vacuumed out, except of course for the people I bring. I love you guys. Hehe.

This morning: woke up and decided not to go the gym... too freakin' tired. Drove to Jamba Juice and tried wheat grass juice for the first time. That shit is NASTY. Nobody try it! It's disgusting, it really does taste like grass. I figured there had to be a

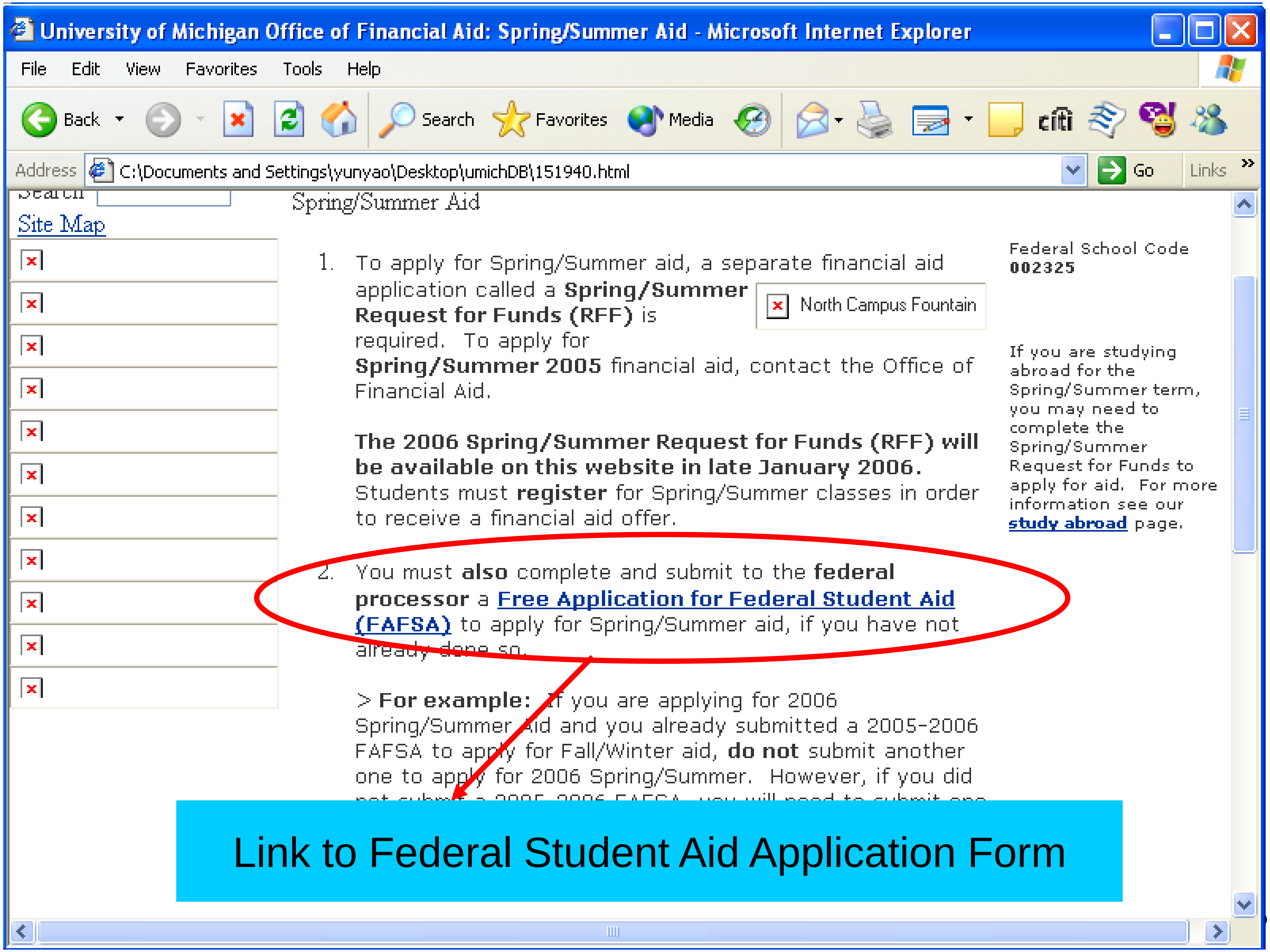
….there were a bunch of other bands …. I loved STAB (….). they were a really weird ska band and people were running around and …

occasion.
I also gave coinage to two homeless dudes today and the second was so sweet:

Man: How are you doing, miss?
E: what?
Man: Well, I asked you how you were doing.

03/04/2001 - 03/10/2001
03/11/2001 - 03/17/2001
03/18/2001 - 03/24/2001
03/25/2001 - 03/31/2001
04/01/2001 - 04/07/2001
04/08/2001 - 04/14/2001
04/15/2001 - 04/21/2001
04/22/2001 - 04/28/2001
04/29/2001 - 05/05/2001
05/06/2001 - 05/12/2001
05/13/2001 - 05/19/2001
05/20/2001 - 05/26/2001
05/27/2001 - 06/02/2001
06/03/2001 - 06/09/2001
06/10/2001 - 06/16/2001
06/17/2001 - 06/23/2001
06/24/2001 - 06/30/2001
07/01/2001 - 07/07/2001
07/15/2001 - 07/21/2001
07/22/2001 - 07/28/2001
07/29/2001 - 08/04/2001
08/12/2001 - 08/18/2001
08/19/2001 - 08/25/2001
08/26/2001 - 09/01/2001
10/28/2001 - 11/03/2001
11/04/2001 - 11/10/2001
11/11/2001 - 11/17/2001
11/18/2001 - 11/24/2001
11/25/2001 - 12/01/2001
12/02/2001 - 12/08/2001
12/09/2001 - 12/15/2001
12/16/2001 - 12/22/2001

Intranet Web: Identify form-entry pages [Li et al, SIGIR, 2006]


University of Michigan Office of Financial Aid: Spring/Summer Aid - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Back
Search
Favorites
Media
Address
C:\Documents and Settings\yunyao\Desktop\umichDB\151940.html
Go
Links
Search
Site Map
North Campus Fountain

Spring/Summer Aid

1. To apply for Spring/Summer aid, a separate financial aid application called a **Spring/Summer Request for Funds (RFF)** is required. To apply for **Spring/Summer 2005** financial aid, contact the Office of Financial Aid.

   The 2006 Spring/Summer Request for Funds (RFF) will be available on this website in late January 2006. Students must **register** for Spring/Summer classes in order to receive a financial aid offer.

2. You must **also** complete and submit to the **federal processor** a **Free Application for Federal Student Aid (FAFSA)** to apply for Spring/Summer aid, if you have not already done so.

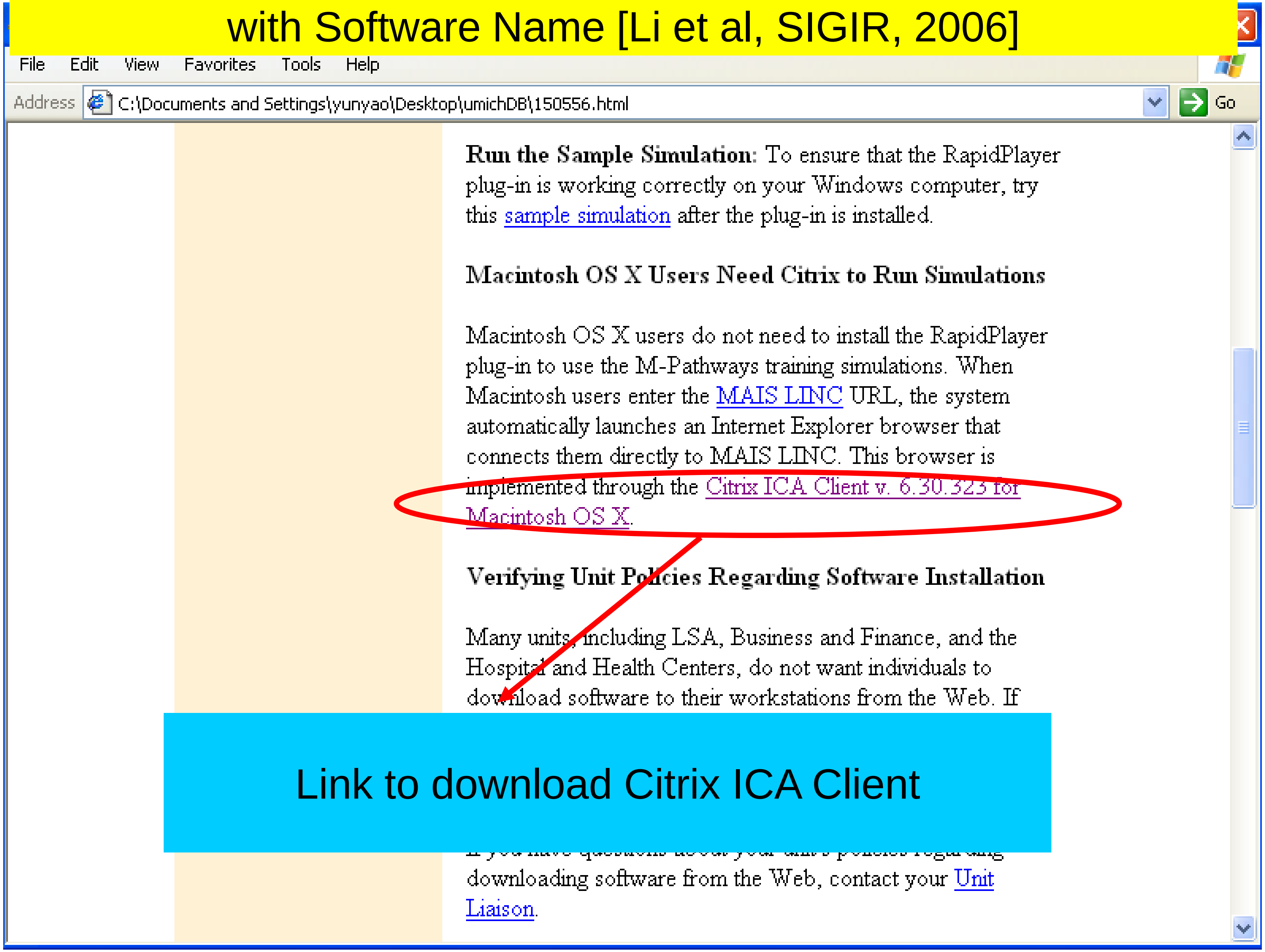
   > **For example:** If you are applying for 2006 Spring/Summer Aid and you already submitted a 2005-2006 FAFSA to apply for Fall/Winter aid, **do not** submit another one to apply for 2006 Spring/Summer. However, if you did not submit a 2005-2006 FAFSA, you will need to submit one

Federal School Code
002325

If you are studying abroad for the Spring/Summer term, you may need to complete the Spring/Summer Request for Funds to apply for aid. For more information see our **study abroad** page.

Link to Federal Student Aid Application Form

Intranet Web: Software download pages along with Software Name [Li et al, SIGIR, 2006]



C:\Documents and Settings\yunyao\Desktop\umichDB\150556.html

Run the Sample Simulation: To ensure that the RapidPlayer plug-in is working correctly on your Windows computer, try this sample simulation after the plug-in is installed.

Macintosh OS X Users Need Citrix to Run Simulations

Macintosh OS X users do not need to install the RapidPlayer plug-in to use the M-Pathways training simulations. When Macintosh users enter the MAIS LINC URL, the system automatically launches an Internet Explorer browser that connects them directly to MAIS LINC. This browser is implemented through the Citrix ICA Client v. 6.30.323 for Macintosh OS X.

Verifying Unit Policies Regarding Software Installation

Many units, including LSA, Business and Finance, and the Hospital and Health Centers, do not want individuals to download software to their workstations from the Web. If

If you have questions about your unit's policies regarding downloading software from the Web, contact your Unit Liaison.

Link to download Citrix ICA Client

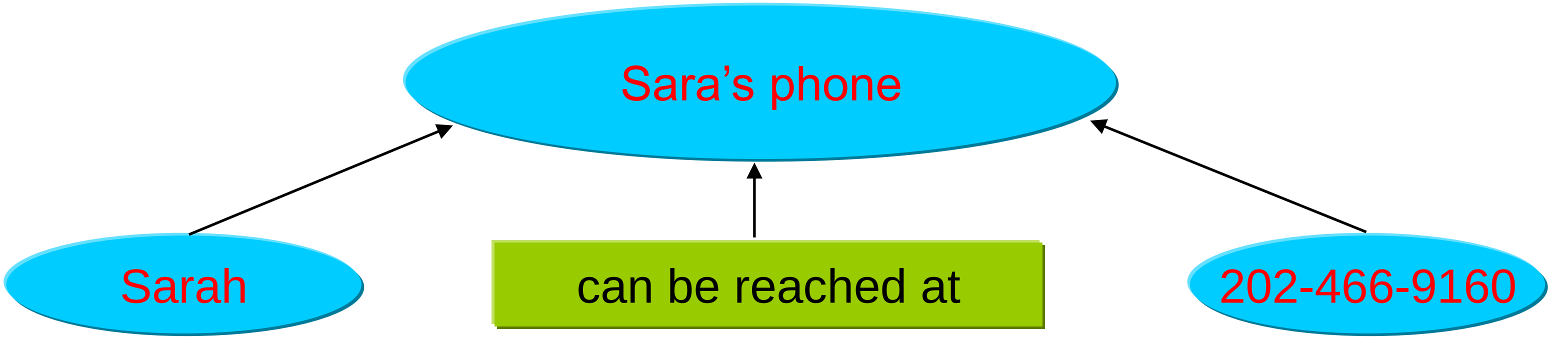
Workflows in Extraction

I will be out Thursday, but back on Friday.
Sarah can be reached at 202-466-9160.
Thanks for your help. Christi 37007.

⇨ Sarah’s phone is 202-466-9160

Single-shot extraction

Multi-step Workflow


Sara’s phone

Sarah

can be reached at

202-466-9160

Broadly-speaking two types of IE systems: hand-coded and learning-based.

What do they look like?
When best to use what?
Where can I learn more?

Lets start with hand-coded systems ...

Generic Template for hand-coded annotators

Document d

Previous annotations on document d

<u>Procedure Annotator (d, A_d)</u>

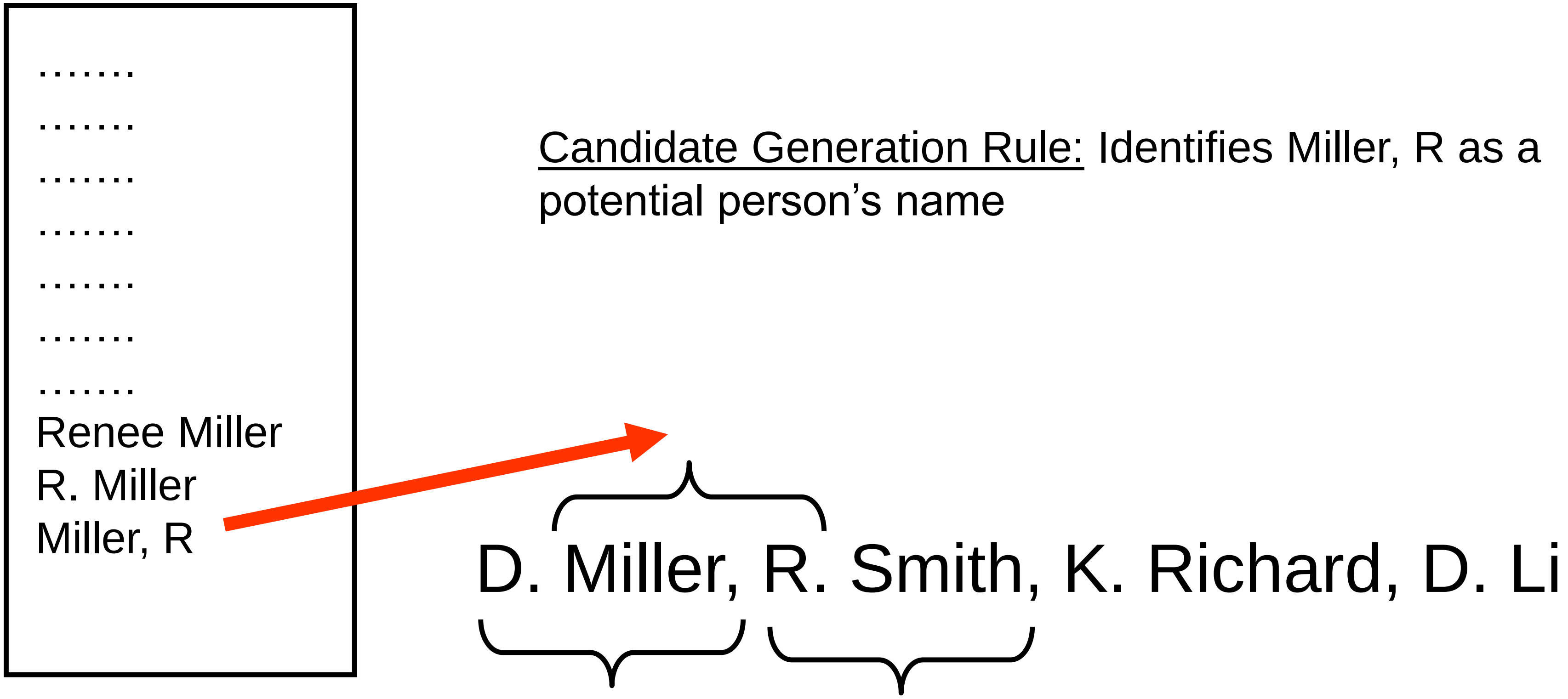
- R_f is a set of rules to generate features
- R_g is a set of rules to create candidate annotations
- R_c is a set of rules to consolidate annotations created by R_g

```
1.  Features = Compute_Features($R_f$, d)
2.  foreach r $\varepsilon$ $R_g$
    Candidates = Candidates U  ApplyRule (r, Features, $A_d$)
3.  Results =  Consolidate ($R_c$, Candidates)
    return Results
```

Simplified Real Example in DBLife

- Goal: build a simple person-name extractor
  - input: a set of Web pages W, DB Research People Dictionary DBN
  - output: all mentions of names in DBN
- Simplified DBLife Person-Name extraction
  - Obtain Features: HTML tags, detect lists of proper-names
  - Candidate Generation:
    - for each name e.g., David Smith
      - generate variants (V): “David Smith”, “D. Smith”, “Smith, D.”, etc.
      - obtain candidate person-names in W using V
  - Consolidation: if an occurrence straddles two proper-names then drop it

Compiled Dictionary



…….
…….
…….
…….
…….
…….
…….
Renee Miller
R. Miller
Miller, R

D. Miller, R. Smith, K. Richard, D. Li

<u>Candidate Generation Rule:</u> Identifies Miller, R as a potential person's name

Detected List of Proper-names

<u>Consolidation Rule:</u> If a candidate straddles two elements of the list then drop it

Example of Hand-coded Extractor [Ramakrishnan. G, 2005]

Rule 1 This rule will find person names with a salutation (e.g. Dr. Laura Haas) and two capitalized words

```
<token> INITIAL</token>
<token>DOT </token>
<token>CAPSWORD</token>
<token>CAPSWORD</token>
```

Rule 2 This rule will find person names where two capitalized words are present in a Person dictionary

```
<token>PERSONDICT, CAPSWORD </token>
<token>PERSONDICT, CAPSWORD</token>
```

```
CAPSWORD : Word starting with uppercase, second letter lowercase
           E.g., DeWitt will satisfy it (DEWITT will not)
           \p{Upper}\p{Lower}[\p{Alpha}]{1,25}
DOT      : The character '.'
```

