

# Best-Effort Data Integration



**AnHai Doan**

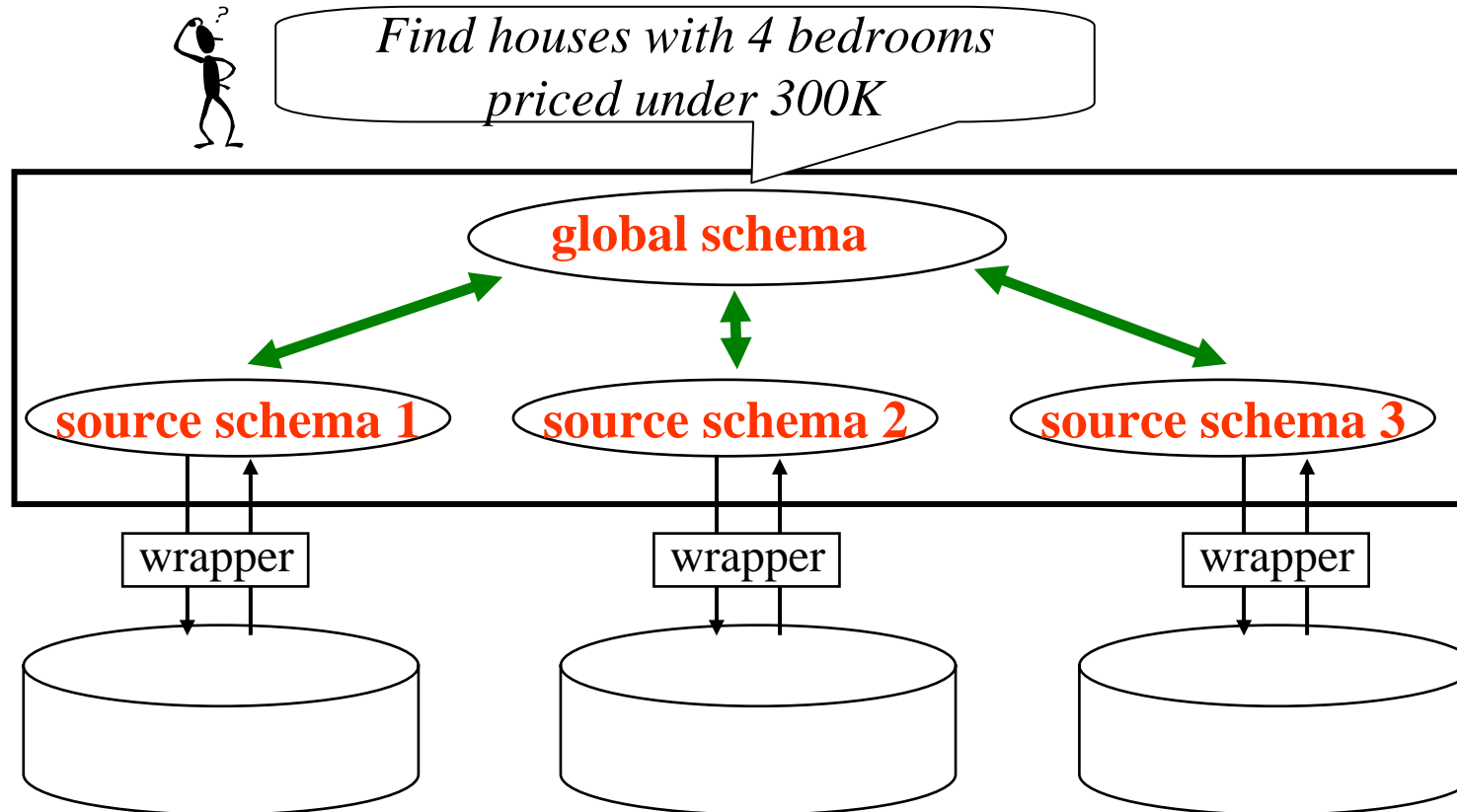
University of Wisconsin-Madison

*Joint work with Fei Chen, Pedro DeRose,  
Robert McCann, Yoonkyong Lee, Mayssam Sayyadian,  
Warren Shen, Luis Gravano, Raghu Ramakrishnan*

# Data Integration: Current Status

- **We have made tremendous progress in the last 30 years**
  - **develop foundations**: mediator model, GAV, LAV
  - **build on the foundation**: query reformulation, provenance, uncertainty, schema matching and mapping, entity resolution, adaptive query processing, managing inconsistent data, P2P, etc.
  - **branch into applications**: bio-informatics, geo-spatial, Web, ...
  - **join forces**: databases, AI, Web
- **But data integration remains hard**
  - intractable, AI-complete, etc.
- **Partly because we often want exact, precise integration**

# Precise Data Integration



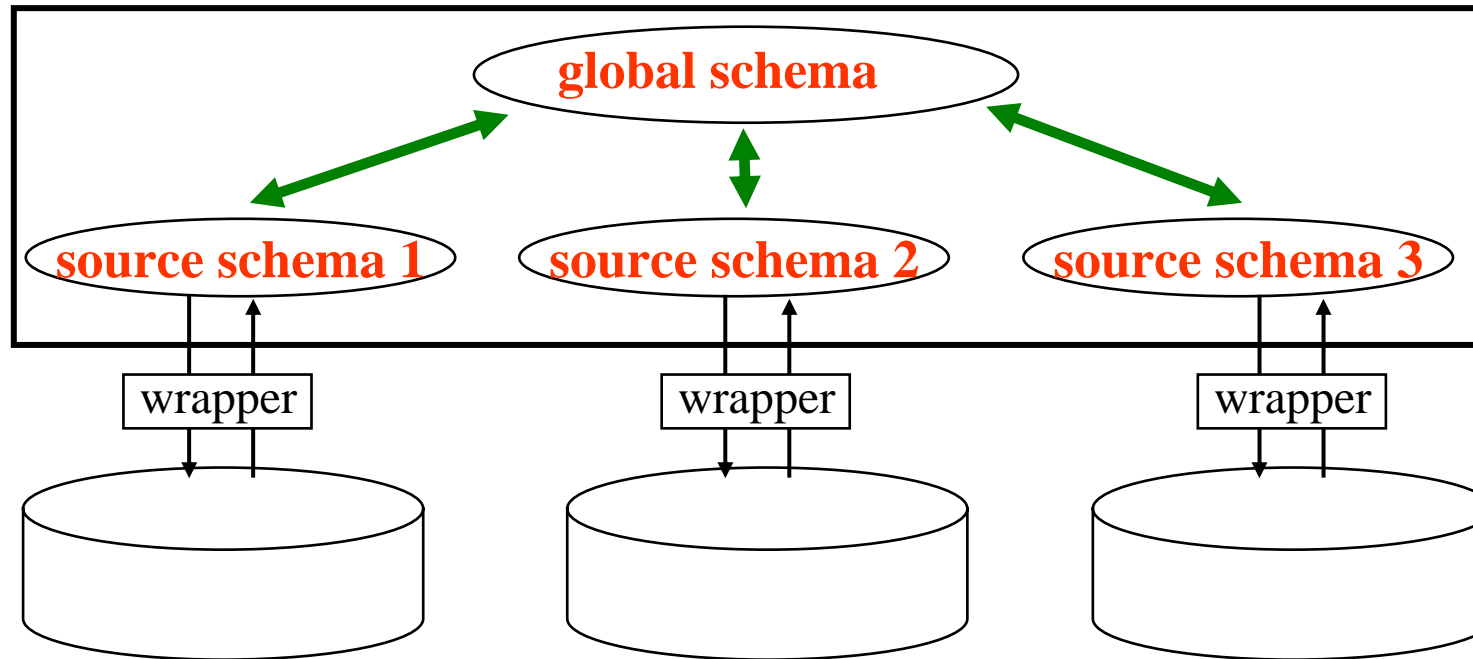
- **Original motivation: business applications**

