



# Managing Inconsistency and Uncertainty in Databases



**Renée J. Miller**

University of Toronto

October 2, 2007  
Bertinoro, Italy

Bertinoro Workshop on Information Integration (INFINT)

# A Rosenthal Decomposition

AUTHOR	TITLE
Persephone	Mythology
Pandora	Secrets of



```
<XML>
<AUTHOR>
</AUTHOR>
<TITLE>
</TITLE>
</XML>
```

Clio

*Nice clean problem*

Create mappings that ensure all target tgds (foreign keys) are satisfied.

*Nice elegant solution.*

*Remainder*

Managing data that violates key constraints

*Need a solution...*

# Contributors

- Ariel Fuxman, PhD Thesis
  - *Microsoft Search Labs*
  - ICDT05, SIGMOD05, VLDB05(demo), ICDE06
- Periklis Andritsos
  - *University of Trento*
  - ICDE06
- Elham Fazli and Jiang Du, MS students
- Diego Fuxman, Undergrad Thesis
- Oktie Hassanzadeh, PhD student

# Dirty Databases

- The presence of dirty data is known to be a major problem in enterprises
  - ▣ “A quarter of data held in commercial enterprises is flawed” (Gartner 2004)
  - ▣ “Three quarters of financial service providers are making business decisions based on suboptimal data” (Gartner 2003)

**Without getting a warning from their database systems**

# Querying a Dirty Database

- Fundamental assumption underlying traditional database systems
  - ▣ Data is clean and consistent
- If we break this assumption
  - ▣ Query results may not be meaningful
- Current Solution
  - ▣ Data Cleaning

# Limitations of Data Cleaning

- Semi-automatic process
  - ▣ Requires highly-qualified domain experts
- Time consuming
  - ▣ May not be possible to wait until the database is clean

# Our Proposal

- A complementary approach to data cleaning
- Extend database systems to obtain meaningful answers from **potentially dirty** databases
  - ▣ Semantics of “dirty data”
  - ▣ Semantics of “meaningful answer”
  - ▣ Efficient query processing