Note that some names will be identified by both rules

Hand-coded rules can be artbitrarily complex

Find conference name in raw text

```
################################################################################
# Regular expressions to construct the pattern to extract conference names
################################################################################

# These are subordinate patterns
my $wordOrdinals="(?:first|second|third|fourth|fifth|sixth|seventh|eighth|ninth|tenth|eleventh|twelfth|thirteenth|fourteenth|fifteenth)";
my $numberOrdinals="(?:\\d?(?:1st|2nd|3rd|1th|2th|3th|4th|5th|6th|7th|8th|9th|0th))";
my $ordinals="(?:$wordOrdinals|$numberOrdinals)";
my $confTypes="(?:Conference|Workshop|Symposium)";
my $words="(?:[A-Z]\\w+\\s*)"; # A word starting with a capital letter and ending with 0 or more spaces
my $confDescriptors="(?:international\\s+|[A-Z]+\\s+)"; # .e.g "International Conference ...' or the conference name for workshops (e.g.
"VLDB Workshop ...")
my $connectors="(?:on|of)";
my $abbreviations="(?:\\([A-Z]\\w\\w+[\\W\\s]*?(?:\\d\\d+)?\\))"; # Conference abbreviations like "(SIGMOD'06)"

# The actual pattern we search for.  A typical conference name this pattern will find is
# "3rd International Conference on Blah Blah Blah (ICBBB-05)"
my
$fullNamePattern="((?:$ordinals\\s+$words*|$confDescriptors)?$confTypes(?:\\s+$connectors\\s+.*?|\\s+)?$abbreviations?)(?:\\n|\\r|\\.|<)";

############################### #################################
# Given a <dbworldMessage>, look for the conference pattern
#################################################################
lookForPattern($dbworldMessage, $fullNamePattern);

############################################################
# In a given <file>, look for occurrences of <pattern>
# <pattern> is a regular expression
############################################################
sub lookForPattern {
my ($file,$pattern) = @_;
```

Example Code of Hand-Coded Extractor

```
# Only look for conference names in the top 20 lines of the file
my $maxLines=20;
my $topOfFile=getTopOfFile($file,$maxLines);

# Look for the match in the top 20 lines - case insenstive, allow matches spanning multiple lines
if($topOfFile=~/(.*?)$pattern/is) {
my ($prefix,$name)=($1,$2);

# If it matches, do a sanity check and clean up the match
# Get the first letter
# Verify that the first letter is a capital letter or number
if(!($name=~/^\W*?[A-Z0-9]/)) { return (); }

# If there is an abbreviation, cut off whatever comes after that
if($name=~/^(.*?$abbreviations)/s) { $name=$1; }

# If the name is too long, it probably isn't a conference
if(scalar($name=~/[^\s]/g) > 100) { return (); }

# Get the first letter of the last word (need to this after chopping off parts of it due to abbreviation
my ($letter,$nonLetter)=("[A-Za-z]","[^A-Za-z]");
" $name"=~/$nonLetter($letter) $letter*$nonLetter*$/; # Need a space before $name to handle the first $nonLetter in the pattern if there is only one word in name

my $lastLetter=$1;
if(!($lastLetter=~/[A-Z]/)) { return (); } # Verify that the first letter of the last word is a capital letter

# Passed test, return a new crutch
return newCrutch(length($prefix),length($prefix)+length($name),$name,"Matched pattern in top $maxLines lines","conference name",getYear($name));
}
return ();
}
```

Some Examples of Hand-Coded Systems

- FRUMP [DeJong 82]
- CIRCUS / AutoSlog [Riloff 93]
- SRI FASTUS [Appelt, 1996]
- OSMX [Embley, 2005]
- DBLife [Doan et al, 2006]
- Avatar [Jayram et al, 2006]

Template for Learning based annotators

Procedure LearningAnnotator (D, L)

- D is the training data
- L is the labels

1. Preprocess D to extract features F
2. Use F,D & L to learn an extraction model E using a learning algorithm A

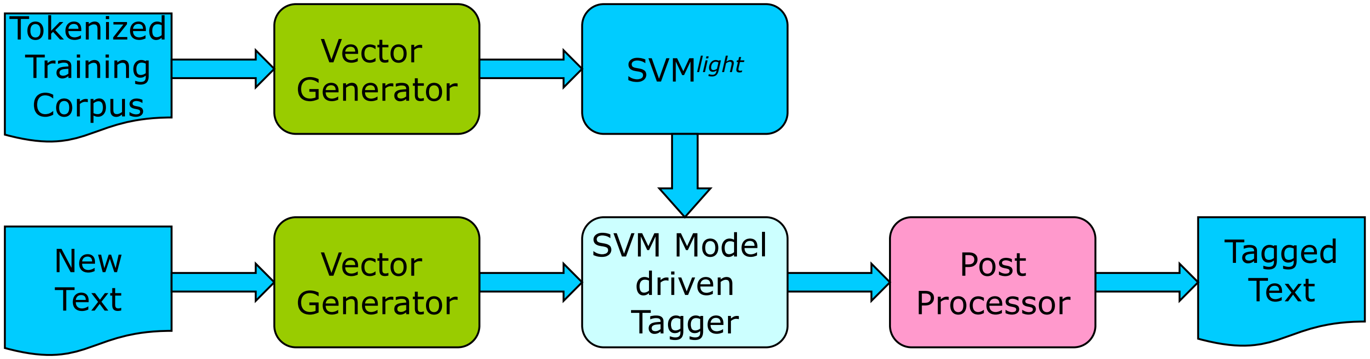
(Iteratively fine-tune parameters of the model and F)

Procedure ApplyAnnotator(d,E)

1. *Features = Compute_Features (d)*
2. *results = ApplyModel (E,Features, d)*
3. *return Results*

Real Example in AliBaba

- Extract gene names from PubMed abstracts
- Use Classifier (Support Vector Machine - SVM)



- Corpus of 7500 sentences
  - 140.000 non-gene words
  - 60.000 gene names
- SVM^{light} on different feature sets
- Dictionary compiled from Genbank, HUGO, MGD, YDB
- Post-processing for compound gene names

Learning-Based Information Extraction

- Naive Bayes
- SRV [Freitag-98], Inductive Logic Programming
- Rapier [Califf & Mooney-97]
- Hidden Markov Models [Leek, 1997]
- Maximum Entropy Markov Models [McCallum et al, 2000]
- Conditional Random Fields [Lafferty et al, 2000]

For an excellent and comprehensive view [Cohen, 2004]

Semi-Supervised IE Systems
Learn to Gather More Training Data

Only a seed set

1. Use labeled data to learn an extraction model E
2. Apply E to find mentions in document collection.
3. Construct more labeled data → T' is the new set.
4. Use T' to learn
   a hopefully better extraction model E'.
5. Repeat.

Expand the seed set

[DIPRE, Brin 98, Snowball, Agichtein & Gravano, 2000]

So there are basically two types of IE systems: hand-coded and learning-based.

What do they look like?

When best to use what?

Where can I learn more?

Hand-Coded Methods

- Easy to construct in many cases
  - e.g., to recognize prices, phone numbers, zip codes, conference names, etc.
- Easier to debug & maintain
  - especially if written in a “high-level” language (as is usually the case)
  - e.g.,

[From Avatar]

```
ContactPattern   ← RegularExpression(Email.body,”can be reached at”)

PersonPhone     ←  Precedes(Person
                             Precedes(ContactPattern, Phone, D),
                             D)
```

- Easier to incorporate / reuse domain knowledge
- Can be quite labor intensive to write

Learning-Based Methods

- Can work well when training data is easy to construct and is plentiful
- Can capture complex patterns that are hard to encode with hand-crafted rules
  - e.g., determine whether a review is positive or negative
  - extract long complex gene names

[From AliBaba]

The human T cell leukemia lymphotropic virus type 1 Tax protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300.“

- Can be labor intensive to construct training data
  - not sure how much training data is sufficient

Complementary to hand-coded methods

Where to Learn More

- Overviews / tutorials
  - Wendy Lehnert [Comm of the ACM, 1996]
  - Appelt [1997]
  - Cohen [2004]
  - Agichtein and Sarawai [KDD, 2006]
  - Andrew McCallum [ACM Queue, 2005]
- Systems / codes to try
  - OpenNLP
  - MinorThird
  - Weka
  - Rainbow

So what are the new IE challenges for IE-based applications?

First, lets discuss several observations, to motivate the new challenges

Observation 1:
We Often Need Complex Workflow

- What we have discussed so far are largely IE components
- Real-world IE applications often require a workflow that glue together these IE components
- These workflows can be quite large and complex
- Hard to get them right!

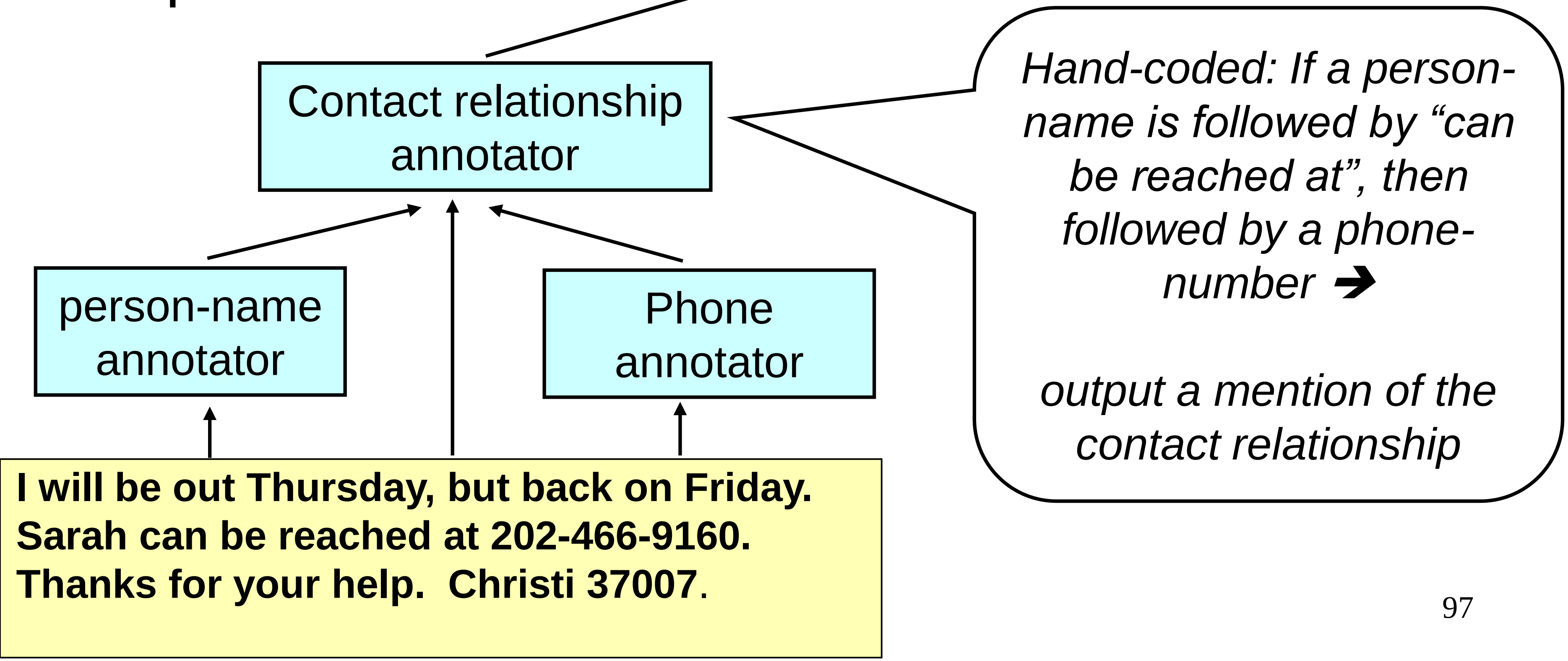
Illustrating Workflows

- Extract person’s contact phone-number from e-mail

I will be out Thursday, but back on Friday.
Sarah can be reached at 202-466-9160.
Thanks for your help. Christi 37007.

⇨ Sarah’s contact number is 202-466-9160

- A possible workflow



Contact relationship annotator

person-name annotator

Phone annotator

I will be out Thursday, but back on Friday.
Sarah can be reached at 202-466-9160.
Thanks for your help. Christi 37007.

Hand-coded: If a person-name is followed by “can be reached at”, then followed by a phone-number ➔

output a mention of the contact relationship

How Workflows are Constructed

- Define the information extraction task
  - e.g., identify people's phone numbers from email
- Identify the text-analysis components
  - E.g., tokenizer, part-of-speech tagger, Person, Phone annotator
- Compose different text-analytic components into a workflow
  - Several open-source plug-and-play architectures such as UIMA, GATE available
- Build domain-specific text-analytic component

How Workflows are Constructed

- Define the information extraction task
  - E.g., identify people’s phone numbers from email
- Identify the generic annotator components
  - E.g., tokenizer, part-of-speech tagger, Person, Phone annotator
- Compose di ferent text-an lytic compon nts into a workflow
  - Several open-so e plug-and-p archite es such as UIMA, GATE available
- Build domain-spe ext-a component

Generic text-analytic tasks.
Use available components

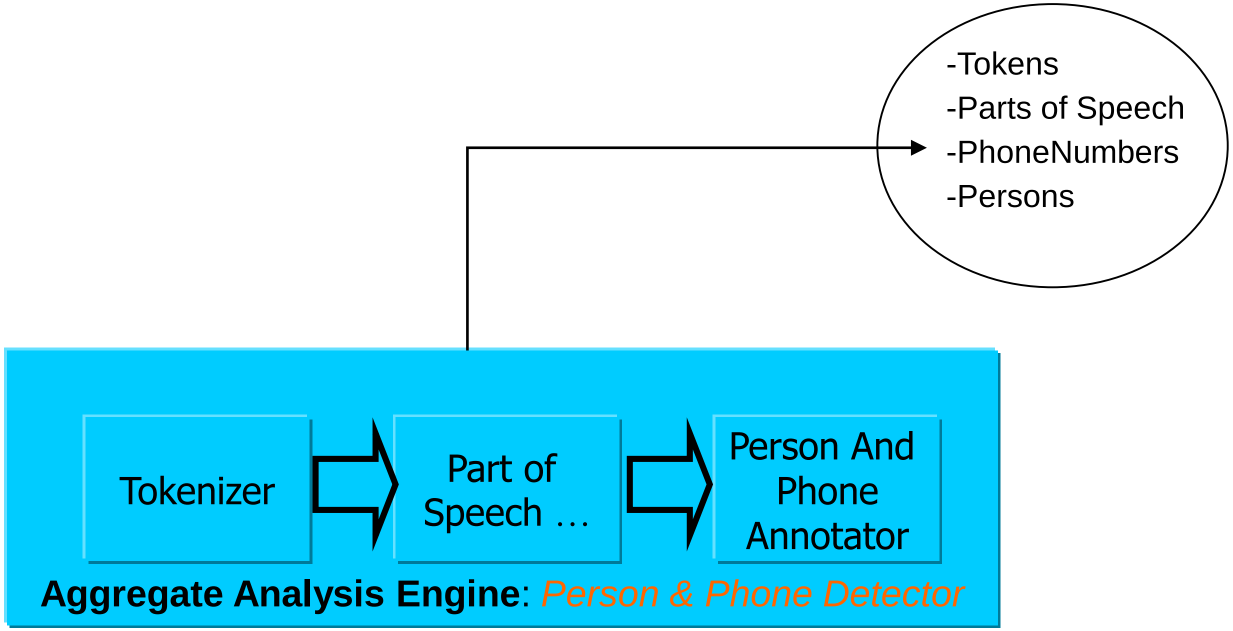
How Workflows are Constructed

- Define the information extraction task
  - E.g., identify people's phone numbers from email
- Identify the text-analysis components
  - E.g., tokenizer, part-of-speech tagger, Person, Phone annotator
- Compose different text-analytic components into a workflow
  - Several open-source plug-and-play architectures such as UIMA, GATE available
- Build domain-specific text-analytic component

How Workflows are Constructed

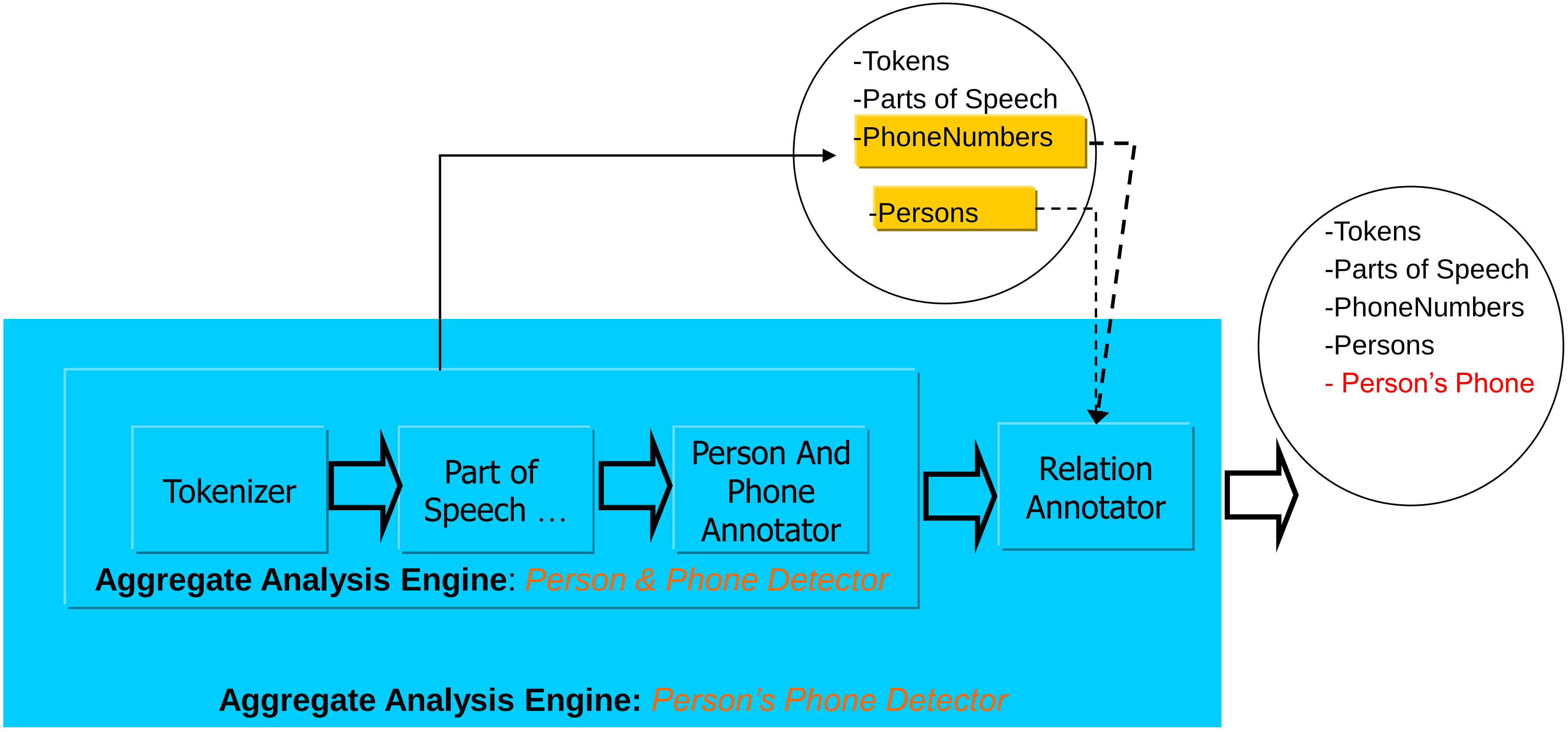
- Define the information extraction task
  - E.g., identify people's phone numbers from email
- Identify the generic text-analysis components
  - E.g., tokenizer, part-of-speech tagger, Person, Phone annotator
- Compose different text-analytic components into a workflow
  - Several open-source plug-and-play architectures such as UIMA, GATE available
- Build domain-specific text-analytic component
  - which is the contact relationship annotator in this example

UIMA & GATE


-Tokens
-Parts of Speech
-PhoneNumbers
-Persons
Tokenizer
Part of Speech …
Person And Phone Annotator
Aggregate Analysis Engine: *Person & Phone Detector*