- e.g., payroll, human resources, banking
- here, anything less is NOT usable

# However, the Application Landscape Has Changed in the Past Decade

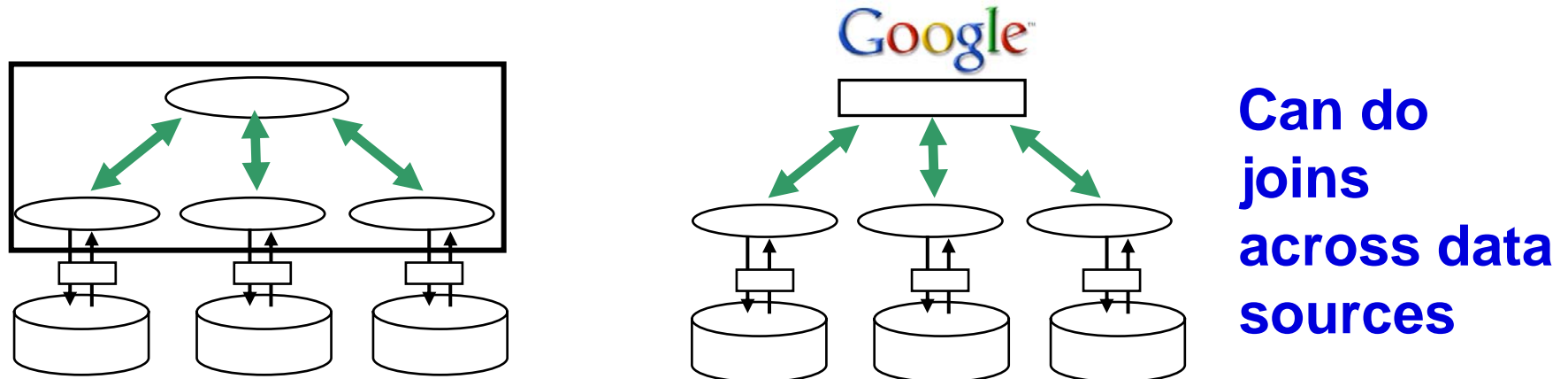
- **Today, precise integration continues to be critical**
  - e.g., expedia.com
- **But for many emerging application domains, best-effort data integration**
  - often incurs far less cost
  - may already prove very useful
- **Examples**
  - citation tracking (e.g., Citeseer, Google Scholar)
  - personal information management
  - scientific, exploratory data analysis
  - intelligence analysis for homeland security
  - business intelligence
  - Web integration scenarios (e.g., Froogle)

# Best-Effort Data Integration



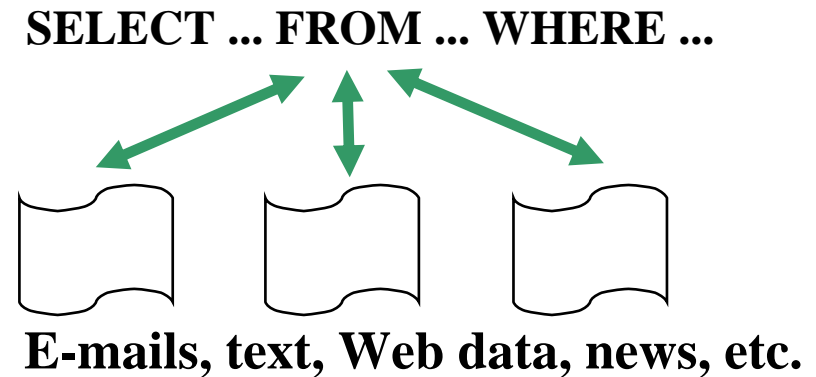
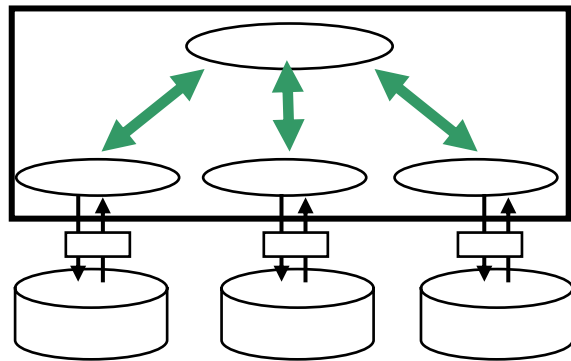
- Remove, simplify, or make “less precise”  
certain components
- Employ automatic techniques
- To go “the last mile”: learn from human interaction

# Example 1: Simplify Global Schema → Keyword Search over Multiple Databases



- **Novel problem**
- **Very useful for urgent / one-time DI needs**
  - also when users are SQL-illiterate
- **Proposed solution in ICDE-07a**
  - combines IR, schema matching, entity resolution, and AI planning

# Example 2: Simplify Wrappers → Structured Queries over Text/Web Data



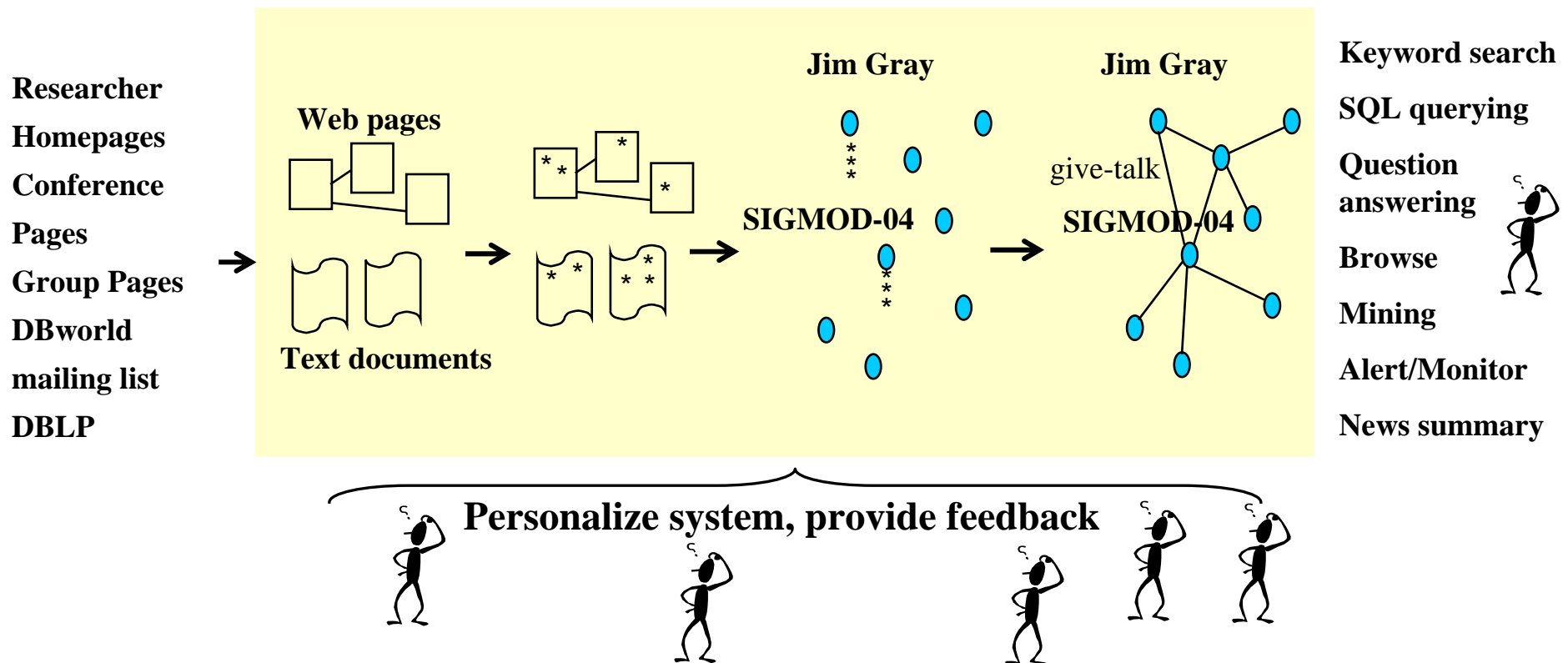
- Novel problem
- Proposed solution in ICDE-07b

# Example 3: Best-Effort Data Integration for Web Communities

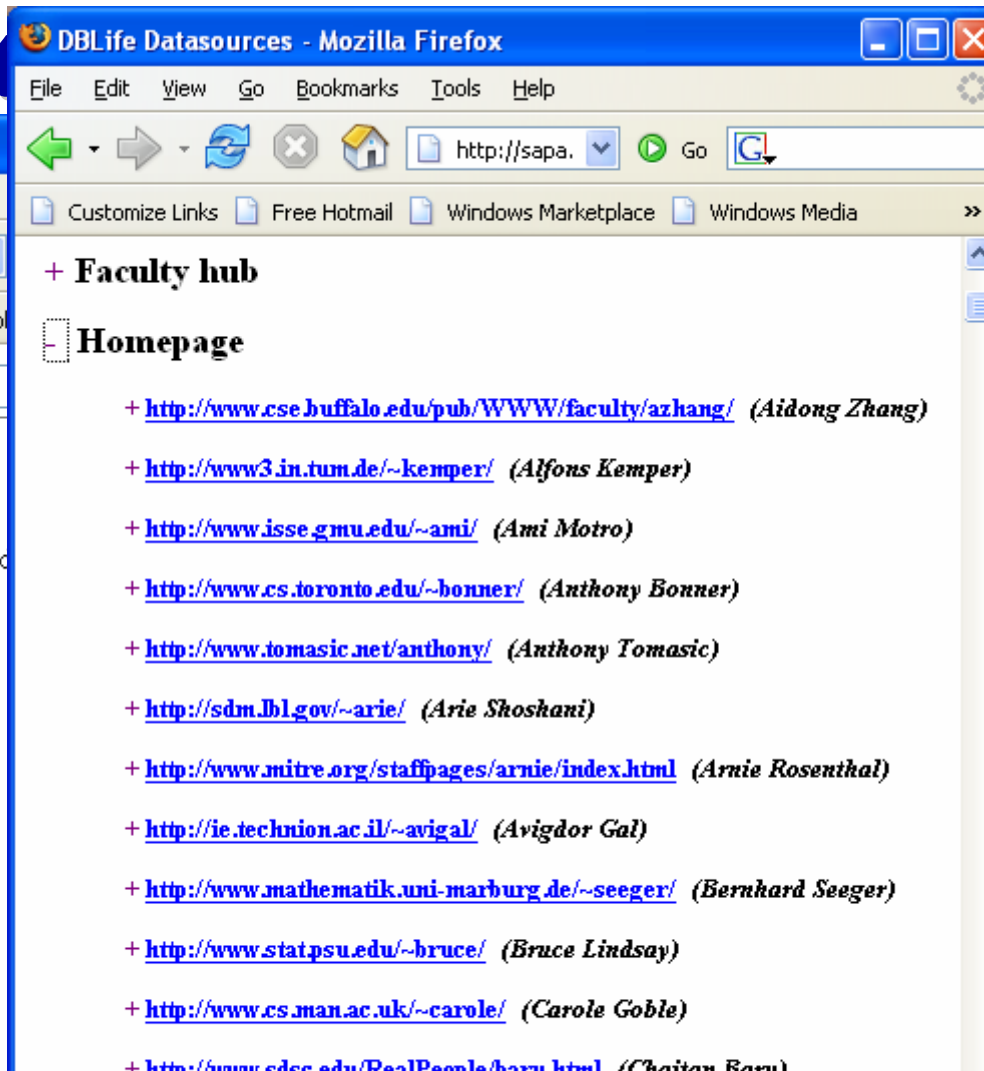
- **Numerous data-rich communities**
  - database researchers, movie fans, legal professionals, bioinformatics, enterprise intranets, etc.
- **Each community = many disparate data sources + people**
- **Members often want to discovery, query, monitor information in the community**
  - any interesting 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?
  - what are current hot topics? who has moved where?