Extracting Persons and Phone Numbers

UIMA & GATE

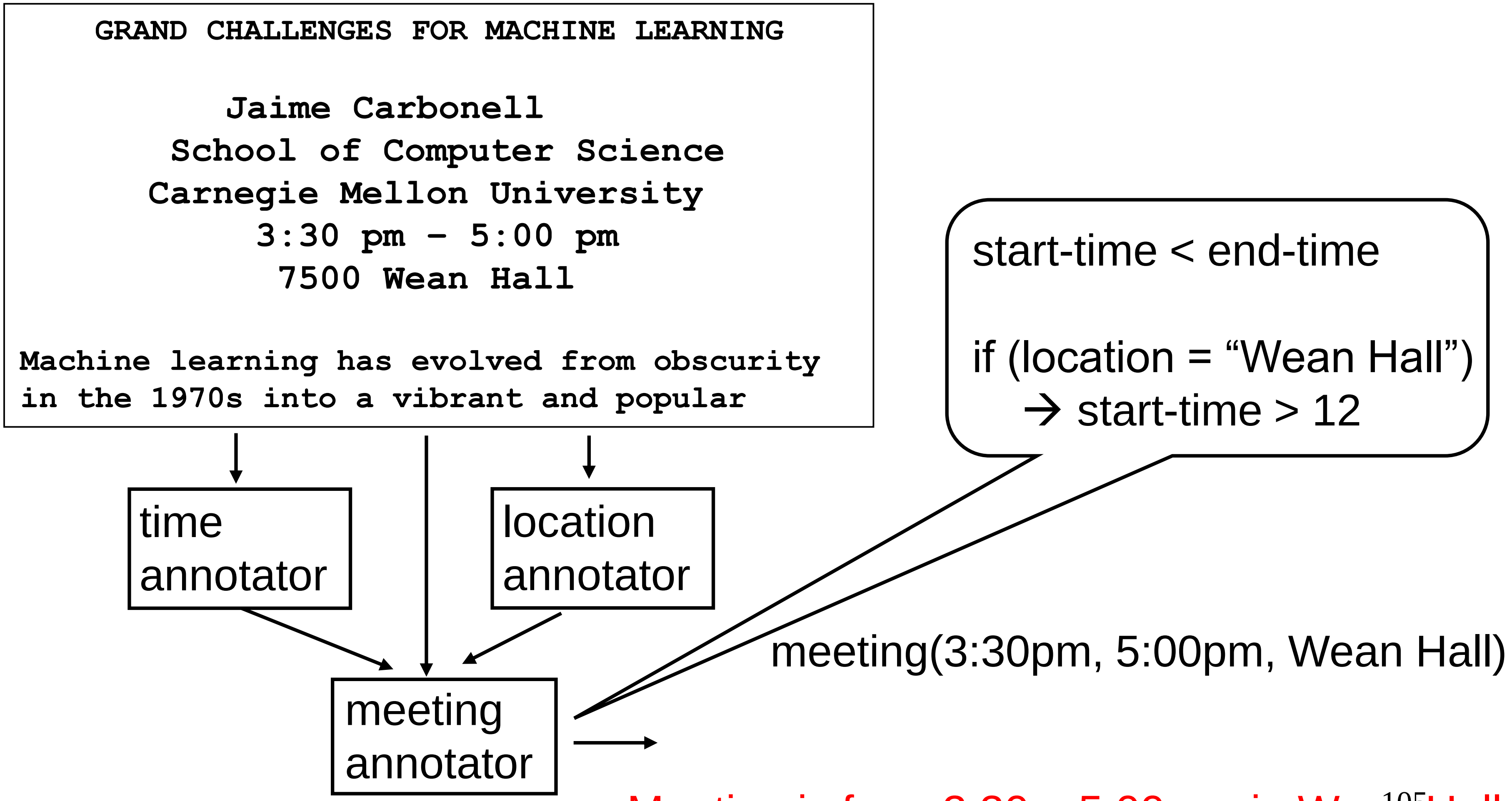


Identifying Person's Phone Numbers from Email

Workflows are often Large and Complex

- In DBLife system
  - between 45 to 90 annotators
  - the workflow is 5 level deep
  - this makes up only half of the DBLife system (this is counting only extraction rules)
- In Avatar
  - 25 to 30 annotators extract a single fact with [SIGIR, 2006]
  - Workflows are 7 level deep

Observation 2: Often Need to Incorporate Domain Constraints



GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell
School of Computer Science
Carnegie Mellon University
3:30 pm – 5:00 pm
7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular

start-time < end-time

if (location = “Wean Hall”)
→ start-time > 12

meeting(3:30pm, 5:00pm, Wean Hall)

Meeting is from 3:30 – 5:00 pm in Wean Hall

Observation 3: The Process is Incremental & Iterative

- During development
  - Multiple versions of the same annotator might need to compared and contrasted before the choosing the right one (e.g., different regular expressions for the same task)
  - Incremental annotator development
- During deployment
  - Constant addition of new annotators; extract new entities, new relations etc.
  - Constant arrival of new documents
  - Many systems are 24/7 (e.g., DBLife)

Observation 4: Scalability is a Major Problem

- DBLife example
  - 120 MB of data / day, running the IE workflow once takes 3-5 hours
  - Even on smaller data sets debugging and testing is a time-consuming process
  - stored data over the past 2 years →magnifies scalability issues
  - write a new domain constraint, now should we rerun system from day one? Would take 3 months.
- AliBaba: query time IE
  - Users expect almost real-time response

These observations lead to
many difficult and important challenges

Efficient Construction of IE Workflow

- What would be the right workflow model ?
  - Help write workflow quickly
  - Helps quickly debug, test, and reuse
  - UIMA / GATE ? (do we need to extend these ?)
- What is a good language to specify a single annotator in this workfow
  - An example of this is CPSL [Appelt, 1998]
  - What are the appropriate list of operators ?
  - Do we need a new data-model ?
  - Help users express domain constraints.

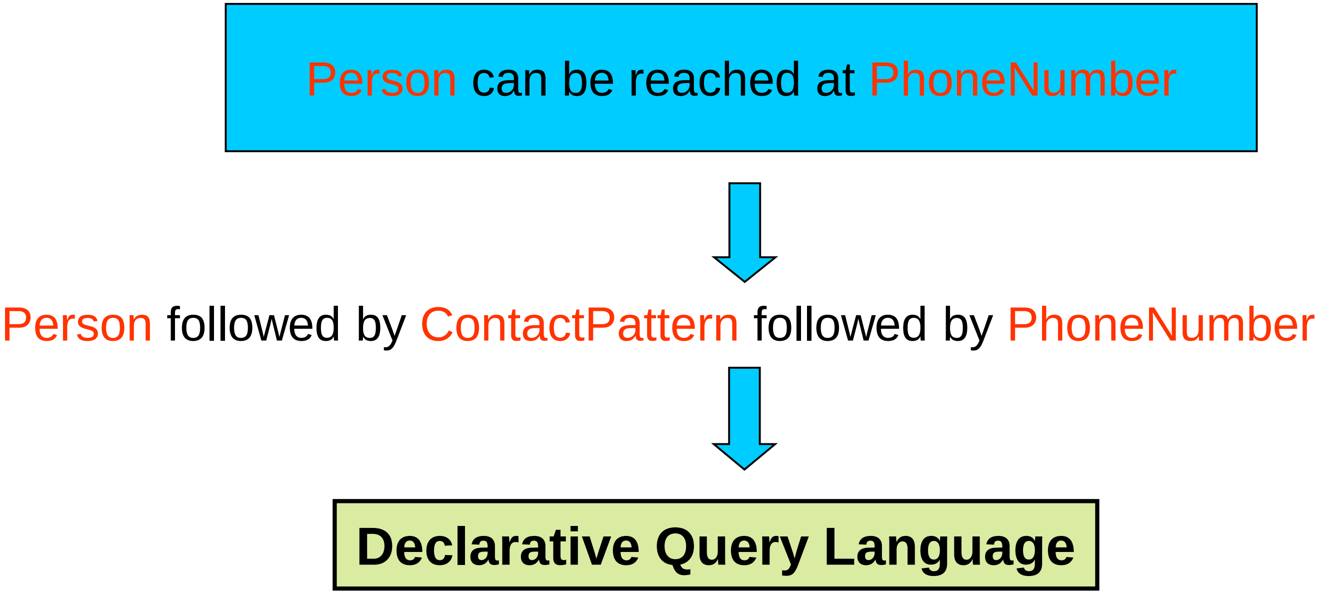
Efficient Compiler for IE Workflows

- What are a good set of “operators” for IE process?
  - Span operations e.g., Precedes, contains etc.
  - Block operations
  - Constraint handler ?
  - Regular expression and dictionary operators
- Efficient implementation of these operators
  - Inverted index constructor? inverted index lookup? [Ramakrishnan, G. et. al, 2006]
- How to compile an efficient execution plan?

Optimizing IE Workflows

- Finding a good execution plan is important !
- Reuse existing annotations
  - E.g., Person's phone number annotator
  - Lower-level operators can ignore documents that do NOT contain Persons and PhoneNumbers → potentially 10-fold speedup in Enron e-mail collection
  - Useful in developing sparse annotators
- Questions ?
  - How to estimate statistics for IE operators?
  - In some cases different execution plans may have different extraction accuracy → not just a matter of optimizing for runtime

Rules as Declarative Queries in Avatar



Person can be reached at PhoneNumber

Person followed by ContactPattern followed by PhoneNumber

Declarative Query Language

```
ContactPattern   ←  RegularExpression(Email.body,”can be reached at”)

PersonPhone      ←   Precedes (
                                      Precedes (Person, ContactPattern, D),
                                      Phone, D)
```

Domain-specific annotator in Avatar

- Identifying people's phone numbers in email

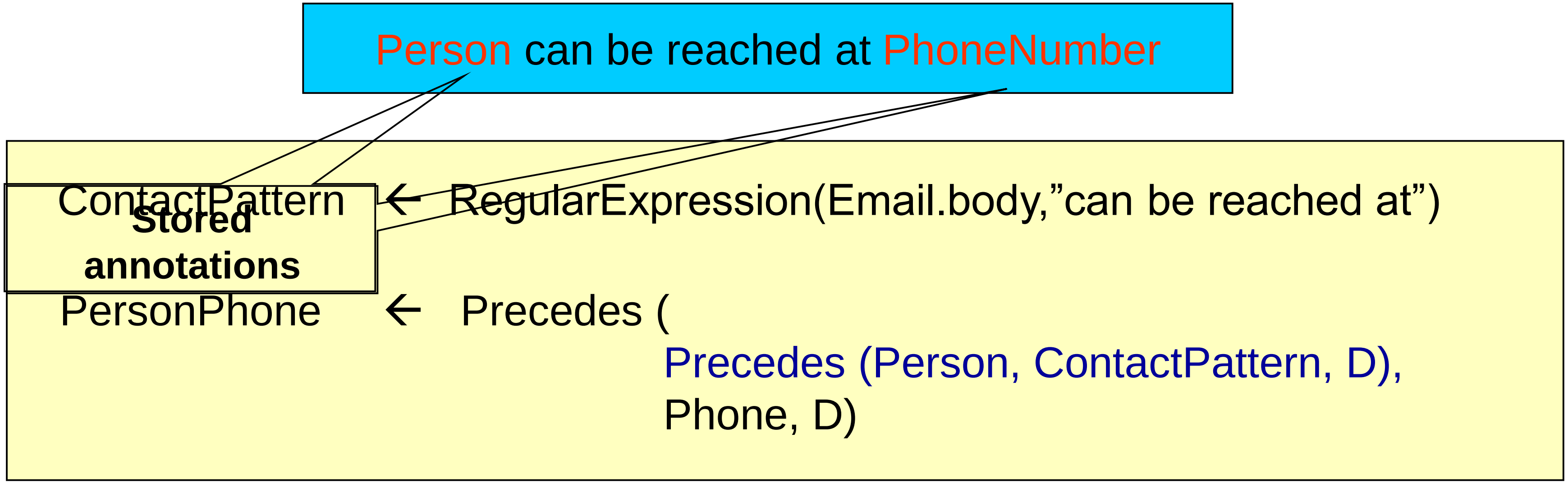
I will be out Thursday, but back on Friday.
Sarah can be reached at 202-466-9160.
Thanks for your help. Christi 37007.

- Generic pattern is

Person can be reached at PhoneNumber

Optimizing IE Workflows in Avatar

- An IE workflow can be compiled into different execution plans
- E.g., two "execution plans" in Avatar:



Person can be reached at PhoneNumber

Stored annotations

ContactPattern ← RegularExpression(Email.body,"can be reached at")

PersonPhone ← Precedes (
Precedes (Person, ContactPattern, D),
Phone, D)

ContactPattern ← RegularExpression(Email.body,"can be reached at")

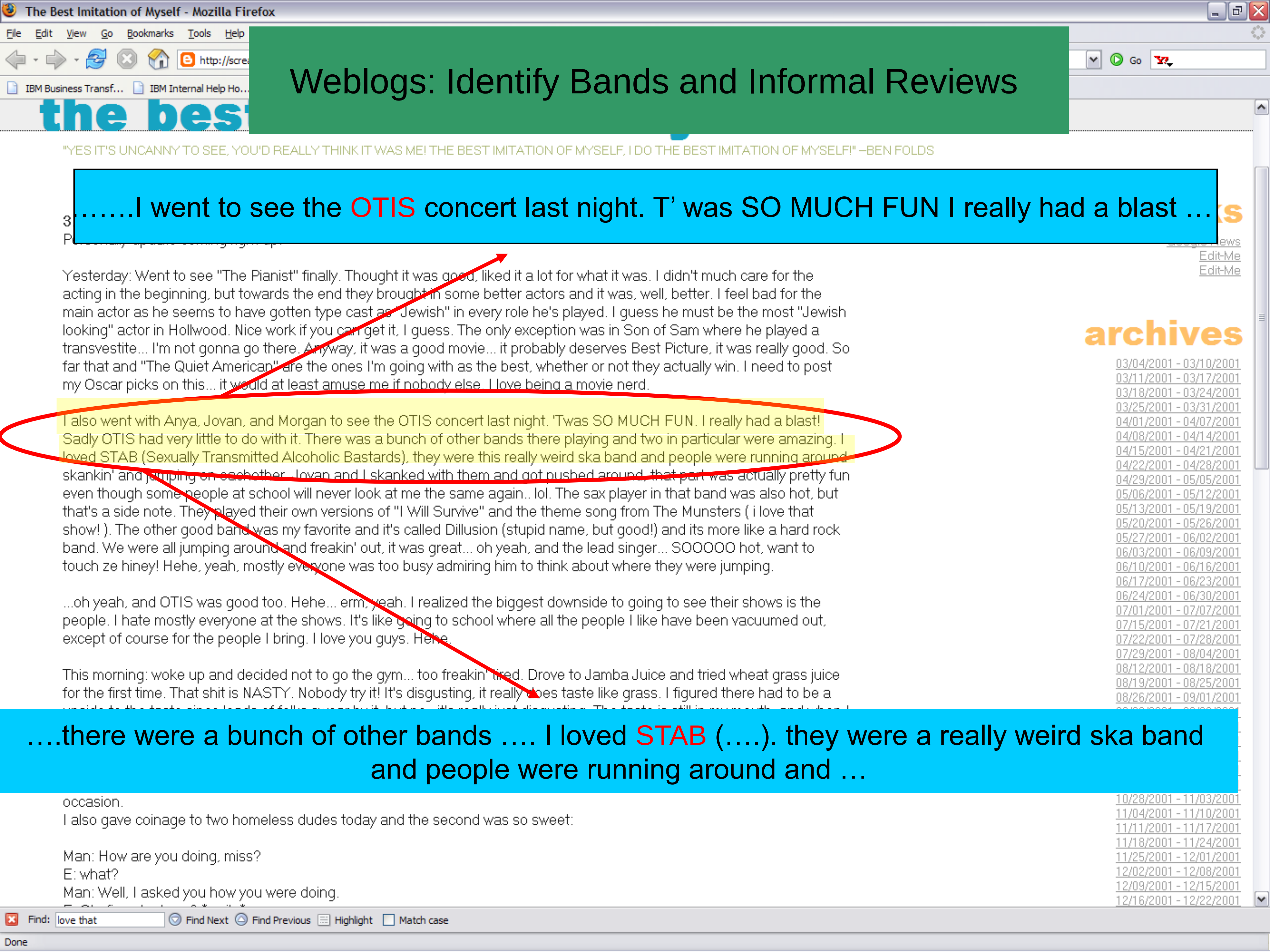
PersonPhone ← Precedes(Person
Precedes(ContactPattern, Phone, D),
D)

Alternative Query in Avatar

```
ContactPattern   ←  RegularExpression(Email.body,”can be reached at”)

PersonPhone      ←   Contains (
                                Precedes (Person, Phone, D),
                                ContactPattern)
```

Weblogs: Identify Bands and Informal Reviews



……I went to see the OTIS concert last night. T' was SO MUCH FUN I really had a blast …

"YES IT'S UNCANNY TO SEE, YOU'D REALLY THINK IT WAS ME! THE BEST IMITATION OF MYSELF, I DO THE BEST IMITATION OF MYSELF!" –BEN FOLDS

Yesterday: Went to see "The Pianist" finally. Thought it was good, liked it a lot for what it was. I didn't much care for the acting in the beginning, but towards the end they brought in some better actors and it was, well, better. I feel bad for the main actor as he seems to have gotten type cast as "Jewish" in every role he's played. I guess he must be the most "Jewish looking" actor in Hollwood. Nice work if you can get it, I guess. The only exception was in Son of Sam where he played a transvestite... I'm not gonna go there. Anyway, it was a good movie... it probably deserves Best Picture, it was really good. So far that and "The Quiet American" are the ones I'm going with as the best, whether or not they actually win. I need to post my Oscar picks on this... it would at least amuse me if nobody else. I love being a movie nerd.

I also went with Anya, Jovan, and Morgan to see the OTIS concert last night. 'Twas SO MUCH FUN. I really had a blast! Sadly OTIS had very little to do with it. There was a bunch of other bands there playing and two in particular were amazing. I loved STAB (Sexually Transmitted Alcoholic Bastards), they were this really weird ska band and people were running around skankin' and jumping on eachother. Jovan and I skanked with them and got pushed around, that part was actually pretty fun even though some people at school will never look at me the same again.. lol. The sax player in that band was also hot, but that's a side note. They played their own versions of "I Will Survive" and the theme song from The Munsters (i love that show!). The other good band was my favorite and it's called Dillusion (stupid name, but good!) and its more like a hard rock band. We were all jumping around and freakin' out, it was great... oh yeah, and the lead singer... SOOOOO hot, want to touch ze hiney! Hehe, yeah, mostly everyone was too busy admiring him to think about where they were jumping.

...oh yeah, and OTIS was good too. Hehe... erm, yeah. I realized the biggest downside to going to see their shows is the people. I hate mostly everyone at the shows. It's like going to school where all the people I like have been vacuumed out, except of course for the people I bring. I love you guys. Hehe.

This morning: woke up and decided not to go the gym... too freakin' tired. Drove to Jamba Juice and tried wheat grass juice for the first time. That shit is NASTY. Nobody try it! It's disgusting, it really does taste like grass. I figured there had to be a

…there were a bunch of other bands …. I loved STAB (….). they were a really weird ska band and people were running around and …

occasion.
I also gave coinage to two homeless dudes today and the second was so sweet:

Man: How are you doing, miss?
E: what?
Man: Well, I asked you how you were doing.