# The Cimple Project @ Wisconsin/Yahoo!

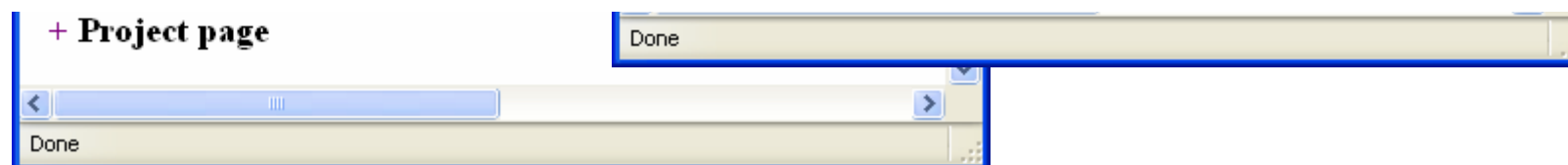
Builds structured data portals using  
**extraction + integration + mass collaboration**



# Protot



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



DBWorld Message - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites

Address <http://sapa.cs.uiuc.edu/cgi-bin/dblife/markUpWebpage.cgi?markurl=http://www.cs.wisc.edu/dbworld/>

Google how to call us from italy

Y! Search Web

This is [DBLife's](#) annotated version of <http://www.cs.wisc.edu/dbworld/>

**Call for Papers**

Label (Previous)

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

**Deadline:** June 15, 2006

Recent advances in computer science

**Topics of interest include:**

\* Cyber-infrastructure

AnHai Doan's HomePage - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites

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

Google how to call us from italy

Y! Search Web

[Schema & Ontology Matching](#)

[Community Information Management](#)

[Data Integration](#)

---

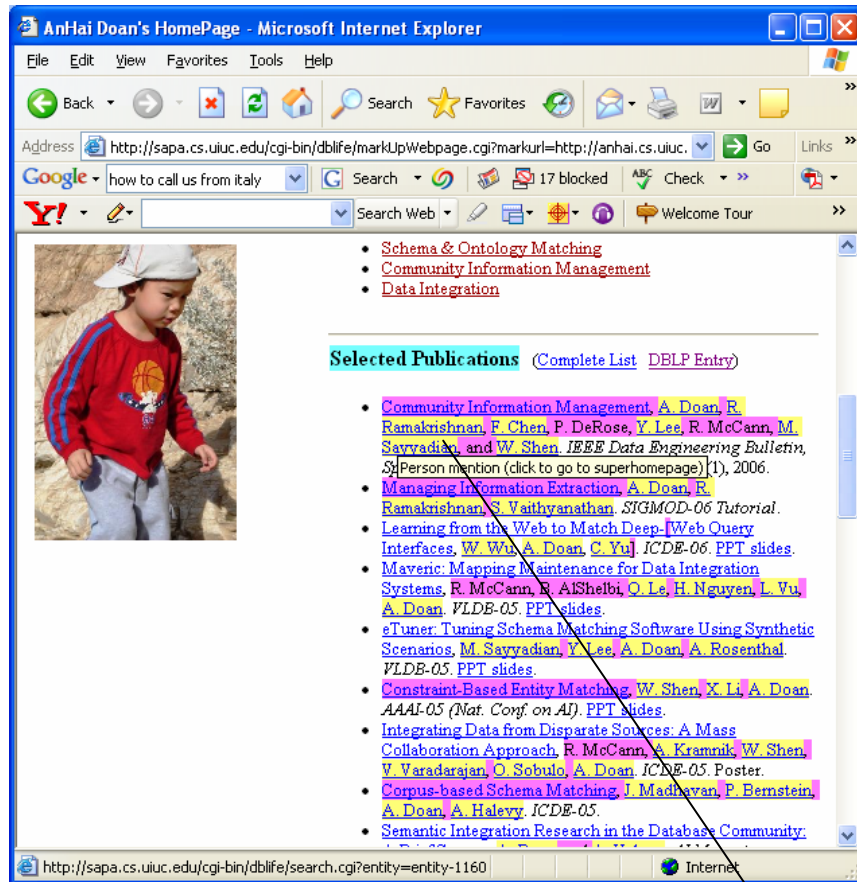
**Selected Publications** ([Complete List](#) [DBLP Entry](#))

- [Community Information Management, A. Doan, R. Ramakrishnan, F. Chen, P. DeRose, Y. Lee, R. McCann, M. Sayyadian, and W. Shen. \*IEEE Data Engineering Bulletin\*, Special 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, O. 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: A Survey](#)

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



# Data Integration



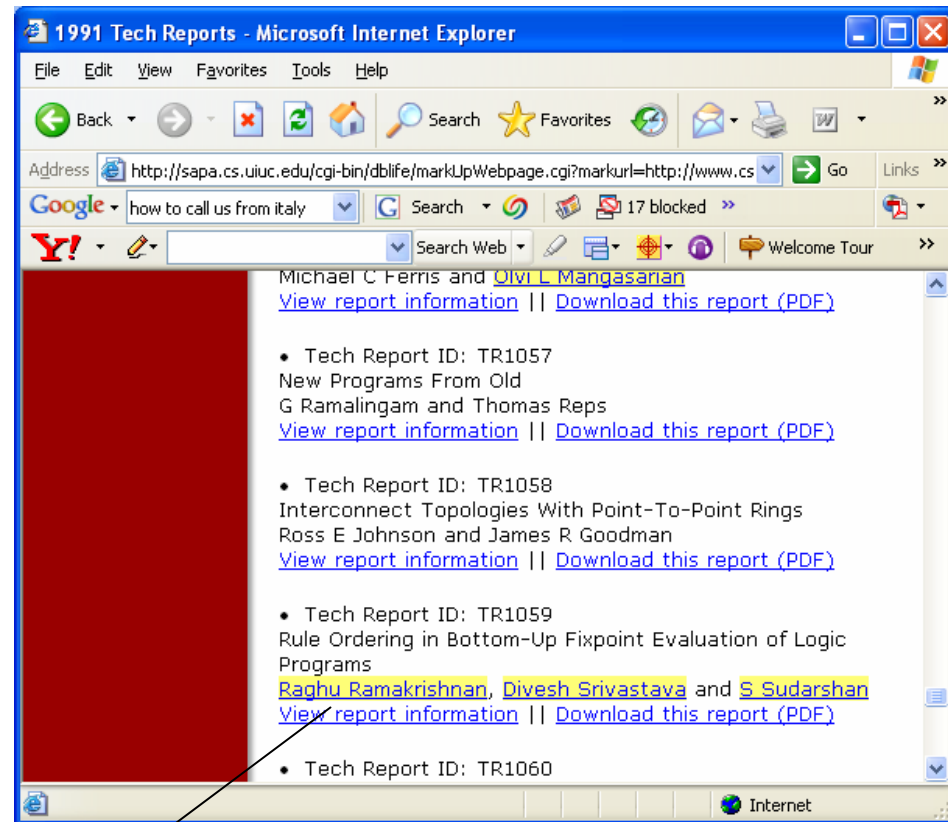
AnHai Doan's Home Page - Microsoft Internet Explorer

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

- [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, Special Person mention \(click to go to superhomepage\) \(1\), 2006.](#)
- [Managing Information Extraction, A. Doan, R. Ramakrishnan, V. 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, O. 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.](#)



1991 Tech Reports - Microsoft Internet Explorer

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

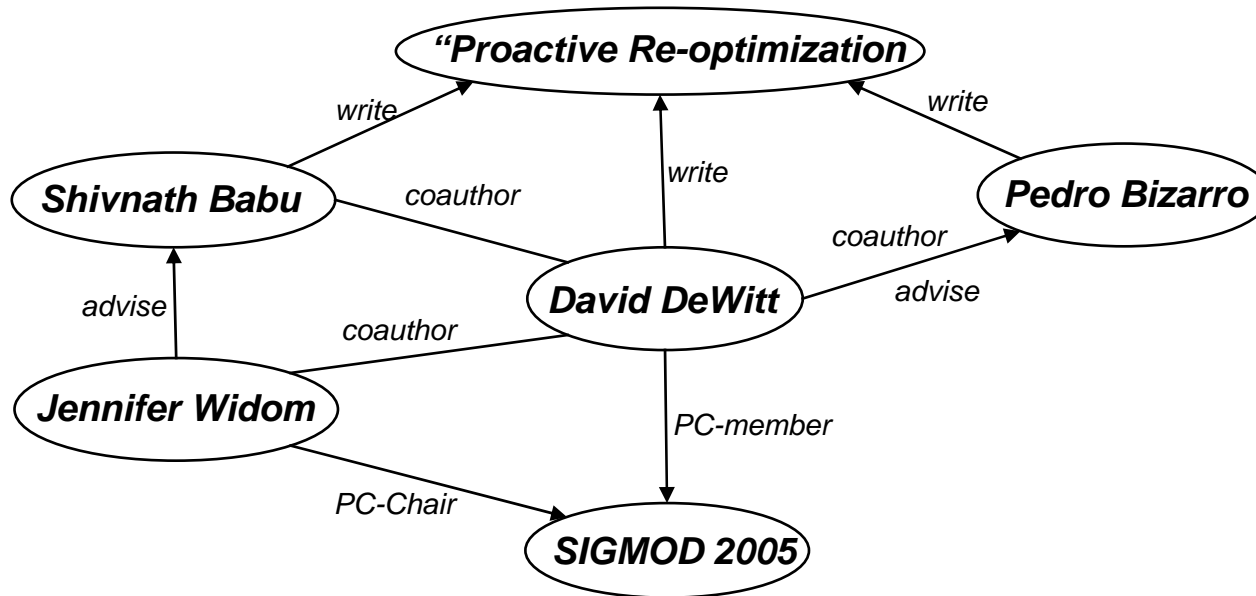
Michael C Ferris and [Olvi L Mangasarian](#)  
[View report information](#) || [Download this report \(PDF\)](#)

- 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

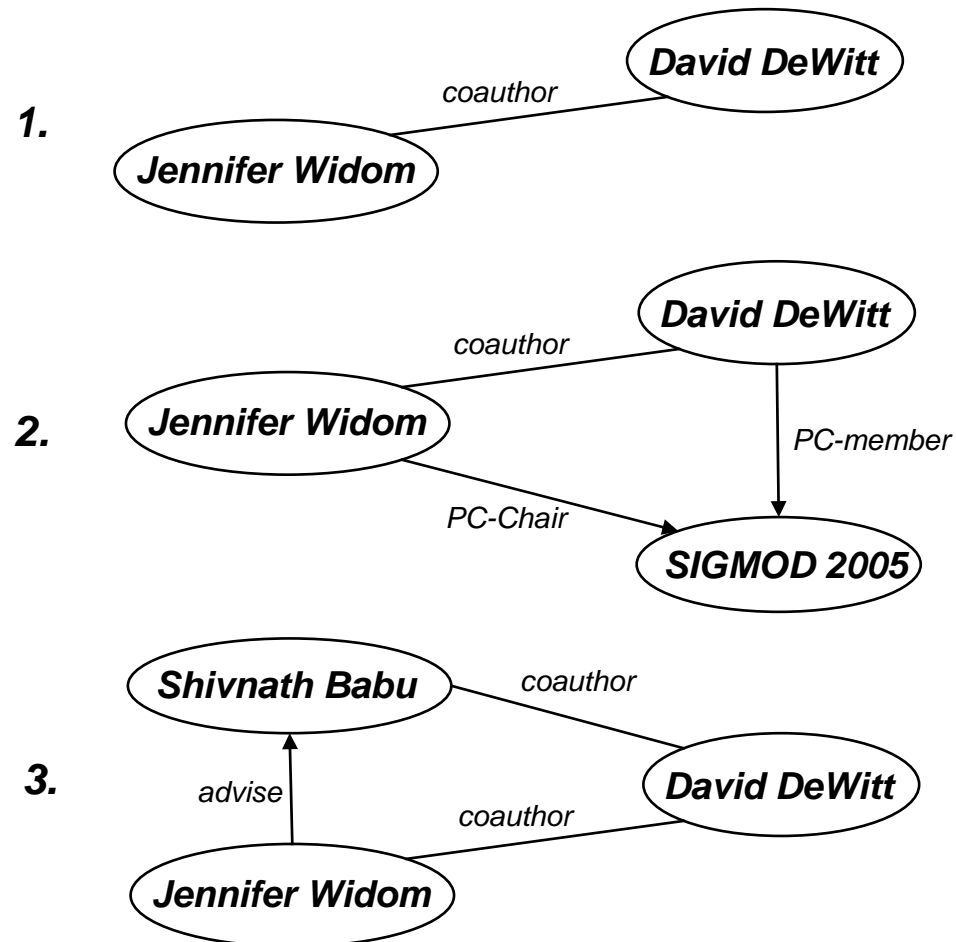
co-authors = A. Doan, Divesh Srivastava, ...