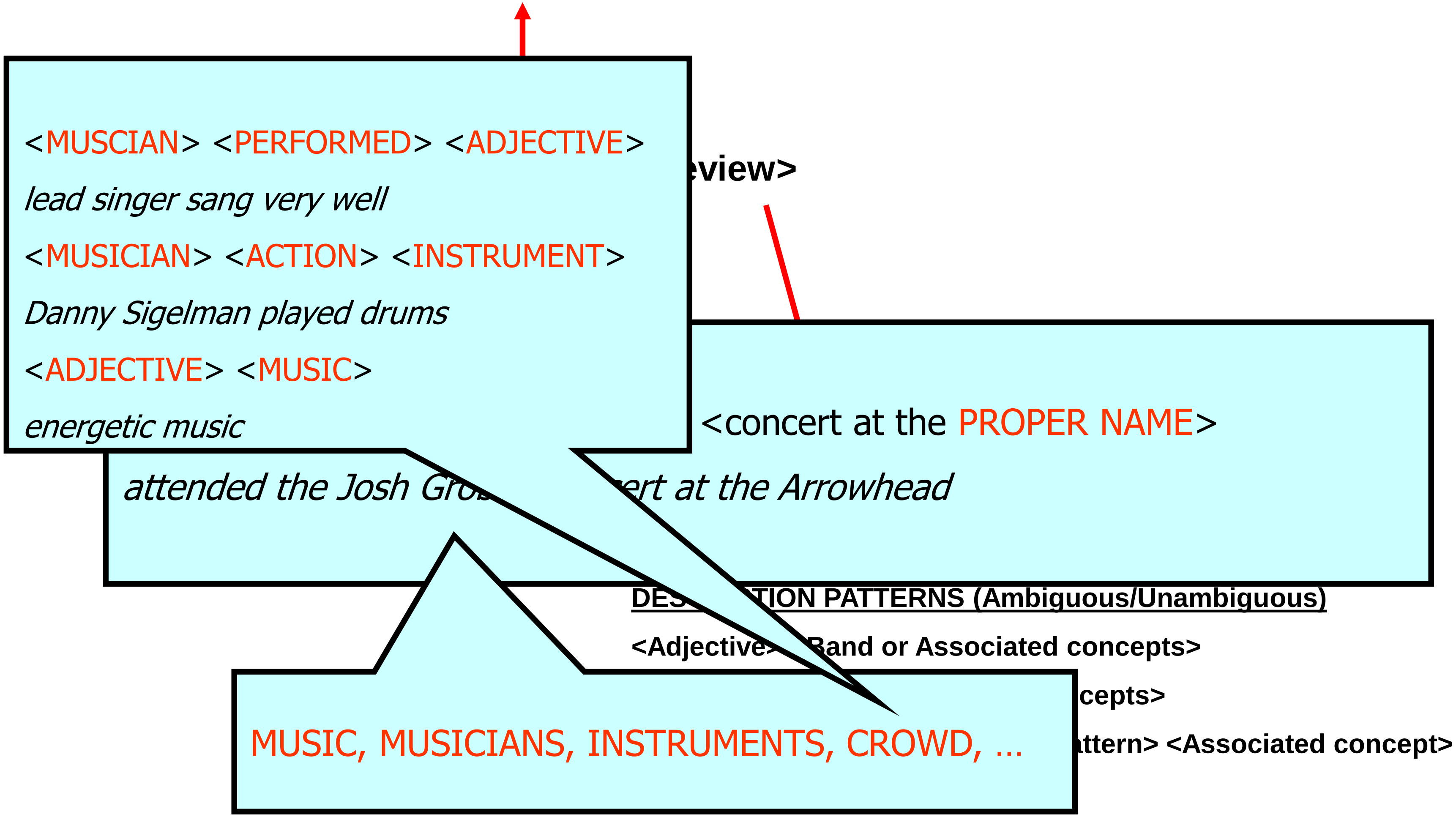
Edit-Me
Edit-Me

archives

03/04/2001 - 03/10/2001
03/11/2001 - 03/17/2001
03/18/2001 - 03/24/2001
03/25/2001 - 03/31/2001
04/01/2001 - 04/07/2001
04/08/2001 - 04/14/2001
04/15/2001 - 04/21/2001
04/22/2001 - 04/28/2001
04/29/2001 - 05/05/2001
05/06/2001 - 05/12/2001
05/13/2001 - 05/19/2001
05/20/2001 - 05/26/2001
05/27/2001 - 06/02/2001
06/03/2001 - 06/09/2001
06/10/2001 - 06/16/2001
06/17/2001 - 06/23/2001
06/24/2001 - 06/30/2001
07/01/2001 - 07/07/2001
07/15/2001 - 07/21/2001
07/22/2001 - 07/28/2001
07/29/2001 - 08/04/2001
08/12/2001 - 08/18/2001
08/19/2001 - 08/25/2001
08/26/2001 - 09/01/2001
10/28/2001 - 11/03/2001
11/04/2001 - 11/10/2001
11/11/2001 - 11/17/2001
11/18/2001 - 11/24/2001
11/25/2001 - 12/01/2001
12/02/2001 - 12/08/2001
12/09/2001 - 12/15/2001
12/16/2001 - 12/22/2001

Band INSTANCE PATTERNS

<Leading pattern> <Band instance> <Trailing pattern>



<MUSCIAN> <PERFORMED> <ADJECTIVE>

lead singer sang very well

<MUSICIAN> <ACTION> <INSTRUMENT>

Danny Sigelman played drums

<ADJECTIVE> <MUSIC>

energetic music

eview>

<concert at the PROPER NAME>

attended the Josh Gro ert at the Arrowhead

DES TION PATTERNS (Ambiguous/Unambiguous)

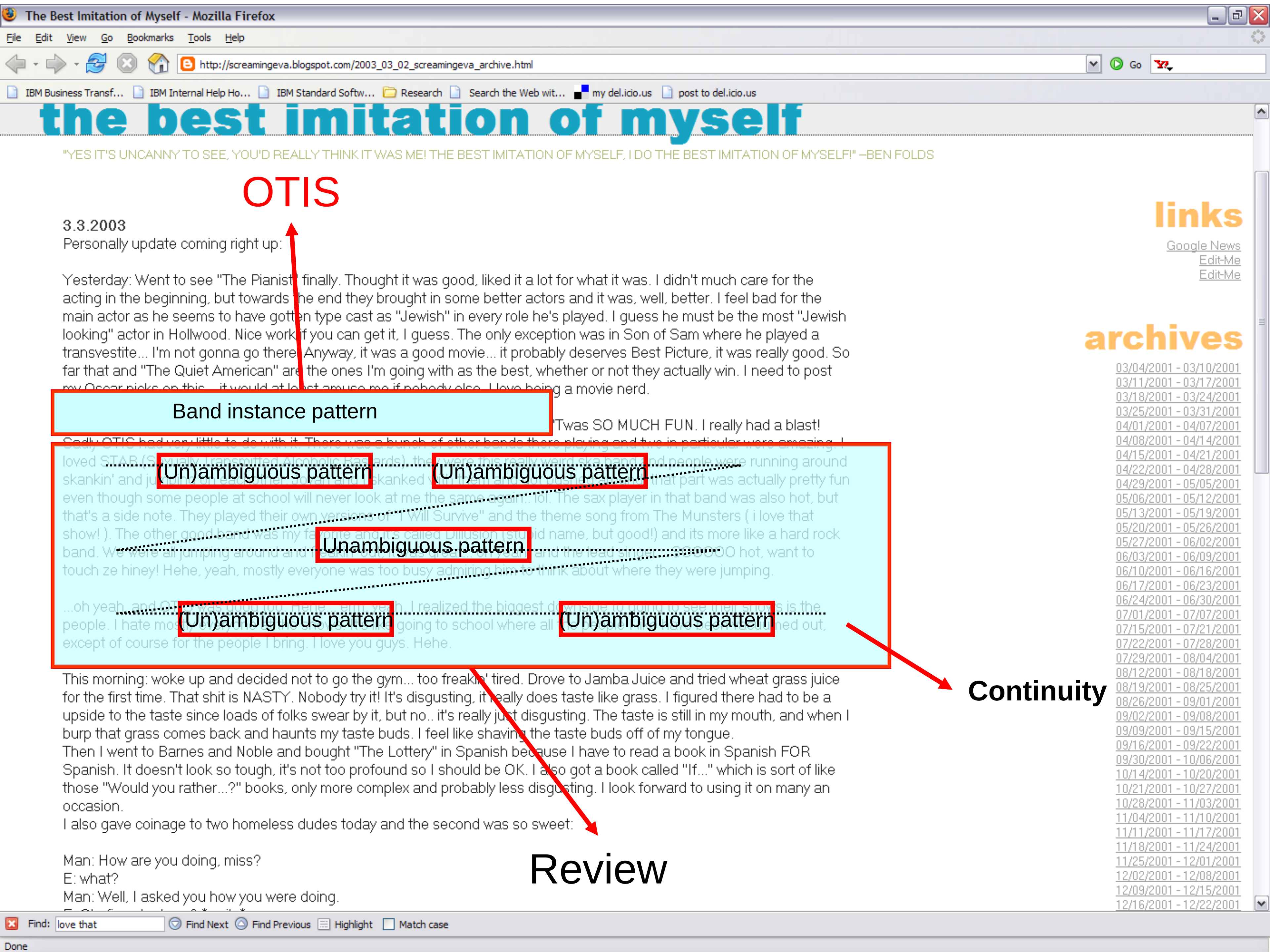
<Adjective> Band or Associated concepts>

cepts>

ttern> <Associated concept>

MUSIC, MUSICIANS, INSTRUMENTS, CROWD, …

Real challenge is in optimizing such complex workflows !!



The Best Imitation of Myself - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://screamingeva.blogspot.com/2003_03_02_screamingeva_archive.html
Go
IBM Business Transf... IBM Internal Help Ho... IBM Standard Softw... Research Search the Web wit... my del.icio.us post to del.icio.us

the best imitation of myself

"YES IT'S UNCANNY TO SEE, YOU'D REALLY THINK IT WAS ME! THE BEST IMITATION OF MYSELF, I DO THE BEST IMITATION OF MYSELF!" –BEN FOLDS

3.3.2003
Personally update coming right up:

Yesterday: Went to see "The Pianist" finally. Thought it was good, liked it a lot for what it was. I didn't much care for the acting in the beginning, but towards the end they brought in some better actors and it was, well, better. I feel bad for the main actor as he seems to have gotten type cast as "Jewish" in every role he's played. I guess he must be the most "Jewish looking" actor in Hollwood. Nice work if you can get it, I guess. The only exception was in Son of Sam where he played a transvestite... I'm not gonna go there. Anyway, it was a good movie... it probably deserves Best Picture, it was really good. So far that and "The Quiet American" are the ones I'm going with as the best, whether or not they actually win. I need to post my Oscar picks on this... it would at least amuse me if nobody else. I love being a movie nerd.

[illegible] 'Twas SO MUCH FUN. I really had a blast! Sadly OTIS had very little to do with it. There was a bunch of other bands there playing and two in particular were amazing. I loved STAR (S[illegible] ards), they were this really weird ska band and people were running around skankin' and ju[illegible] skanked w[illegible] that part was actually pretty fun even though some people at school will never look at me the same again.. lol. The sax player in that band was also hot, but that's a side note. They played their own versions of "I Will Survive" and the theme song from The Munsters (i love that show!). The other good band was my fa[illegible] id name, but good!) and its more like a hard rock band. We were all jumping around and [illegible] and the lead singer... SOOOOO hot, want to touch ze hiney! Hehe, yeah, mostly everyone was too busy admiring him to think about where they were jumping.

...oh yeah, and O[illegible] h. I realized the biggest d[illegible] s is the people. I hate mo[illegible] going to school where all [illegible] ned out, except of course for the people I bring. I love you guys. Hehe.

This morning: woke up and decided not to go the gym... too freakin' tired. Drove to Jamba Juice and tried wheat grass juice for the first time. That shit is NASTY. Nobody try it! It's disgusting, it really does taste like grass. I figured there had to be a upside to the taste since loads of folks swear by it, but no.. it's really just disgusting. The taste is still in my mouth, and when I burp that grass comes back and haunts my taste buds. I feel like shaving the taste buds off of my tongue.
Then I went to Barnes and Noble and bought "The Lottery" in Spanish because I have to read a book in Spanish FOR Spanish. It doesn't look so tough, it's not too profound so I should be OK. I also got a book called "If..." which is sort of like those "Would you rather...?" books, only more complex and probably less disgusting. I look forward to using it on many an occasion.
I also gave coinage to two homeless dudes today and the second was so sweet:

Man: How are you doing, miss?
E: what?
Man: Well, I asked you how you were doing.

links

Google News
Edit-Me
Edit-Me

archives

03/04/2001 - 03/10/2001
03/11/2001 - 03/17/2001
03/18/2001 - 03/24/2001
03/25/2001 - 03/31/2001
04/01/2001 - 04/07/2001
04/08/2001 - 04/14/2001
04/15/2001 - 04/21/2001
04/22/2001 - 04/28/2001
04/29/2001 - 05/05/2001
05/06/2001 - 05/12/2001
05/13/2001 - 05/19/2001
05/20/2001 - 05/26/2001
05/27/2001 - 06/02/2001
06/03/2001 - 06/09/2001
06/10/2001 - 06/16/2001
06/17/2001 - 06/23/2001
06/24/2001 - 06/30/2001
07/01/2001 - 07/07/2001
07/15/2001 - 07/21/2001
07/22/2001 - 07/28/2001
07/29/2001 - 08/04/2001
08/12/2001 - 08/18/2001
08/19/2001 - 08/25/2001
08/26/2001 - 09/01/2001
09/02/2001 - 09/08/2001
09/09/2001 - 09/15/2001
09/16/2001 - 09/22/2001
09/30/2001 - 10/06/2001
10/14/2001 - 10/20/2001
10/21/2001 - 10/27/2001
10/28/2001 - 11/03/2001
11/04/2001 - 11/10/2001
11/11/2001 - 11/17/2001
11/18/2001 - 11/24/2001
11/25/2001 - 12/01/2001
12/02/2001 - 12/08/2001
12/09/2001 - 12/15/2001
12/16/2001 - 12/22/2001

Tutorial Roadmap

- Introduction to managing IE [RR]
  - Motivation
  - What’s different about *managing* IE?
- Major research directions
  - Extracting mentions of entities and relationships [SV]
    - Uncertainty management
  - Disambiguating extracted mentions [AD]
    - Tracking mentions and entities over time
  - Understanding, correcting, and maintaining extracted data [AD]
    - Provenance and explanations
    - Incorporating user feedback

Uncertainty Management

Uncertainty During Extraction Process

- **Annotators make mistakes !**
- **Annotators provide confidence scores with each annotation**
- **Simple named-entity annotator**

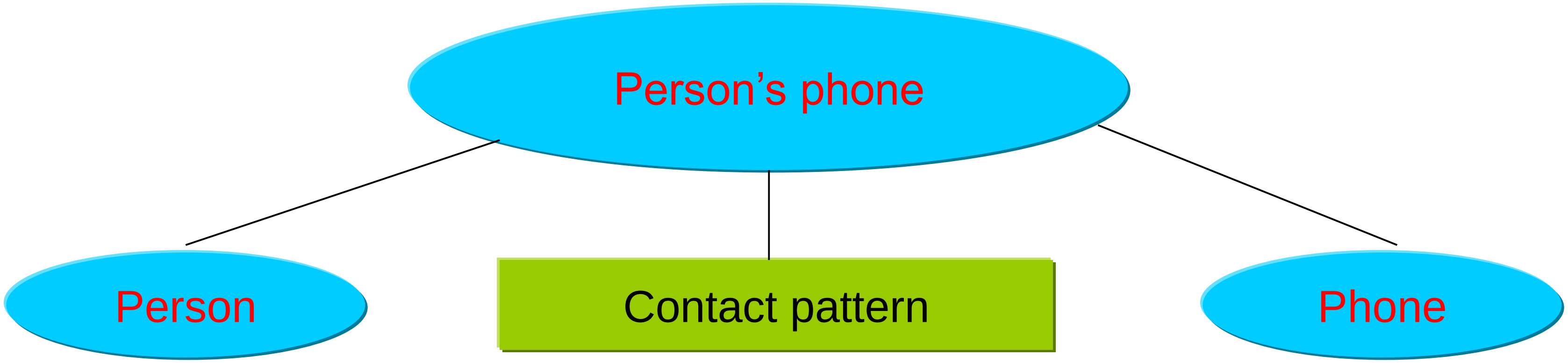
C = Word with first letter capitalized
D = Matches an entry in a person name dictionary

Annotator Rules		**Precision**
1.	[CD] [CD]	0.9
2.	[CD]	0.6

Last evening I met the candidate Shiv Vaithyanathan for dinner. We had an interesting conversation and I encourage you to get an update. His host Bill can be reached at X-2465.

Text-mention	Probability	
Shiv Vaithyanathan	0.9	[CD] [CD]
Bill	0.6	[CD]

Composite Annotators [Jayram et al, 2006]



Person's phone

Person

Contact pattern

Phone

Person can be reached at PhoneNumber

- **Question:** How do we compute probabilities for the output of composite annotators from base annotators ?

With Two Annotators

Person Table

ID	Text-mention	
1	Shiv Vaithyanathan	0.9
2	Bill	0.6

Telephone Table

ID	Text-mention	
1	(408)-927-2465	0.95
2	X-2465	0.3

These annotations are kept in separate tables

Problem at Hand

Last evening I met the candidate Shiv Vaithyanathan for dinner. We had an interesting conversation and I encourage you to get an update. His host Bill can be reached at X-2465.

Person can be reached at PhoneNumber

Person Table

ID	Text-mention	
1	**Shiv Vaithyanathan**	0.9
2	**Bill**	0.6

Telephone Table

ID	Text-mention	
1	**(408)-927-2465**	0.95
2	**X-2465**	0.3

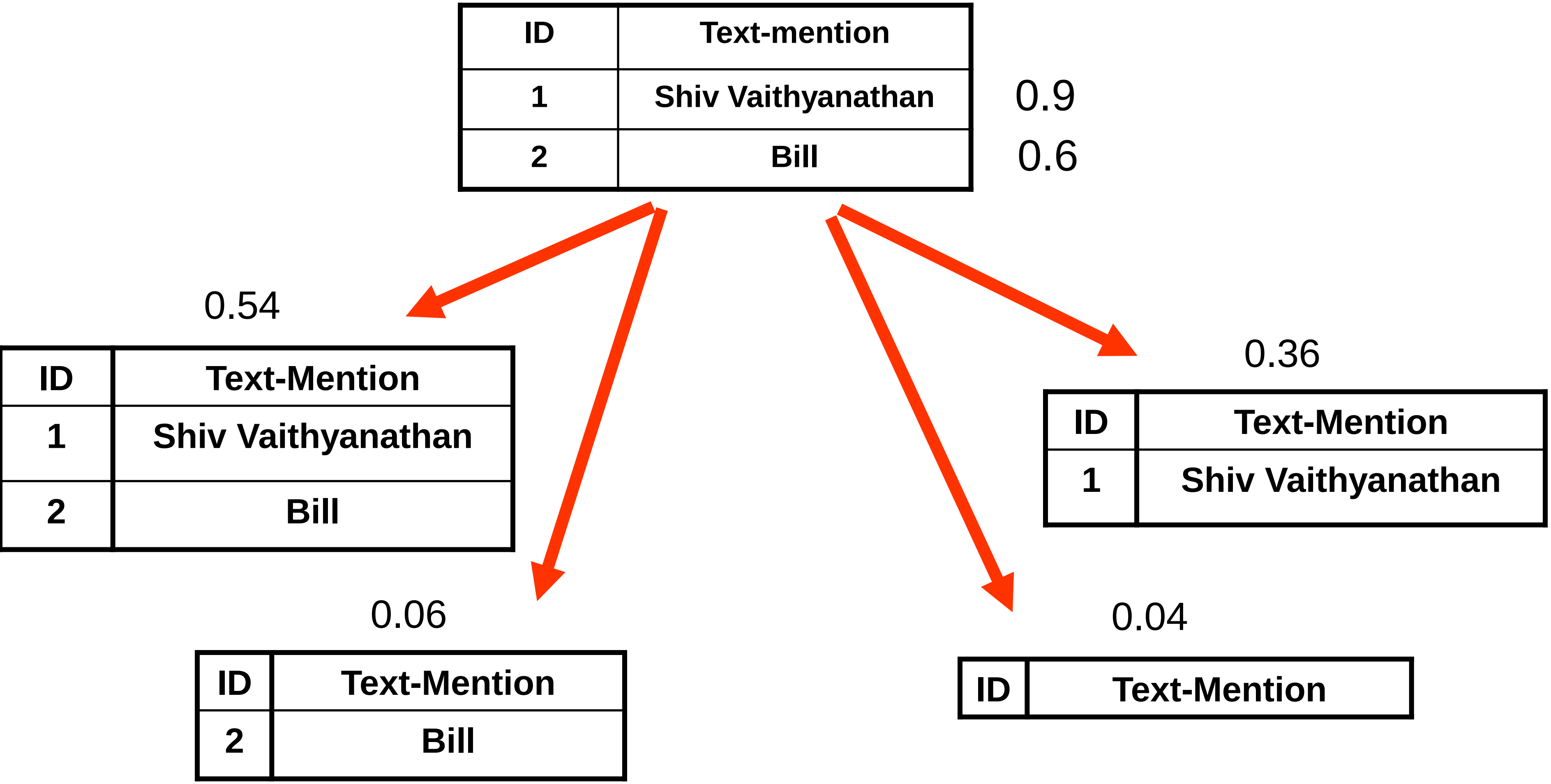
ID	Person	Telephone
1	Bill	X-2465

?