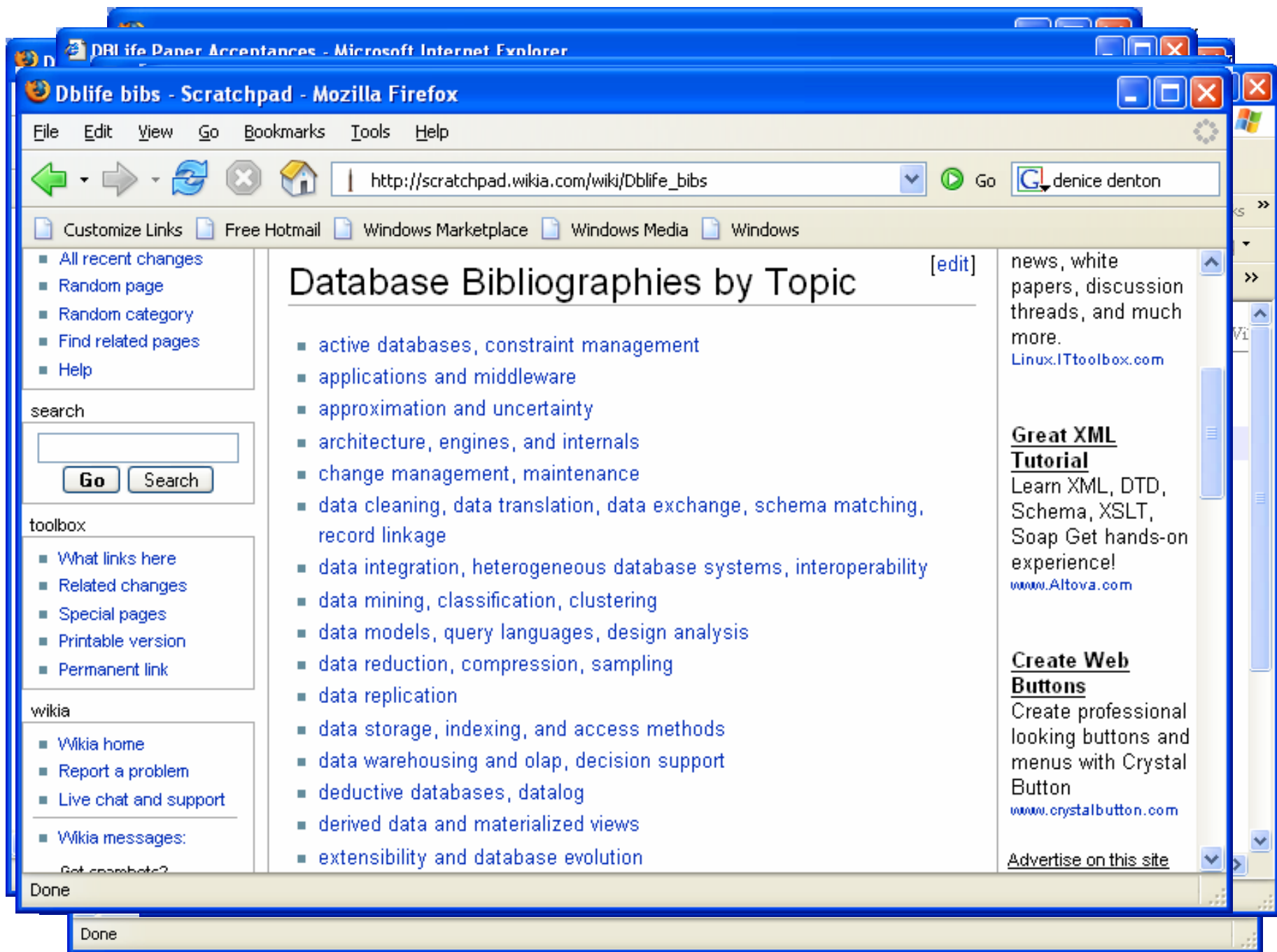
# Resulting ER Graph



# Querying The ER Graph

Query: "David DeWitt Jennifer Widom"



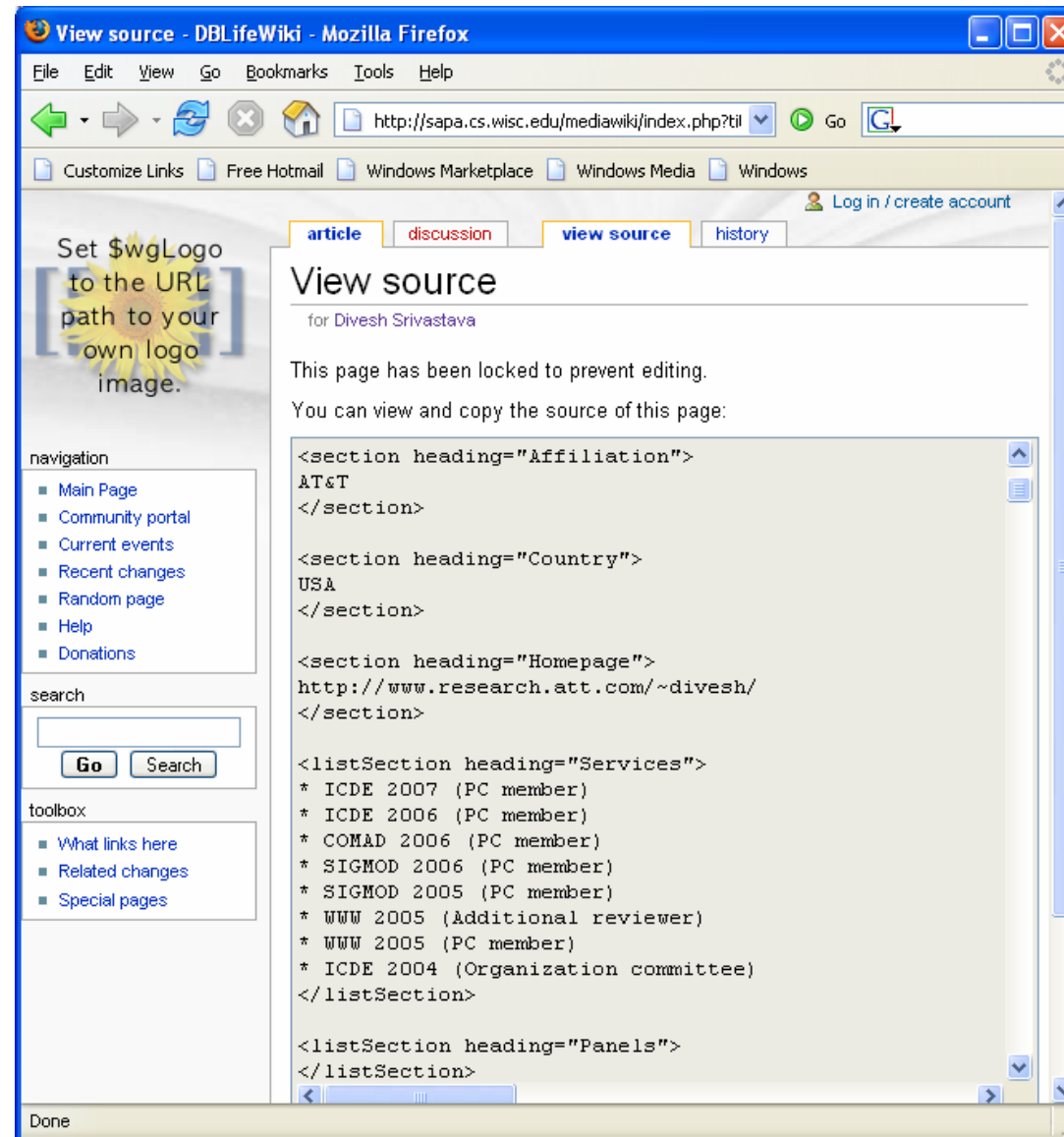


# Mass Collaboration: A Simplified Example



Picture is removed if enough users vote “no”.

# More on Mass Collaboration



The screenshot shows a Mozilla Firefox browser window titled "View source - DBLifeWiki - Mozilla Firefox". The address bar contains the URL "http://sapa.cs.wisc.edu/mediawiki/index.php?tit". The page content includes a navigation menu with links like "Main Page", "Community portal", and "Recent changes". The main content area is titled "View source" and contains the following text:

This page has been locked to prevent editing.  
You can view and copy the source of this page:

```
<section heading="Affiliation">
AT&T
</section>

<section heading="Country">
USA
</section>

<section heading="Homepage">
http://www.research.att.com/~divesh/
</section>

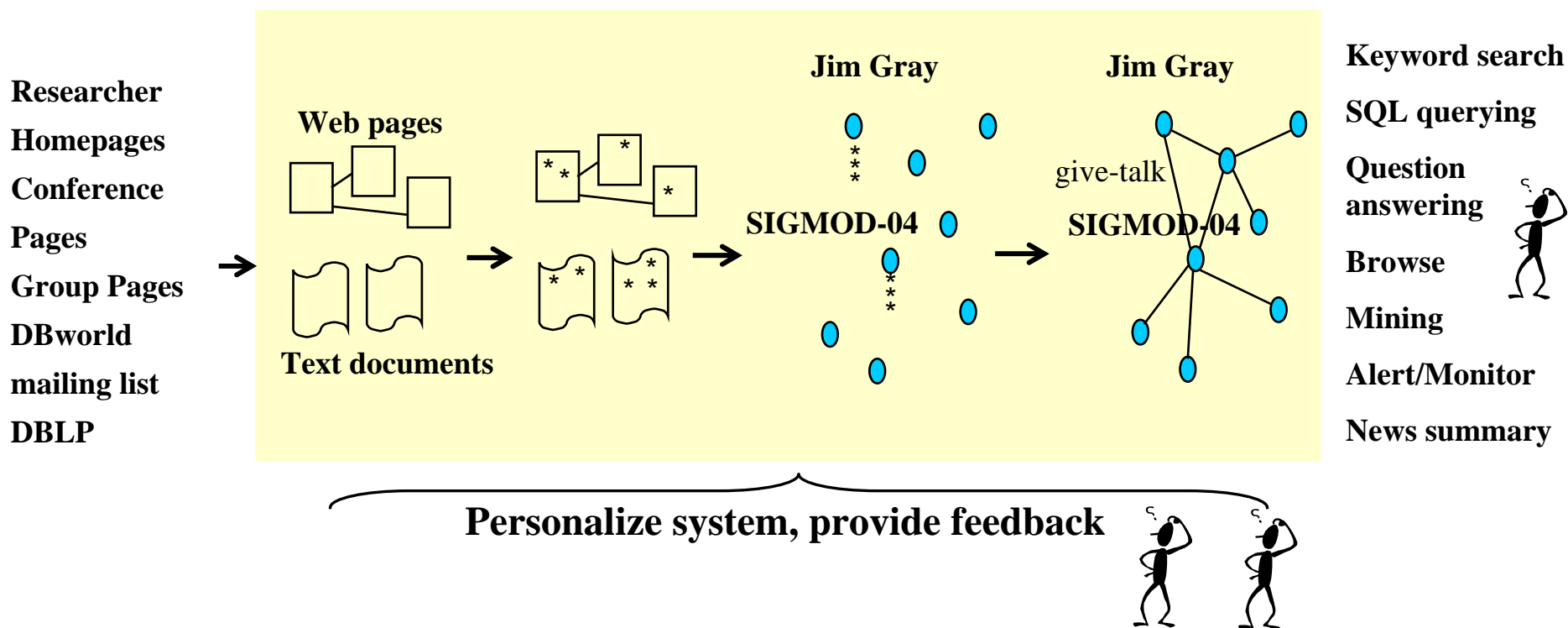
<listSection heading="Services">
* ICDE 2007 (PC member)
* ICDE 2006 (PC member)
* COMAD 2006 (PC member)
* SIGMOD 2006 (PC member)
* SIGMOD 2005 (PC member)
* WWW 2005 (Additional reviewer)
* WWW 2005 (PC member)
* ICDE 2004 (Organization committee)
</listSection>

<listSection heading="Panels">
</listSection>
```

# DBLife: Key Lessons Learned

- **Built relatively simple best-effort integration tools**
- **Combined them in a flexible, bottom-up fashion**
- **System appears already interesting/useful**
  - see [dblifec.s.wisc.edu](http://dblifec.s.wisc.edu) (still very preliminary & slow)
- **Hence possible strategy for best-effort integration:**
  - build relatively simple integration tools
  - learn how to combine them effectively
- **Relative simple integration tools = Lego blocks**
  - easier to build, debug, work with, enable quick tech transfer?
- **Building systems bring much benefits**
  - suggests many interesting / unexpected research challenges
  - helps bridge the research/tech transfer gap

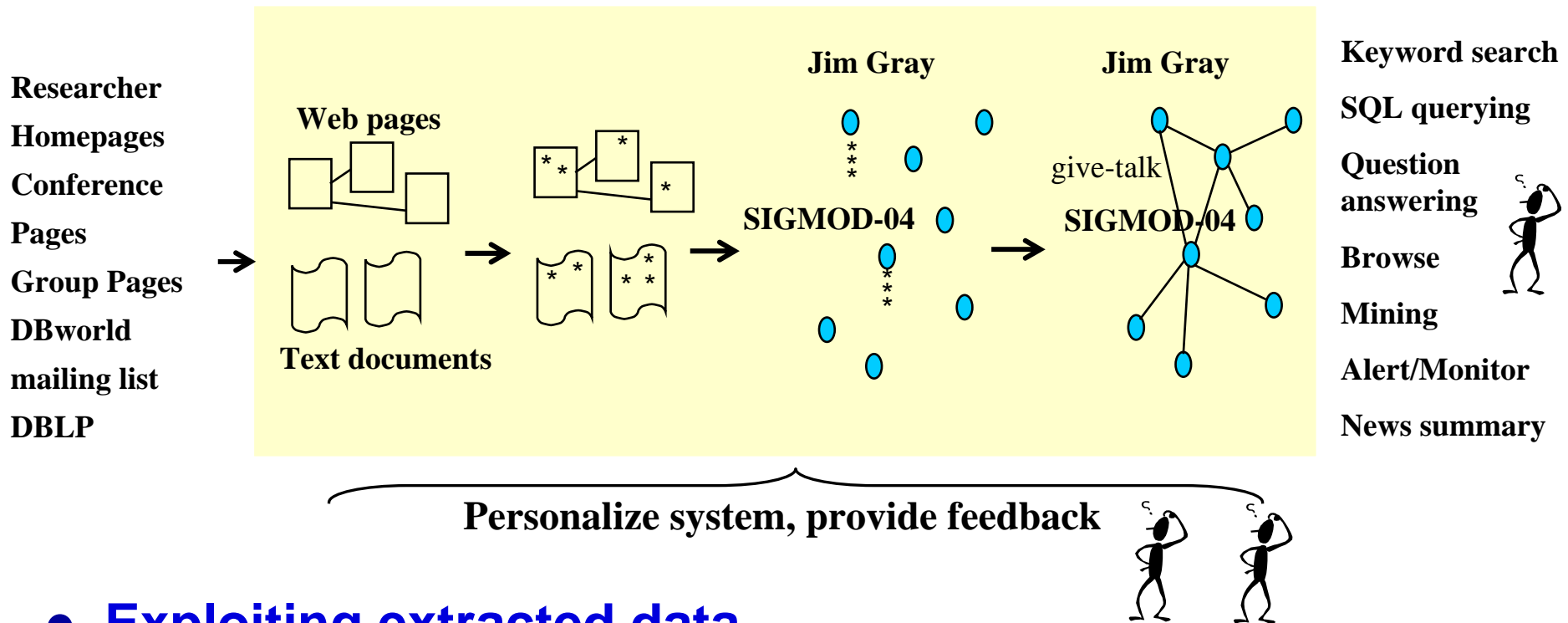
# Research Challenges (1)