What is the probability ?

One Potential Approach: Possible Worlds [Dalvi-Suciu, 2004]

Person example



ID	Text-mention
1	Shiv Vaithyanathan
2	Bill

0.9
0.6

0.54

ID	Text-Mention
1	Shiv Vaithyanathan
2	Bill

0.36

ID	Text-Mention
1	Shiv Vaithyanathan

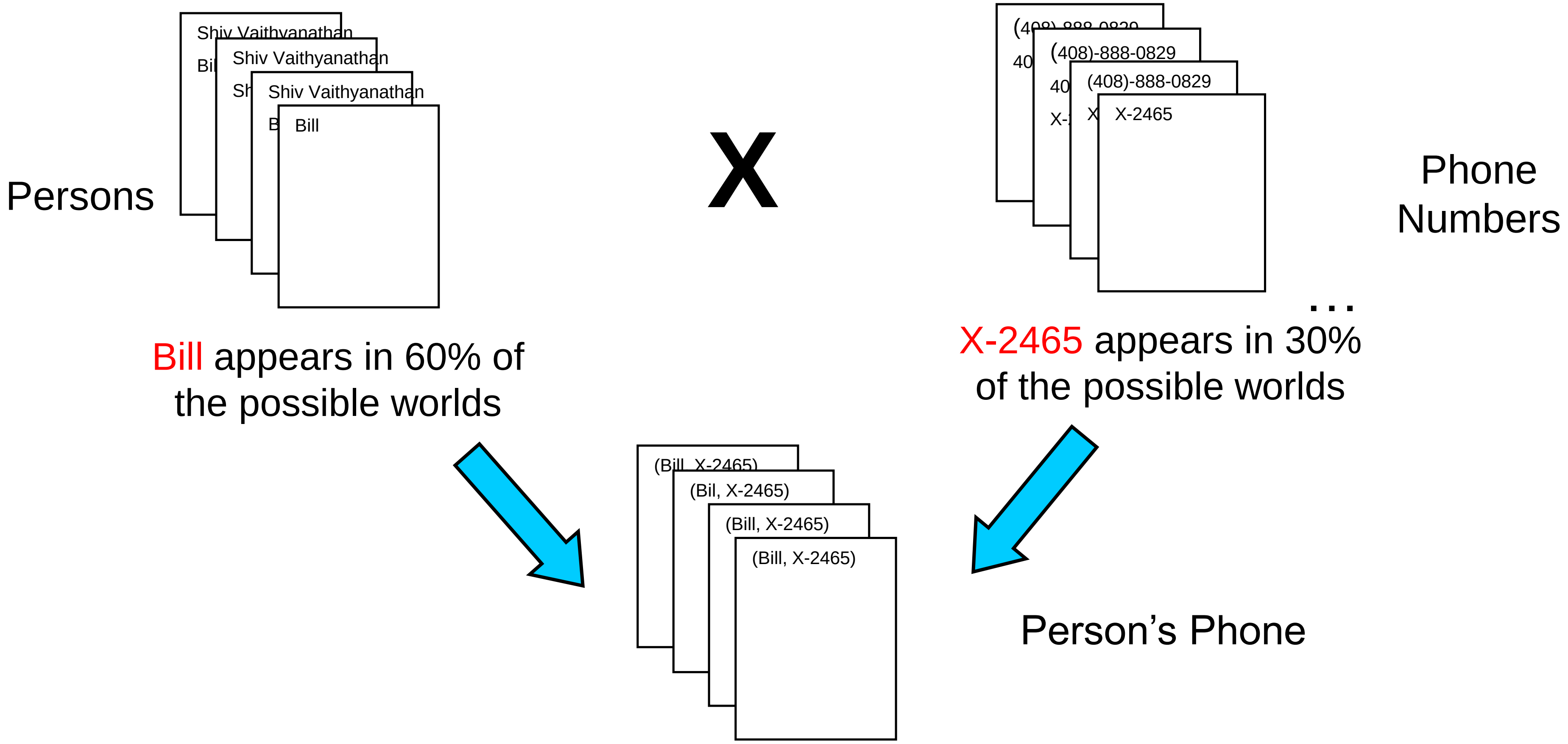
0.06

ID	Text-Mention
2	Bill

0.04

ID	Text-Mention

Possible Worlds Interpretation [Dalvi-Suciu, 2004]



Bill appears in 60% of the possible worlds

X-2465 appears in 30% of the possible worlds

(Bill, X-2465) appears in at most 18% of the possible worlds

Annotation (Bill, X-2465) can have a probability of at most 0.18

But Real Data Says Otherwise …. [Jayram et al, 2006]

- With Enron collection using Person instances with a low probability the following rule

  Person can be reached at PhoneNumber

  produces annotations that are correct more than 80% of the time
- Relaxing independence constraints [Fuhr-Roelleke, 95] does not help since X-2465 appears in only 30% of the worlds

More powerful probabilistic database constructs are needed to capture the dependencies present in the Rule above !

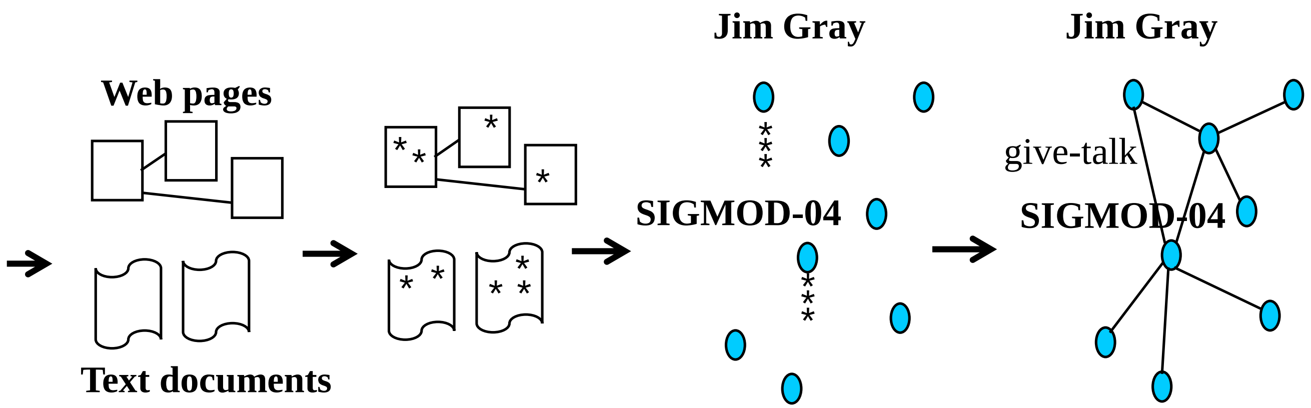
Databases and Probability

- Probabilistic DB
  - Fuhr [F&R97, F95] : uses events to describe possible worlds
  - [Dalvi&Suciu04] : query evaluation assuming independence of tuples
  - Trio System [Wid05, Das06] : distinguishes between data lineage and its probability
- Relational Learning
  - Bayesian Networks, Markov models: assumes tuples are independently and identically distributed
  - Probabilistic Relational Models [Koller+99]: accounts for correlations between tuples
- Uncertainty in Knowledge Bases
  - [GHK92, BGHK96] generating possible worlds probability distribution from statistics
  - [BGHK94] updating probability distribution based on new knowledge
- Recent work
  - MauveDB [D&M 2006], Gupta & Sarawagi [G&S, 2006]

Disambiguate, aka match, extracted mentions

Once mentions have been extracted, matching them is the next step

Researcher
Homepages
Conference
Pages
Group Pages
DBworld
mailing list
DBLP



Keyword search
SQL querying
Question answering
Browse
Mining
Alert/Monitor
News summary

Mention Matching: Problem Definition

- Given extracted mentions $M = \{m_1, ..., m_n\}$
- Partition M into groups $M_1, ..., M_k$
  - All mentions in each group refer to the same real-world entity

- Variants are known as
  - Entity matching, record deduplication, record linkage, entity resolution, reference reconciliation, entity integration, fuzzy duplicate elimination

Another Example

Document 1: *The Justice Department has officially ended its inquiry into the assassinations of* **John F. Kennedy** *and Martin Luther King Jr., finding ``no persuasive evidence'' to support conspiracy theories, according to department documents. The House Assassinations Committee concluded in 1978 that* **Kennedy** *was ``probably'' assassinated as the result of a conspiracy involving a second gunman, a finding that broke from the* ***Warren Commission*** *'s belief that Lee Harvey Oswald acted alone in* ***Dallas*** *on Nov. 22, 1963.*

Document 2: *In 1953, Massachusetts* **Sen. John F. Kennedy** *married Jacqueline Lee Bouvier in Newport, R.I. In 1960, Democratic presidential candidate* **John F. Kennedy** *confronted the issue of his Roman Catholic faith by telling a Protestant group in Houston, ``I do not speak for my church on public matters, and the church does not speak for me.''*

Document 3: David Kennedy *was born in Leicester, England in 1959. …***Kennedy** *co-edited The New Poetry (Bloodaxe Books 1993), and is the author of New Relations: The Refashioning Of British Poetry 1980-1994 (Seren 1996).*

[From Li, Morie, & Roth, AI Magazine, 2005]

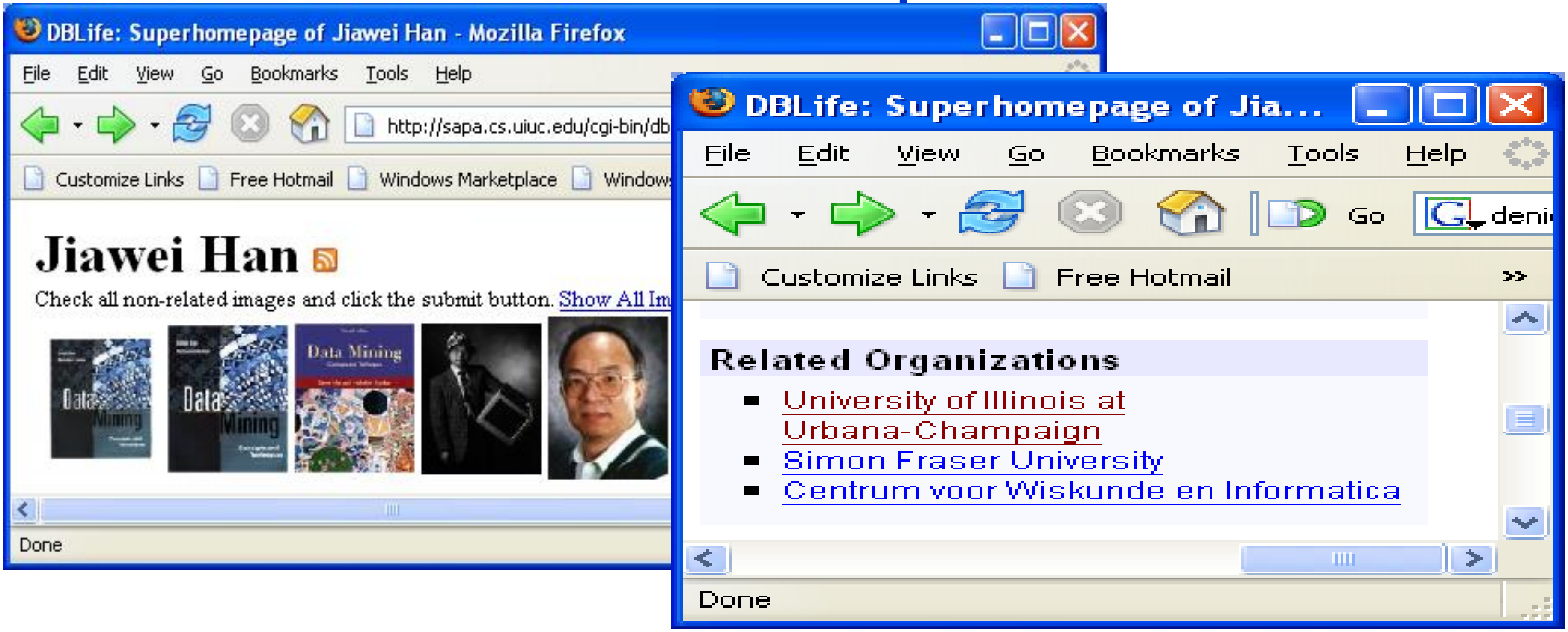
Extremely Important Problem!

- Appears in numerous real-world contexts
- Plagues many applications that we have seen
  - Citeseer, DBLife, AliBaba, Rexa, etc.

Why so important?

- Many useful services rely on mention matching being right
- If we do not match mentions with sufficient accuracy ➔ errors cascade, greatly reducing the usefulness of these services

An Example


DBLife: Superhomepage of Jiawei Han - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://sapa.cs.uiuc.edu/cgi-bin/db
Customize Links
Free Hotmail
Windows Marketplace
Window
Done
DBLife: Superhomepage of Jia...
File Edit View Go Bookmarks Tools Help
Go
deni
Customize Links
Free Hotmail
Done

Jiawei Han

Check all non-related images and click the submit button. Show All Im

Related Organizations

- University of Illinois at Urbana-Champaign
- Simon Fraser University
- Centrum voor Wiskunde en Informatica

Discover related organizations using occurrence analysis:

“J. Han ... Centrum voor Wiskunde en Informatica”

DBLife incorrectly matches this mention “J. Han” with “Jiawei Han”, but it actually refers to “Jianchao Han”.

The Rest of This Section

- To set the stage, briefly review current solutions to mention matching / record linkage
  - **a comprehensive tutorial is provided tomorrow Wed 2-5:30pm, by Nick Koudas, Sunita Sarawagi, & Divesh Srivastava**
- Then focus on novel challenges brought forth by IE over text
  - developing matching workflow, optimizing workflow, incorporating domain knowledge
  - tracking mentions / entities, detecting interesting events

A First Matching Solution: String Matching

m_{11} = “John F. Kennedy”
m_{12} = “Kennedy”

m_{21} = “Senator John F. Kennedy”
m_{22} = “John F. Kennedy”

m_{31} = “David Kennedy”
m_{32} = “Kennedy”

$sim(m_i,m_j) > 0.8$ ➔ m_i and m_j match.

sim = edit distance, q-gram, TF/IDF, etc.

- A recent survey:
  - **Adaptive Name Matching in Information Integration, by M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, & S. Fienberg, *IEEE Intelligent Systems*, 2003.**
  - **Other recent work: [Koudas, Marathe, Srivastava, VLDB-04]**
- Pros & cons
  - conceptually simple, relatively fast
  - often insufficient for achieving high accuracy

A More Common Solution

- For each mention m, extract additional data
  - transform m into a record
- Match the records
  - leveraging the wealth of existing record matching solutions

Document 3: **David Kennedy** *was born in Leicester, England in 1959. …* **Kennedy** *co-edited The New Poetry (Bloodaxe Books 1993), and is the author of New Relations: The Refashioning Of British Poetry 1980-1994 (Seren 1996).*

first-name	last-name	birth-date	birth-place
David	Kennedy	1959	Leicester
D.	Kennedy	1959	England

Two main groups of
record matching solutions

- hand-crafted rules
- learning-based

which we will discuss next

Hand-Crafted Rules

If R_1.last-name = R_2.last-name
R_1.first-name ~ R_2.first-name
R_1.address ~ R_2.address
➔ R_1 matches R_2

[Hernandez & Stolfo, SIGMOD-95]

$$\begin{aligned} \mathrm{sim}(R_1,R_2) = \; & \mathrm{alpha}_1 * \mathrm{sim}_1(R_1.\text{last-name},R_2.\text{last-name}) + \\ & \mathrm{alpha}_2 * \mathrm{sim}_2(R_1.\text{first-name},R_2.\text{first-name}) + \\ & \mathrm{alpha}_3 * \mathrm{sim}_3(R_1.\text{address}, R_2.\text{address}) \end{aligned}$$

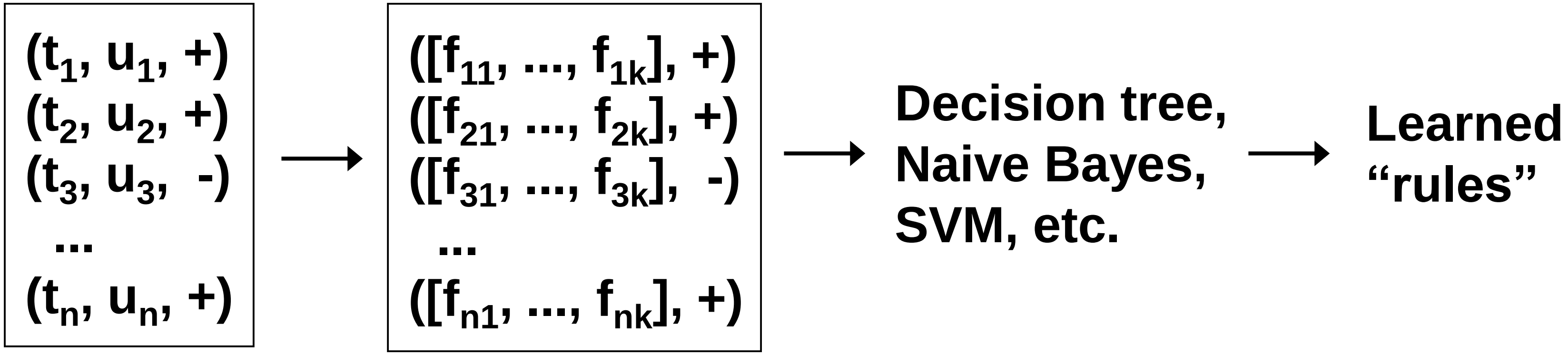
If $\mathrm{sim}(R_1,R_2) > 0.7$ ➔ match

- Pros and cons
  - relatively easy to craft rules in many cases
  - easy to modify, incorporate domain knowledge
  - laborious tuning
  - in certain cases may be hard to create rules manually

Learning-Based Approaches

$(t_1, u_1, +)$
$(t_2, u_2, +)$
$(t_3, u_3, -)$
...
$(t_n, u_n, +)$

- Learn matching rules from training data
- Create a set of features: $f_1, \ldots, f_k$
  - each feature is a function over (t,u)
  - e.g., t.last-name = u.last-name?
    edit-distance(t.first-name,u.first-name)
- Convert each tuple pair to a feature vector, then apply a machine learning algorithm



$(t_1, u_1, +)$
$(t_2, u_2, +)$
$(t_3, u_3, -)$
...
$(t_n, u_n, +)$

→

$([f_{11}, \ldots, f_{1k}], +)$
$([f_{21}, \ldots, f_{2k}], +)$
$([f_{31}, \ldots, f_{3k}], -)$
...
$([f_{n1}, \ldots, f_{nk}], +)$