- **Information extraction**
- **Data integration**
- **Mass collaboration**

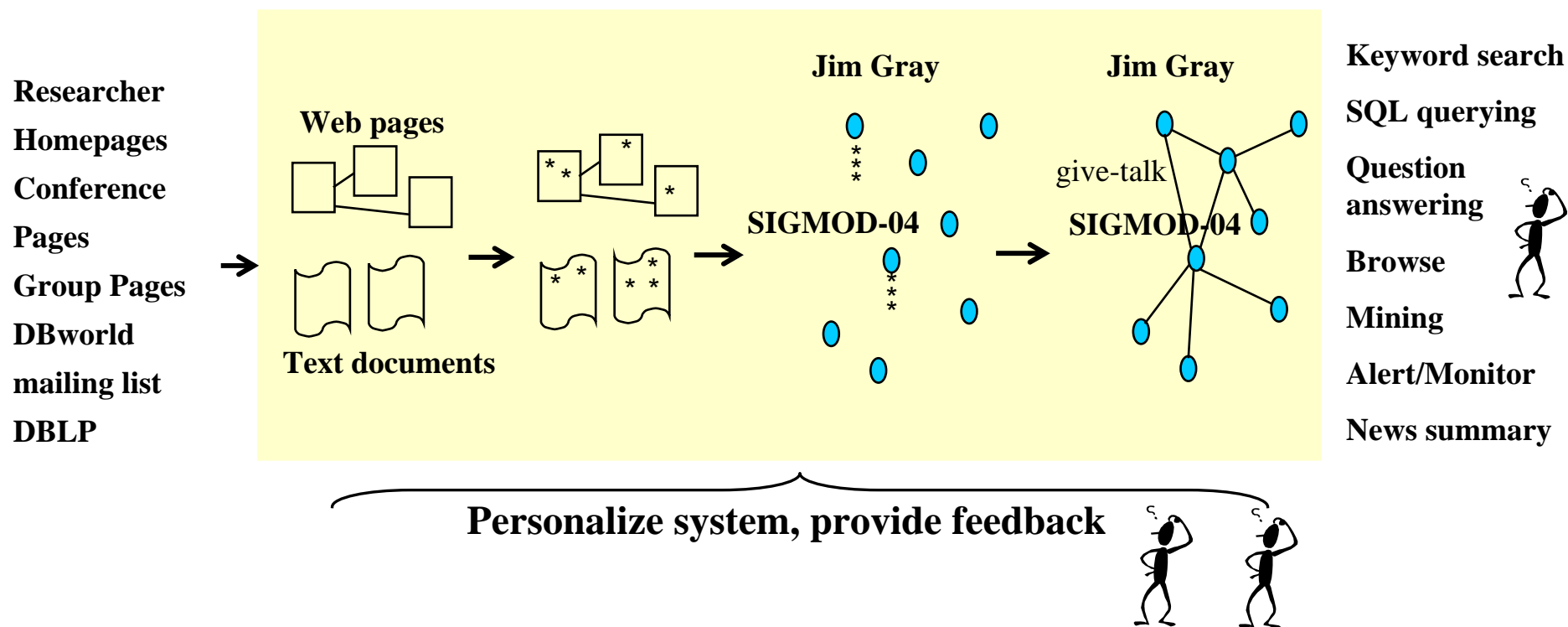
– how to collectively edit extracted and integrated data?

# Research Challenges (2)



- **Exploiting extracted data**
  - keyword search, structured querying, mining, monitoring
  - how to seamlessly transition among these?
- **Handling uncertainty / provenance / explanation**
- **Dealing with evolving data**

# Research Challenges (3)



- **New data model?**
- **Should we use / extend relational databases?**
- **How to build continuously running systems?**

# Summary



- **Precise vs. best-effort data integration**
- **Sample research**
  - keyword search over multiple databases
  - SQL queries over text
  - Cimple project @ Wisconsin/Yahoo! Research
- **The topic is wide open**
- **Our community can contribute much**
- **Prototype system: DBlife**
  - can serve as a data integration challenge / testbed / benchmark
  - potentially provides useful service to our community (as DBWorld+)
  - provides data for researchers (on a variety of topics)

***More details: search “anhai cimple”***

# Mass Collaboration Meets Jeff Naughton

The screenshot shows a Mozilla Firefox browser window titled "DBLife: Superhomepage of David J. DeWitt". The address bar contains the URL "http://dblife.cs.wisc.edu/search.cgi:". The search bar shows the query "david dewitt" and a "Search" button. The search results display the name "David J. DeWitt" with an RSS icon. Below the name is a row of images with vote counts: a woman (2 votes), a group of people (1 vote), a man (3 votes), a man (3 votes), a book cover (3 votes), and a man (1 vote). A yellow callout bubble points to the man with 3 votes, containing the text "Jeffrey F. Naughton swears that this is David J. DeWitt". To the right, there is a "Login" button and a list of related links including "http://www", "Professo", "Compute", "Wiscons", "410 total", and "0 new m". Below the images, there is a section for "Related" links: "Jeffre", "Ragh", "Micha", "Jayav", and "more". At the bottom, there is a "Related" section with links for "xml", "optim", and "perfo". The browser's status bar at the bottom shows "Done".

DBLife: Superhomepage of David J. DeWitt - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://dblife.cs.wisc.edu/search.cgi: Go

Customize Links Free Hotmail Windows Marketplace Windows Media Windows

**DBLife** david dewitt Search Login

**David J. DeWitt**

Check all *non-related* images and click the submit button. [Show All Images](#)

2 votes 1 votes 3 votes 3 votes 3 votes 1 votes

Jeffrey F. Naughton swears that this is David J. DeWitt

Wednesday Sep 27, 2006 ["Shivnath Babu's Home Page", Shivnath Babu's homepage](#)

...the 2006 ACM Intl. Conf. on Management of Data (SIGMOD 2006), June 2006 S. Babu, P. Bizarro, and D. DeWitt. Proactive Re-optimization with Rio Demonstrated at the 2005 ACM Intl. Conf. on Management of Data...

Related

- Jeffre
- Ragh
- Micha
- Jayav
- more

Related

- xml
- optim
- perfo

Done