→

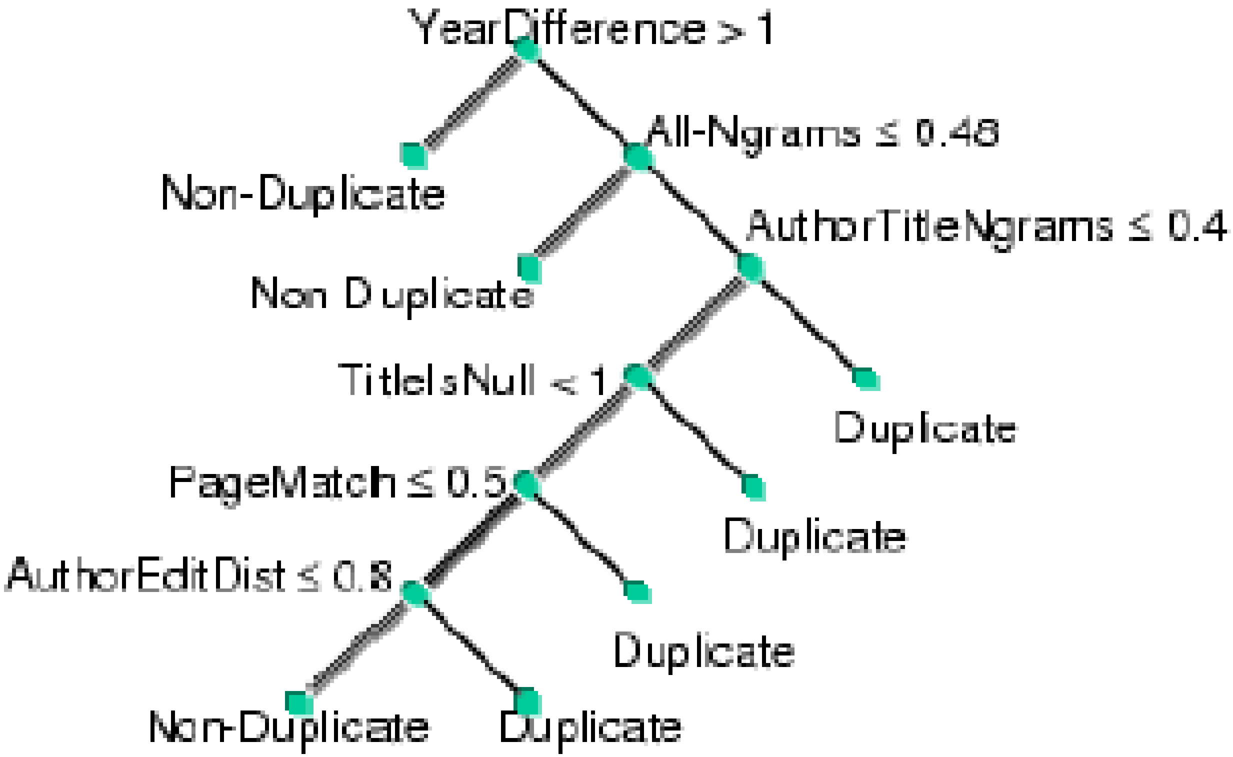
Decision tree,
Naive Bayes,
SVM, etc.

→

Learned
“rules”

Example of Learned Matching Rules

- Produced by a decision-tree learner, to match paper citations



[Sarawagi & Bhamidipaty, KDD-02]

Twists on the Basic Methods

- Compute transitive closures
  - [Hernandez & Stolfo, SIGMOD-95]
- Learn all sorts of other thing (not just matching rules)
  - e.g., transformation rules [Tejada, Knoblock, & Minton, KDD-02]
- Ask users to label selected tuple pairs (active learning)
  - [Sarawagi & Bhamidipaty, KDD-02]
- Can we leverage relational database?
  - [Gravano et. al., VLDB-01]

Twists on the Basic Methods

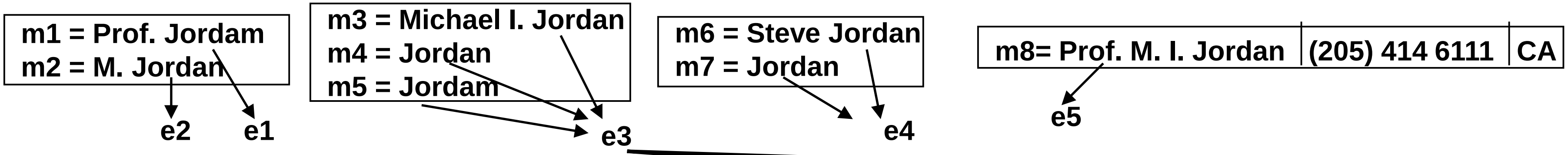
- Record matching in data warehouse contexts
  - Tuples can share values for subsets of attributes
  - [Ananthakrishna, Chaudhuri, & Ganti, VLDB-02]
- Combine mention extraction and matching
  - [Wellner et. al., UAI-04]
- And many more
  - e.g., [Jin, Li, Mehrotra, DASFAA-03]
  - TAILOR record linkage project at Purdue
    [Elfeky, Elmagarmid, Verykios]

Collective Mention Matching: A Recent Trend

- Prior solutions
  - assume tuples are immutable (can’t be changed)
  - often match tuples of just one type
- Observations
  - can enrich tuples along the way → improve accuracy
  - often must match tuples of interrelated types → can leverage matching one type to improve accuracy of matching other types
- This leads to a flurry of recent work on collective mention matching
  - which builds upon the previous three solution groups
- Will illustrate enriching tuples
  - Using [Li, Morie, & Roth, AAAI-04]

Example of Collective Mention Matching

1. Use a simple matching measure to cluster mentions in each document. Each cluster → an entity. Then learn a “profile” for each entity.


m1 = Prof. Jordam
m2 = M. Jordan

m3 = Michael I. Jordan
m4 = Jordan
m5 = Jordam

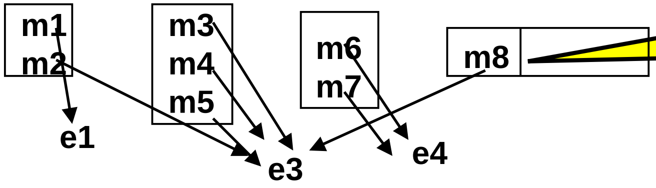
m6 = Steve Jordan
m7 = Jordan

m8= Prof. M. I. Jordan | (205) 414 6111 | CA

e2 e1 e3 e4 e5

first name = Michael, last name = Jordan, middle name = I, can be misspelled as Jordam

2. Reassign each mention to the best matching entity.


m1
m2

m3
m4
m5

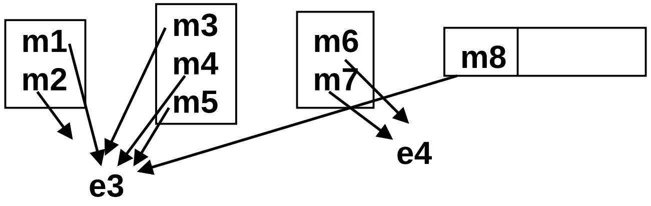
m6
m7

m8

e1 e3 e4

m8 now goes to e3 due to shared middle initial and last name. Entity e5 becomes empty and is dropped.

3. Recompute entity profiles. 4. Repeat Steps 2-3 until convergence.


m1
m2

m3
m4
m5

m6
m7

m8

e3 e4

Collective Mention Matching

1. **Match tuples**
2. **“Enrich” each tuple with information from other tuples that match it; or create “super tuples” that represent groups of matching tuples.**
3. **Repeat Steps 1-2 until convergence.**

Key ideas: enrich each tuple, iterate

Some recent algorithms that employ these ideas:

Pedro Domingos group at Washington, Dan Roth group at Illinois, Andrew McCallum group at UMass, Lise Getoor group at Maryland, Alon Halevy group at Washington (SEMEX), Ray Mooney group at Texas-Austin, Jiawei Han group at Illinois, and more

What new mention matching challenges does IE over text raise?

1. Static data: challenges similar to those in extracting mentions.

2. Dynamic data: challenges in tracking mentions / entities

Classical Mention Matching

- Applies just a single “matcher”
- Focuses mainly on developing matchers with higher accuracy

Real-world IE applications need more

We Need a Matching Workflow

To illustrate with a simple example:

Only one Luis Gravano

d_1: Luis Gravano's Homepage

L. Gravano, K. Ross.
Text Databases. SIGMOD 03

L. Gravano, J. Sanz.
Packet Routing. SPAA 91

d_2: Columbia DB Group Page

Members
L. Gravano K. Ross J. Zhou

L. Gravano, J. Zhou.
Text Retrieval. VLDB 04

d_3: DBLP

Luis Gravano, Kenneth Ross.
Digital Libraries. SIGMOD 04

Luis Gravano, Jingren Zhou.
Fuzzy Matching. VLDB 01

Luis Gravano, Jorge Sanz.
Packet Routing. SPAA 91

Chen Li, Anthony Tung.
Entity Matching. KDD 03

Chen Li, Chris Brown.
Interfaces. HCI 99

d_4: Chen Li's Homepage

C. Li.
Machine Learning. AAAI 04

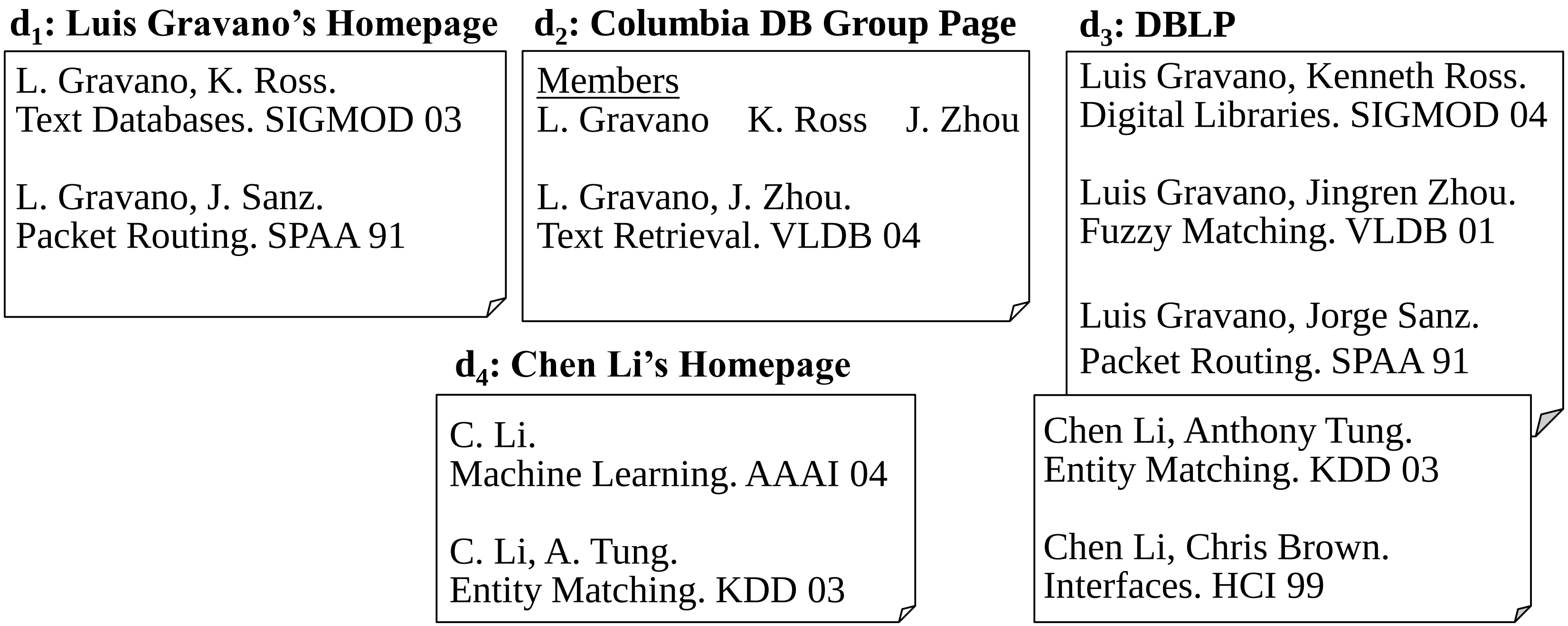
C. Li, A. Tung.
Entity Matching. KDD 03

Two
Chen Li-s

What is the best way to match mentions here?

A liberal matcher: correctly predicts that there is one Luis Gravano, but incorrectly predicts that there is one Chen Li

s_0 matcher: two mentions match if they share the same name.



d_1: Luis Gravano's Homepage

L. Gravano, K. Ross.
Text Databases. SIGMOD 03

L. Gravano, J. Sanz.
Packet Routing. SPAA 91

d_2: Columbia DB Group Page

Members
L. Gravano K. Ross J. Zhou

L. Gravano, J. Zhou.
Text Retrieval. VLDB 04

d_3: DBLP

Luis Gravano, Kenneth Ross.
Digital Libraries. SIGMOD 04

Luis Gravano, Jingren Zhou.
Fuzzy Matching. VLDB 01

Luis Gravano, Jorge Sanz.
Packet Routing. SPAA 91

Chen Li, Anthony Tung.
Entity Matching. KDD 03

Chen Li, Chris Brown.
Interfaces. HCI 99

d_4: Chen Li's Homepage

C. Li.
Machine Learning. AAAI 04

C. Li, A. Tung.
Entity Matching. KDD 03

A conservative matcher: predicts multiple Gravanos and Chen Lis

s_1 matcher: two mentions match if they share the same name and at least one co-author name.

d_1: Luis Gravano's Homepage

L. Gravano, K. Ross.
Text Databases. SIGMOD 03

L. Gravano, J. Sanz.
Packet Routing. SPAA 91

d_2: Columbia DB Group Page

Members
L. Gravano   K. Ross   J. Zhou

L. Gravano, J. Zhou.
Text Retrieval. VLDB 04

d_3: DBLP

Luis Gravano, Kenneth Ross.
Digital Libraries. SIGMOD 04

Luis Gravano, Jingren Zhou.
Fuzzy Matching. VLDB 01

Luis Gravano, Jorge Sanz.
Packet Routing. SPAA 91

Chen Li, Anthony Tung.
Entity Matching. KDD 03

Chen Li, Chris Brown.
Interfaces. HCI 99

d_4: Chen Li's Homepage

C. Li.
Machine Learning. AAAI 04

C. Li, A. Tung.
Entity Matching. KDD 03

Better solution:
apply both matchers in a workflow

d_1: Luis Gravano's Homepage

L. Gravano, K. Ross.
Text Databases. SIGMOD 03

L. Gravano, J. Sanz.
Packet Routing. SPAA 91

d_2: Columbia DB Group Page

Members
L. Gravano K. Ross J. Zhou

L. Gravano, J. Zhou.
Text Retrieval. VLDB 04

d_3: DBLP

Luis Gravano, Kenneth Ross.
Digital Libraries. SIGMOD 04

Luis Gravano, Jingren Zhou.
Fuzzy Matching. VLDB 01

Luis Gravano, Jorge Sanz.
Packet Routing. SPAA 91

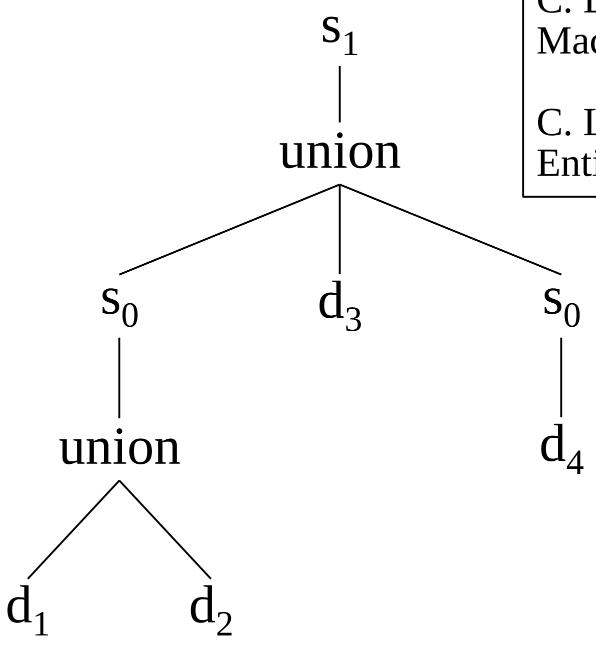
Chen Li, Anthony Tung.
Entity Matching. KDD 03

Chen Li, Chris Brown.
Interfaces. HCI 99

d_4: Chen Li's Homepage

C. Li.
Machine Learning. AAAI 04

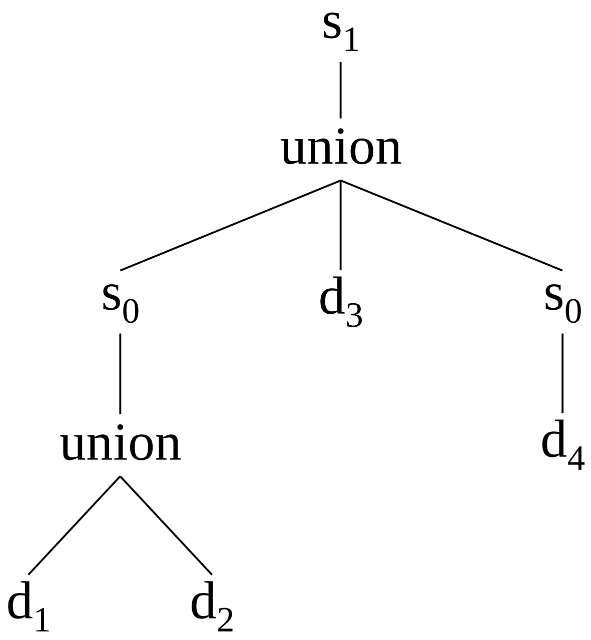
C. Li, A. Tung.
Entity Matching. KDD 03



s_0 matcher: two mentions match if they share the same name.

s_1 matcher: two mentions match if they share the same name and at least one co-author name.

Intuition Behind This Workflow



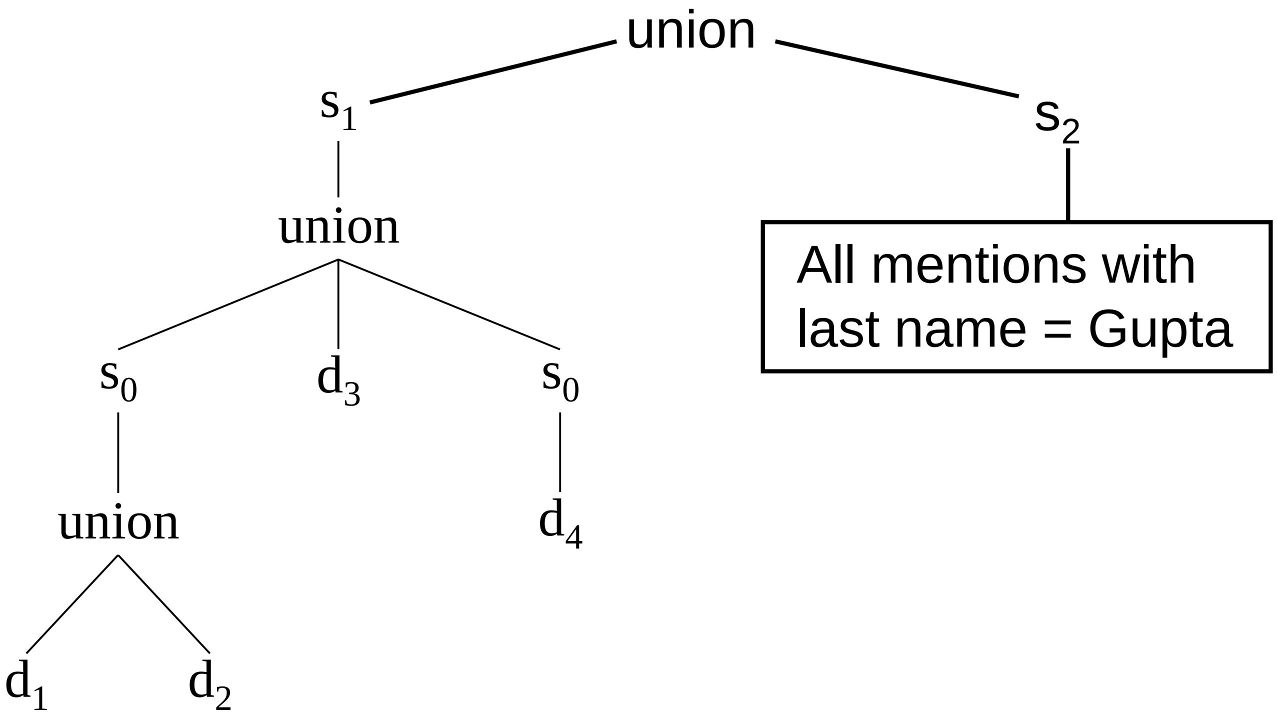
We control how tuple enrichment happens, using different matchers.

Since homepages are often unambiguous, we first match homepages using the simple matcher s_0. This allows us to collect co-authors for Luis Gravano and Chen Li.

So when we finally match with tuples in DBLP, which is more ambiguous, we (a) already have more evidence in form of co-authors, and (b) use the more conservative matcher s_1.

Another Example

- Suppose distinct researchers X and Y have very similar names, and share some co-authors
  - e.g., Ashish Gupta and Ashish K. Gupta
- Then s_1 matcher does not work, need a more conservative matcher s_2



Need to Exploit a Lot of Domain Knowledge in the Workflow

[From Shen, Li, Doan, AAAI-05]

Type	Example
Aggregate	No researcher has chaired more than 3 conferences in a year
Subsumption	If a citation X from DBLP matches a citation Y in a homepage, then each author in Y matches some author in X
Neighborhood	If authors X and Y share similar names and some coauthors, they are likely to match
Incompatible	No researcher exists who has published in both HCI and numerical analysis
Layout	If two mentions in the same document share similar names, they are likely to match
Uniqueness	Mentions in the PC listing of a conference refer to different researchers
Ordering	If two citations match,then their authors will be matched in order
Individual	The researcher named “Mayssam Saria” has fewer than five mentions in DBLP (e.g. being a new graduate student with fewer than five papers)

Need Support for Incremental update of matching workflow

- We have run a matching workflow E on a huge data set D
- Now we modified E a little bit into E’
- How can we run E’ efficiently over D?
  - exploiting the results of running E over D
- Similar to exploiting materialized views
- Crucial for many settings:
  - testing and debugging
  - expansion during deployment
  - recovering from crash

Research Challenges

- Similar to those in extracting mentions
- Need right model / representation language
- Develop basic operators: matcher, merger, etc.
- Ways to combine them → match execution plan

- Ways to optimize plan for accuracy/runtime
  - challenge: estimate their performance
- Akin to relational query optimization

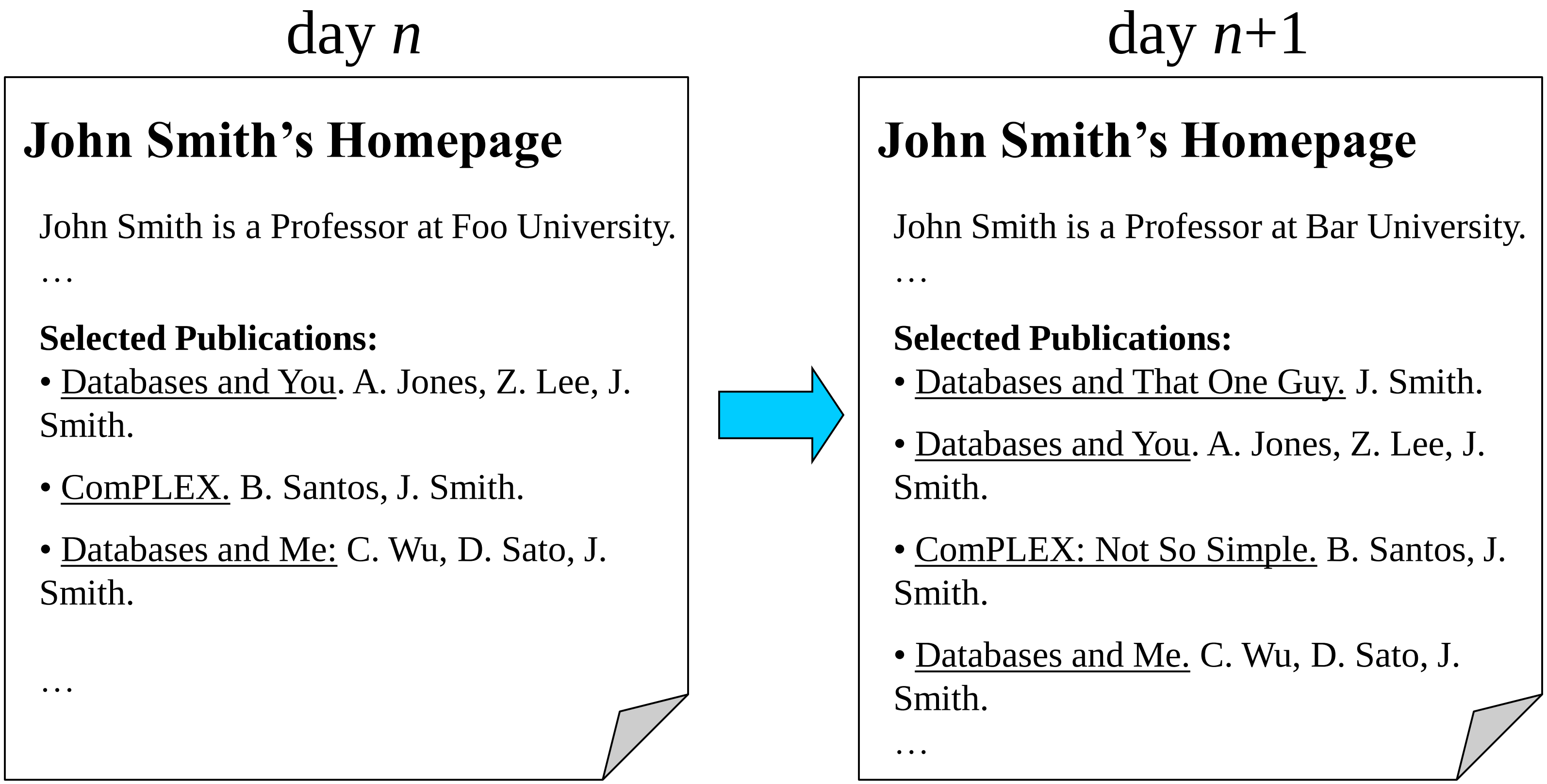
The Ideal Entity Matching Solution

- We throw in all types of information
  - training data (if available)
  - domain constraints
- and all types of matchers + other operators
  - SVM, decision tree, etc.
- Must be able to do this as declaratively as possible (similar to writing a SQL query)
- System automatically compile a good match execution plan
  - with respect to accuracy/runtime, or combination thereof
- Easy for us to debug, maintain, add domain knowledge, add patches

Recent Work / Starting Point

- SERF project at Stanford
  - Develops a generic infrastructure
  - Defines basic operators: match, merge, etc.
  - Finds fast execution plans
- Data cleaning project at MSR
  - Solution to match incoming records against existing groups
  - E.g., [Chaudhuri, Ganjam, Ganti, Motwani, SIGMOD-03]
- Cimple project at Illinois / Wisconsin
  - SOCCER matching approach
  - Defines basic operators, finds highly accurate execution plans
  - Methods to exploit domain constraints [Shen, Li, Doan, AAAI-05]
- Semex project at Washington
  - Methods to expoit domain constraints [Dong et. al., SIGMOD-05]

Mention Tracking



day *n*

John Smith's Homepage

John Smith is a Professor at Foo University.

…

Selected Publications:

• Databases and You. A. Jones, Z. Lee, J. Smith.

• ComPLEX. B. Santos, J. Smith.

• Databases and Me: C. Wu, D. Sato, J. Smith.

…

day *n*+1

John Smith's Homepage

John Smith is a Professor at Bar University.

…

Selected Publications:

• Databases and That One Guy. J. Smith.

• Databases and You. A. Jones, Z. Lee, J. Smith.

• ComPLEX: Not So Simple. B. Santos, J. Smith.

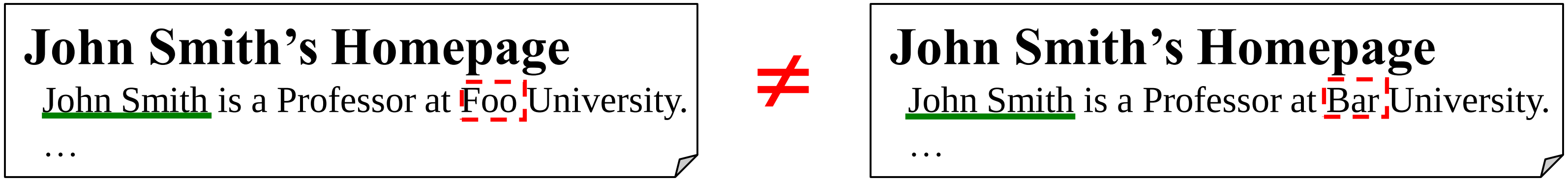
• Databases and Me. C. Wu, D. Sato, J. Smith.

…

- How do you tell if a mention is old or new?
  - Compare mention semantics between days
  - How do we determine a mention's semantics?

Mention Tracking

- Using fixed-width context windows often works …



John Smith's Homepage

John Smith is a Professor at Foo University.

…

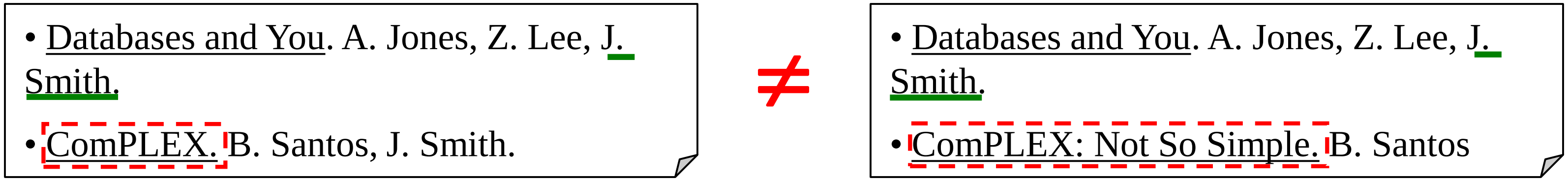
≠

John Smith's Homepage

John Smith is a Professor at Bar University.

…

- But not always.

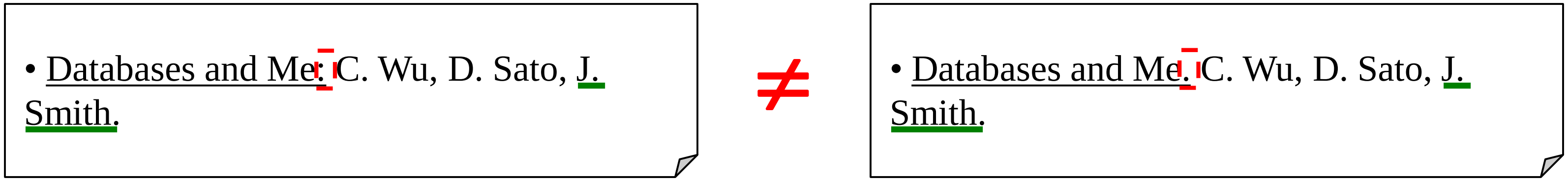


- Databases and You. A. Jones, Z. Lee, J. Smith.
- ComPLEX. B. Santos, J. Smith.

≠

- Databases and You. A. Jones, Z. Lee, J. Smith.
- ComPLEX: Not So Simple. B. Santos

- Even intelligent windows can use help with semantics



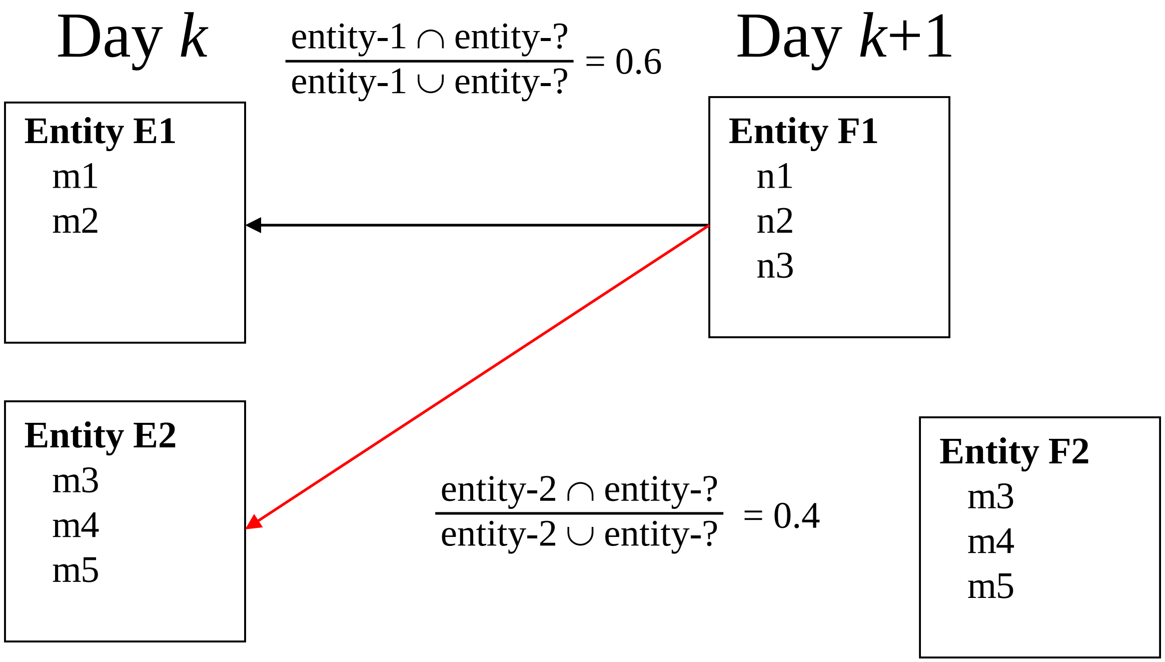
- Databases and Me: C. Wu, D. Sato, J. Smith.

≠

- Databases and Me. C. Wu, D. Sato, J. Smith.

Entity Tracking

- Like mention tracking, how do you tell if an entity is old or new?
- Entities are sets of mentions, so we use a Jaccard distance:

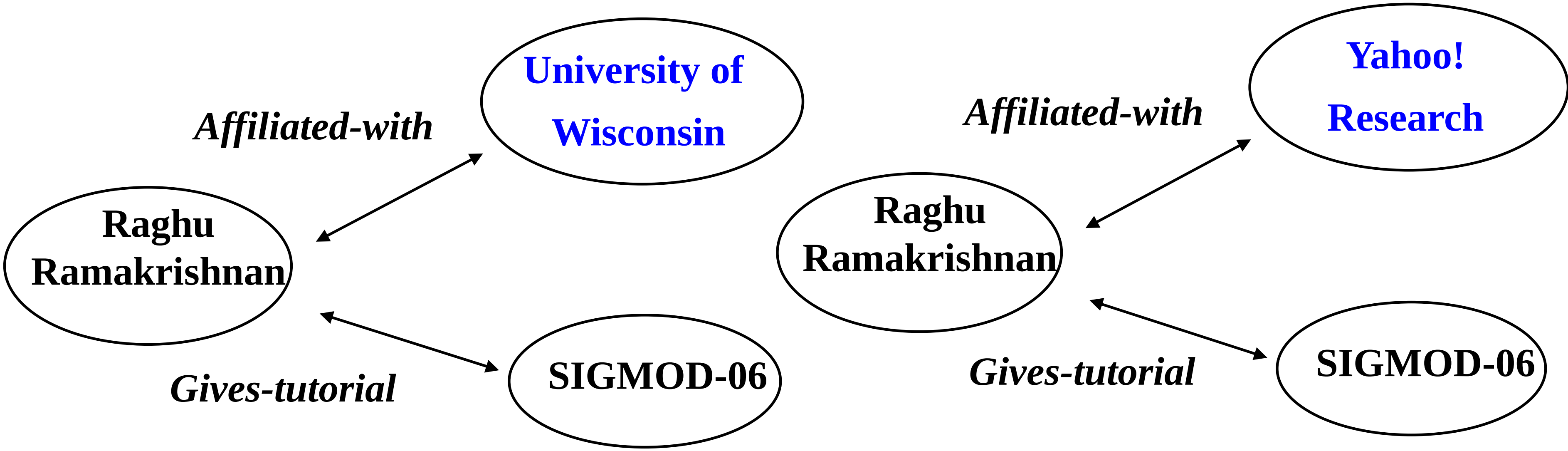


$$\frac{\text{entity-1} \cap \text{entity-?}}{\text{entity-1} \cup \text{entity-?}} = 0.6$$

$$\frac{\text{entity-2} \cap \text{entity-?}}{\text{entity-2} \cup \text{entity-?}} = 0.4$$

Monitoring and Event Detection

- The real world might have changed!
  - And we need to detect this by analyzing changes in extracted information



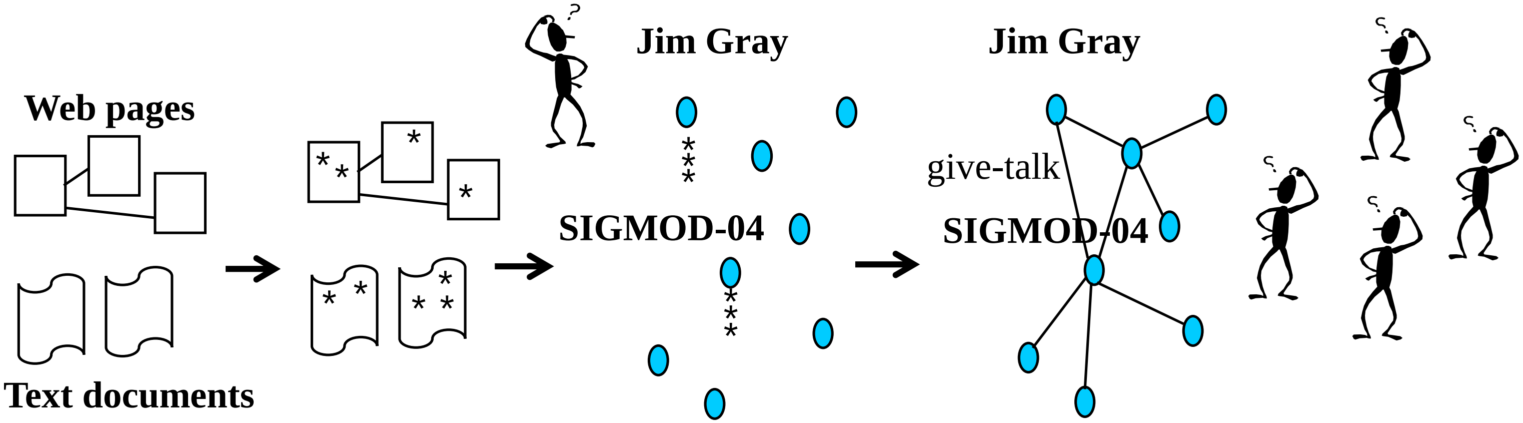
Infer that Raghu Ramakrishnan has moved to Yahoo! Research

Tutorial Roadmap

- Introduction to managing IE [RR]
  - Motivation
  - What’s different about *managing* IE?
- Major research directions
  - Extracting mentions of entities and relationships [SV]
    - Uncertainty management
  - Disambiguating extracted mentions [AD]
    - Tracking mentions and entities over time
  - Understanding, correcting, and maintaining extracted data [AD]
    - Provenance and explanations
    - Incorporating user feedback

Understanding, Correcting, and Maintaining Extracted Data

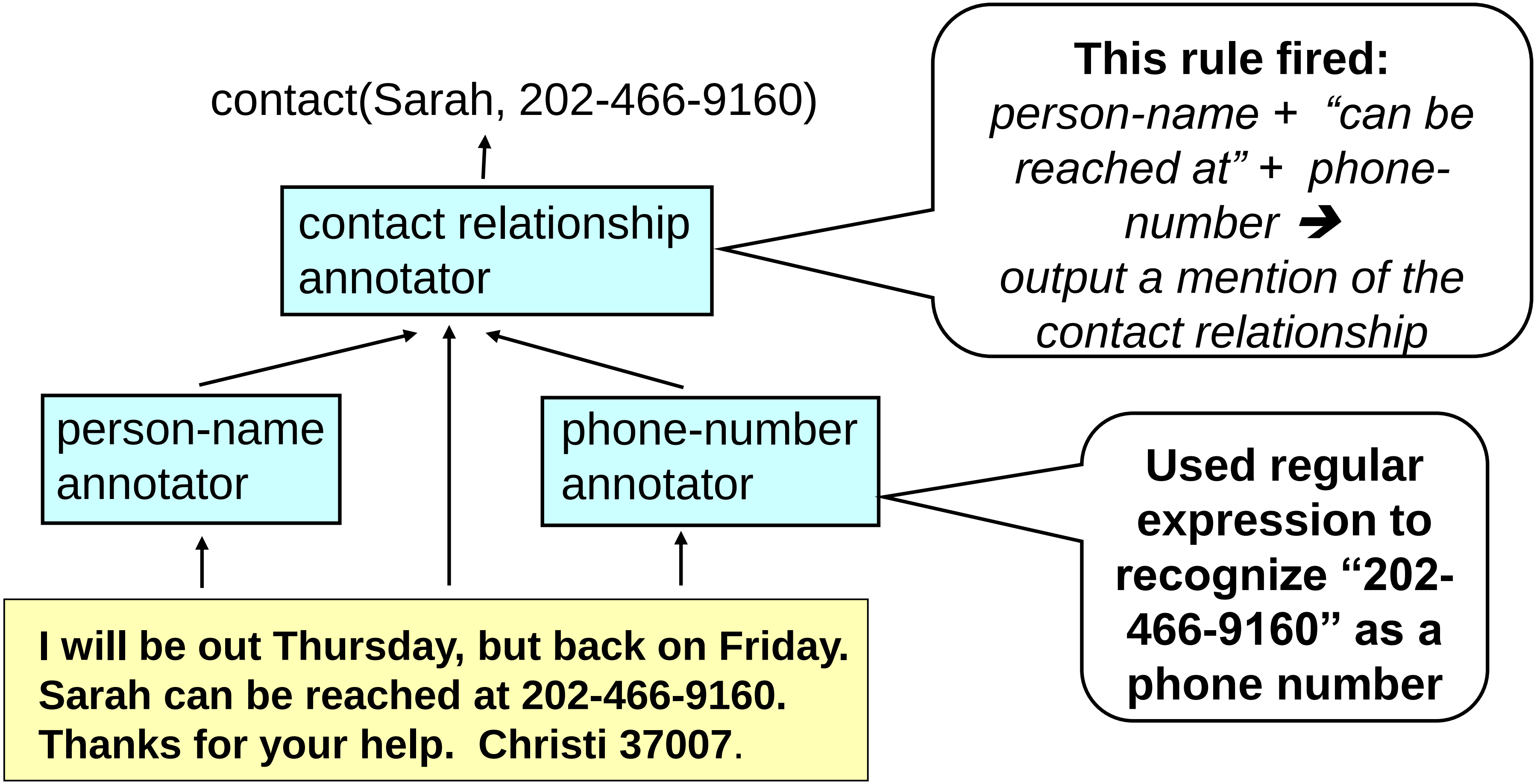
Understanding Extracted Data



- Important in at least three contexts
  - Development ➔ developers can fine tune system
  - Provide services (keyword search, SQL queries, etc.)
    ➔users can be confident in answers
  - Provide feedback
    ➔ developers / users can provide good feedback
- Typically provided as provenance (aka lineage)
  - Often a tree showing the origin and derivation of data

An Example

System extracted contact(Sarah, 202-466-9160). Why?



contact(Sarah, 202-466-9160)

contact relationship annotator

This rule fired:
person-name + “can be reached at” + phone-number ➔
output a mention of the contact relationship

person-name annotator

phone-number annotator

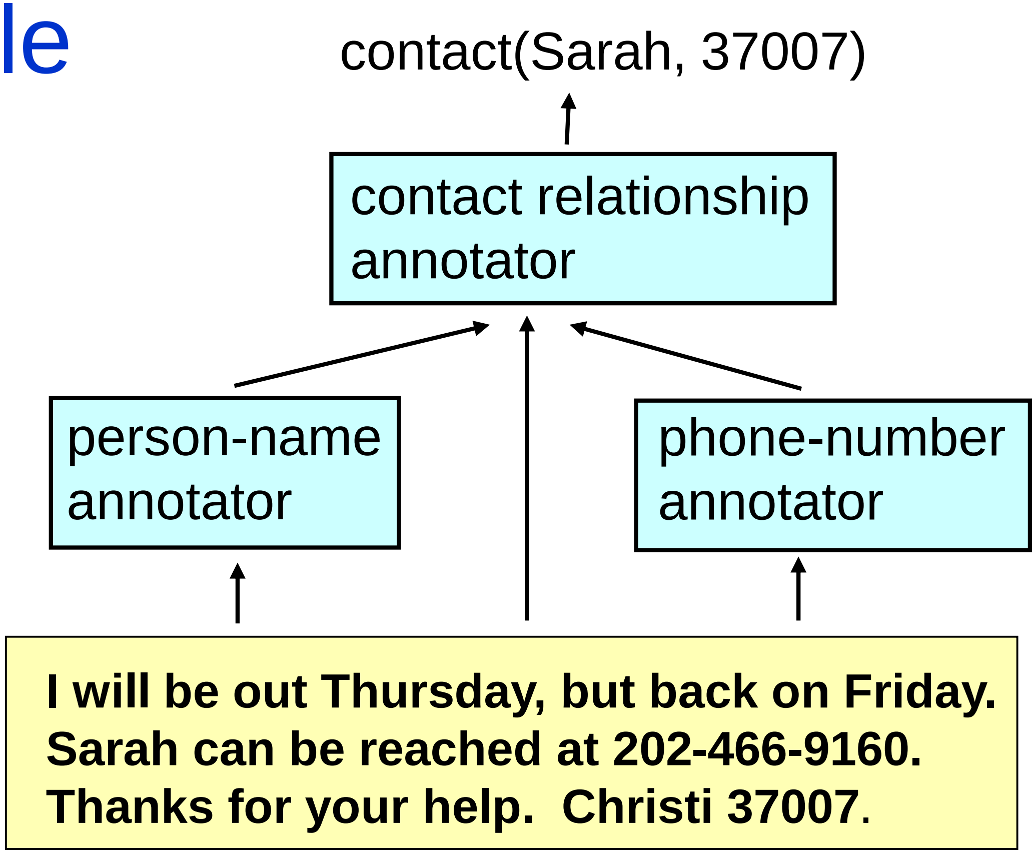
Used regular expression to recognize “202-466-9160” as a phone number

I will be out Thursday, but back on Friday.
Sarah can be reached at 202-466-9160.
Thanks for your help. Christi 37007.

In Practice, Need More than Just Provenance Tree

- Developer / user often want **explanations**
  - why X was extracted?
  - why Y was not extracted?
  - why system has higher confidence in X than in Y?
  - what if ... ?
- Explanations thus are related to,
  but different from provenance

An Example



I will be out Thursday, but back on Friday. Sarah can be reached at 202-466-9160. Thanks for your help. Christi 37007.

Why was “202-466-9160” not extracted?

Explanation:

(1) The relationship annotator uses the following rule to extract 37007:

person name + at most 10 tokens +
“can be reached at” +
at most 6 tokens + phone number ➔ contact(person name, phone number).

(2) “202-466-9160” fits into the part “at most 6 tokens”.

Generating Explanations is Difficult

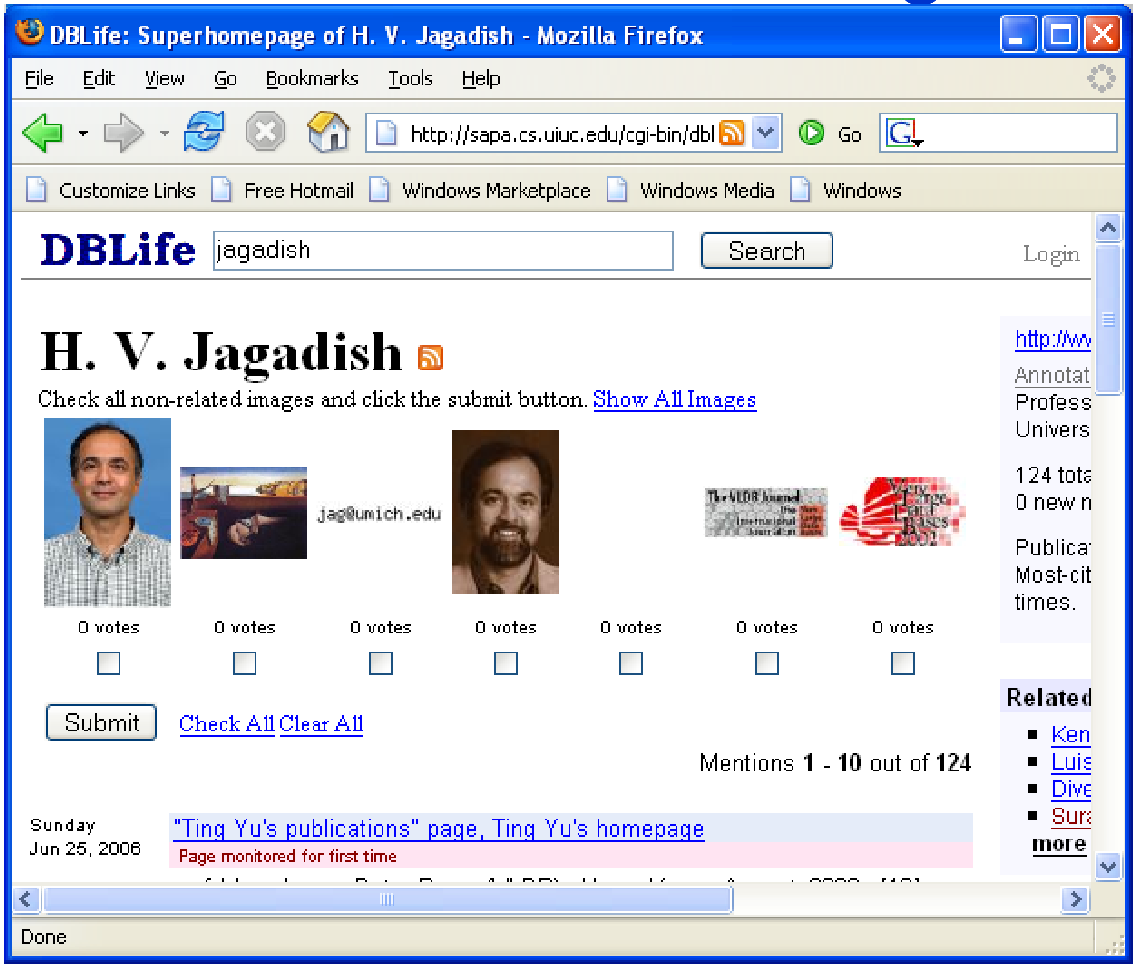
- Especially for
  - why was A not extracted?
  - why does system rank A higher than B?
- Reasons
  - many possible causes for the fact that “A was not extracted”
  - must examine the provenance tree to know which components are chiefly responsible for causing A to be ranked higher than B
  - provenance trees can be huge, especially in continuously running systems, e.g., DBLife
- Some work exist in related areas, but little on generating explanations for IE over text
  - see [Dhamankar et. al., SIGMOD-04]:
    generating explanations for schema matching

System developers and users can use explanations / provenance to provide feedback to system (i.e., this extracted data piece is wrong), or manually correct data pieces

This raises many serious challenges.

Consider the case of multiple users' providing feedback ...

Motivating Example



http://sapa.cs.uiuc.edu/cgi-bin/dbl

DBLife jagadish

H. V. Jagadish

Check all non-related images and click the submit button. Show All Images

jag@umich.edu

0 votes	0 votes	0 votes	0 votes	0 votes	0 votes	0 votes
□	□	□	□	□	□	□

Check All Clear All

Mentions **1** - **10** out of **124**

Sunday
Jun 25, 2006

"Ting Yu's publications" page, Ting Yu's homepage

Page monitored for first time

http://ww
Annotat
Profess
Univers

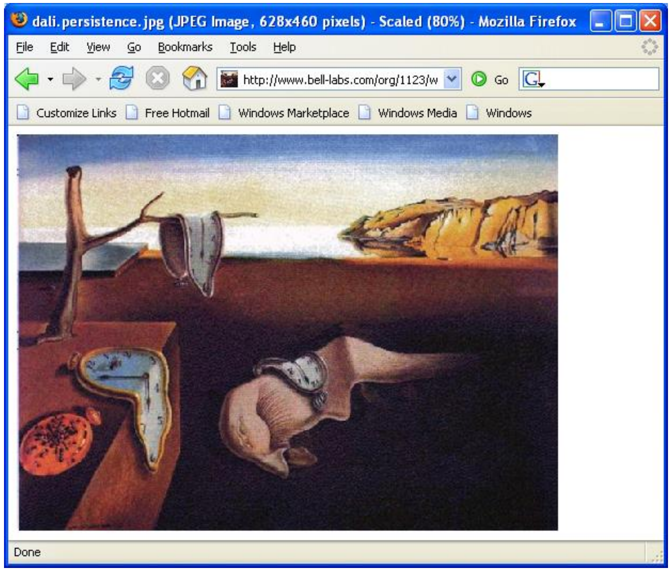
124 tota
0 new n

Publica
Most-cit
times.

Related

- Ken
- Luis
- Dive
- Sura

more


dali.persistence.jpg (JPEG Image, 628x460 pixels) - Scaled (80%) - Mozilla Firefox
File
Edit
View
Go
Bookmarks
Tools
Help
http://www.bell-labs.com/org/1123/w
Go
Customize Links
Free Hotmail
Windows Marketplace
Windows Media
Windows
Done

The General Idea

- Many real-world applications inevitably have multiple developers and many users
- How to exploit feedback efforts from all of them?
- Variants of this is known as
  - collective development of system, mass collaboration, collective curation, Web 2.0 applications, etc.
- Has been applied to many applications
  - open-source software, bug detection, tech support group, Yahoo! Answers, Google Co-op, and many more
- Little has been done in IE contexts
  - except in industry, e.g., epinions.com

Challenges

- If X and Y both edit a piece of extracted data D, they may edit the same data unit differently
- How would X and Y reconcile / share their edition?
- E.g., the ORCHESTRA project at Penn

  [Taylor & Ives, SIGMOD-06]

- How to entice people to contribute?
- How to handle malicious users?
- What types of extraction tasks are most amenable to mass collaboration?
- E.g., see MOBS project at Illinois [WebDB-03, ICDE-05]

Maintenance

- As data evolves, extractors often break

```
<HTML>
<TITLE>Some Country Codes</TITLE>
<B>Congo</B> <I>242</I> <BR>
<B>Egypt</B> <I>20</I> <BR>
<B>Belize</B> <I>501</I> <BR>
<B>Spain</B> <I>34</I> <BR>
</BODY></HTML>
```

(Congo, 242)
(Egypt, 20)
(Belize, 501)
(Spain, 34)

```
<HTML>
<TITLE>Some Country Codes</TITLE>
<B>Congo</B> <I>Africa</I> <I>242</I> <BR>
<B>Egypt</B> <I>Africa</I><I>20</I> <BR>
<B>Belize</B> <I>N. America</I> <I>501</I> <BR>
<B>Spain</B> <I>Europe</I><I>34</I> <BR>
</BODY></HTML>
```

(Congo, Africa)
(Egypt, Africa)
(Belize, N. America)
(Spain, Europe)

Maintenance: Key Challenges

- Detect if an extractor or a set of extractors is broken
- Pinpoint the source of errors
- Suggest repairs or automatically repairs extractors
- Build semantic debuggers?
- Scalability issues

Related Work / Starting Points

- Detect broken extractors
  - Nick Kushmerick group in Ireland, Craig Knoblock group at ISI, Chen Li group at UCI, AnHai Doan group at Illinois
- Repair broken extractors
  - Craig Knoblock group at ISI
- Mapping maintenance
  - Renee Miller group at Toronto, Lucian Popa group at Almaden

Summary: Key Points of Tutorial

- Lot of future activity in text / Web management

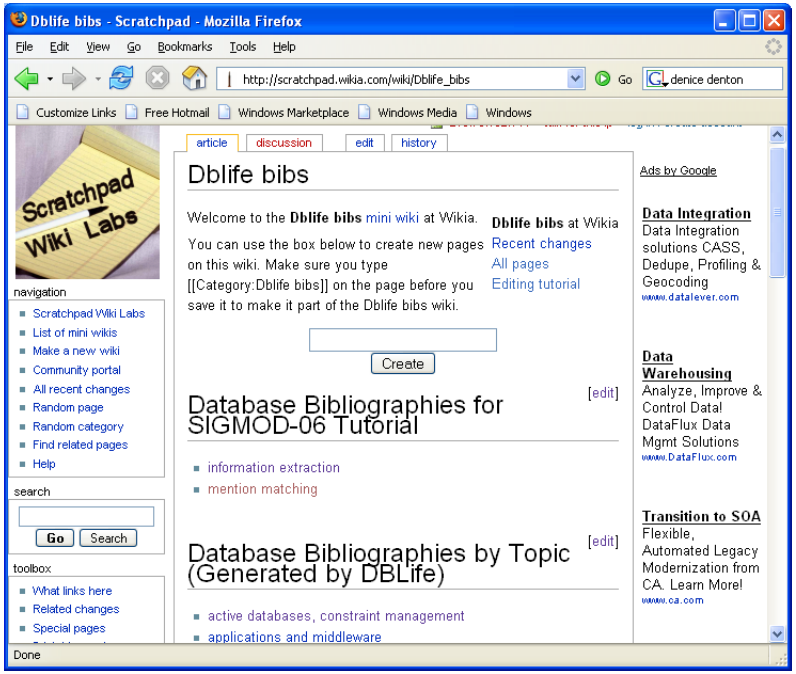
- To build IE-based applications ➔ must go beyond developing IE components, to **managing the entire IE process**:
  - Manage the IE workflow, manage mention matching
  - Provide useful services over extracted data
  - Manage uncertainty, understand, correct, and maintain extracted data

- Solutions here + IR components ➔ can significantly extend the footprint of DBMSs

Think "System R" for IE-based applications!

How Can You Start

- We are putting pointers to literature, tools, & data at http://scratchpad.wikia.com/wiki/Dblife_bibs (all current DBLife bibliographies also reside here)



Dblife bibs

Welcome to the **Dblife bibs** mini wiki at Wikia.

You can use the box below to create new pages on this wiki. Make sure you type [[Category:Dblife bibs]] on the page before you save it to make it part of the Dblife bibs wiki.

Dblife bibs at Wikia
Recent changes
All pages
Editing tutorial

Create

Database Bibliographies for SIGMOD-06 Tutorial [edit]

- information extraction
- mention matching

Database Bibliographies by Topic (Generated by DBLife) [edit]

- active databases, constraint management
- applications and middleware

- **Please contribute!**
- Also watch that space
  - Tutorial slides will be put there
  - Data will be available from DBLife, Avatar project, and Yahoo, in significant amount
- Will be able to navigate there from our homepages