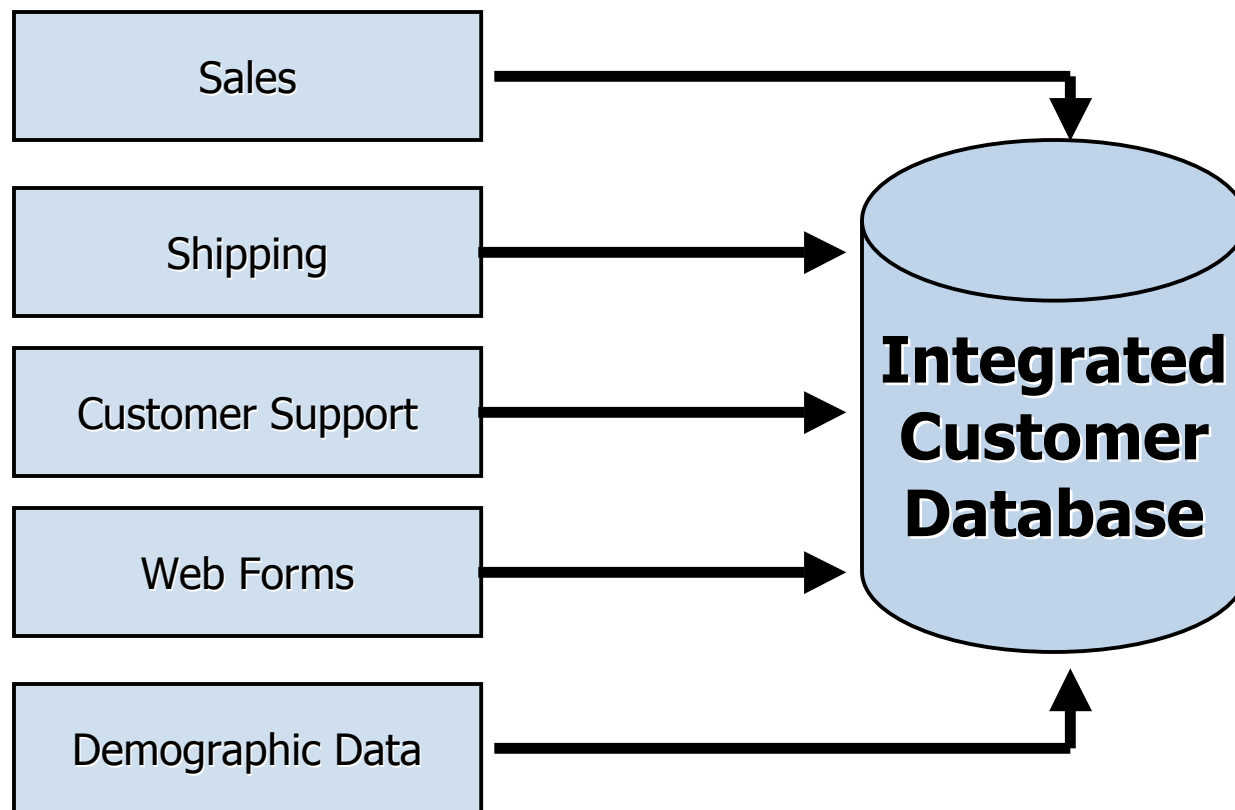
# Outline

- **Data Integration Example**
- Semantics
- Contributions
- Experimental Evaluation
- Future Challenges



# A Data Integration Example

Integrating customer data...



# Matching and Merging

- Matching is well supported by tools
  - ▣ Industrial tools (e.g., IBM's QualityStage, SQL Server Integration Services)
  - ▣ Research work - approximate joins
- Less tool support for **merging**

<i>custid</i>	<i>name</i>	<i>address</i>	<i>...</i>	<i>income</i>
<b>Peter</b>	Peter Yarrow	276 College Street	....	40K
	Paul Stookey	100 Bloor Street	...	400K
	Mary Travers	20 Union Street	...	110K

*web*

<i>custid</i>	<i>name</i>	<i>address</i>	<i>...</i>	<i>income</i>
<b>Peter</b>	P. Yarrow	276 College St.	....	200K
	P. Stookey	100 Bloor St.	...	400K
	M. Travers	20 Union St.	...	130K

*sales*

# Merging (Resolution)

- Conflict resolution rules may be difficult to design
- True disagreement between sources

<i>custid</i>	<i>name</i>	<i>address</i>	...	<i>income</i>
Peter	Peter Yarrow	276 College Street	....	40K
	Paul Stookey	100 Bloor Street	...	400K
	Mary Travers	20 Union Street	...	110K

*web*

<i>custid</i>	<i>name</i>	<i>address</i>	...	<i>income</i>
Peter	Peter Yarrow	276 College Street	....	200K
	P. Stookey	100 Bloor St.	...	400K
	M. Travers	20 Union St.	...	130K

*sales*

# Inconsistent Integrated Databases

In the absence of resolution rules...

*SATISFY custid KEY*

*VIOLATES custid KEY*

*Web*

<i>custid</i>	...	<i>income</i>
Peter	...	40K
Paul	...	400K
Mary	...	110K

*Sales*

<i>custid</i>	...	<i>income</i>
Peter	...	200K
Paul	...	400K
Mary	...	130K

*Inconsistent Integrated Database*

<i>custid</i>	...	<i>income</i>
Peter	...	40K
Peter	...	200K
Paul	...	400K
Mary	...	110K
Mary	...	130K

# Querying an Inconsistent Database

*Example: Offering a Platinum credit card...*

$q_1$  = "Get customers who make more than 100K"

Peter, Paul, Mary

**Are we sure that we want to offer a card to Peter?**

<u>custid</u>	income	
Peter	40K	web
Peter	200K	sales
Paul	400K	sales/web
Mary	110K	web
Mary	130K	sales

# Querying an Inconsistent Database

- **Aggressive:** Get customers who **possibly** make more than 100K
  - ▣ *Peter, Paul, Mary*
- **Conservative:** Get customers who **certainly** make more than 100K
  - ▣ *Paul, Mary*

<i>custid</i>	<i>income</i>	
Peter	40K	<i>web</i>
Peter	200K	<i>sales</i>
Paul	400K	<i>sales/web</i>
Mary	110K	<i>web</i>
Mary	130K	<i>sales</i>

## Answer

Peter	possibly
Paul	certainly
Mary	certainly

# Outline

- Data Integration Example
- **Semantics**
- Contributions
- Experimental Evaluation
- Future Challenges

# Formal Semantics

- Models for incomplete data
  - ▣ [Imielinski Lipski 84]
- Querying incomplete data
  - ▣ [Imielinski Lipski 84, Abiteboul Duschka 98]
  - ▣ Possible world: “complete” database
- Querying inconsistent data
  - ▣ [Arenas Bertossi Chomicki 99]
  - ▣ Possible world: (maximal) “consistent” database
  - ▣ **Consistent answers:** conservative semantics



# Consistent Answers

## Inconsistent database

<i>custid</i>	<i>income</i>	
Peter	40K	<i>web</i>
Peter	200K	<i>sales</i>
Paul	400K	<i>sales/web</i>
Mary	110K	<i>web</i>
Mary	130K	<i>sales</i>

Key: *custid*

## Repairs

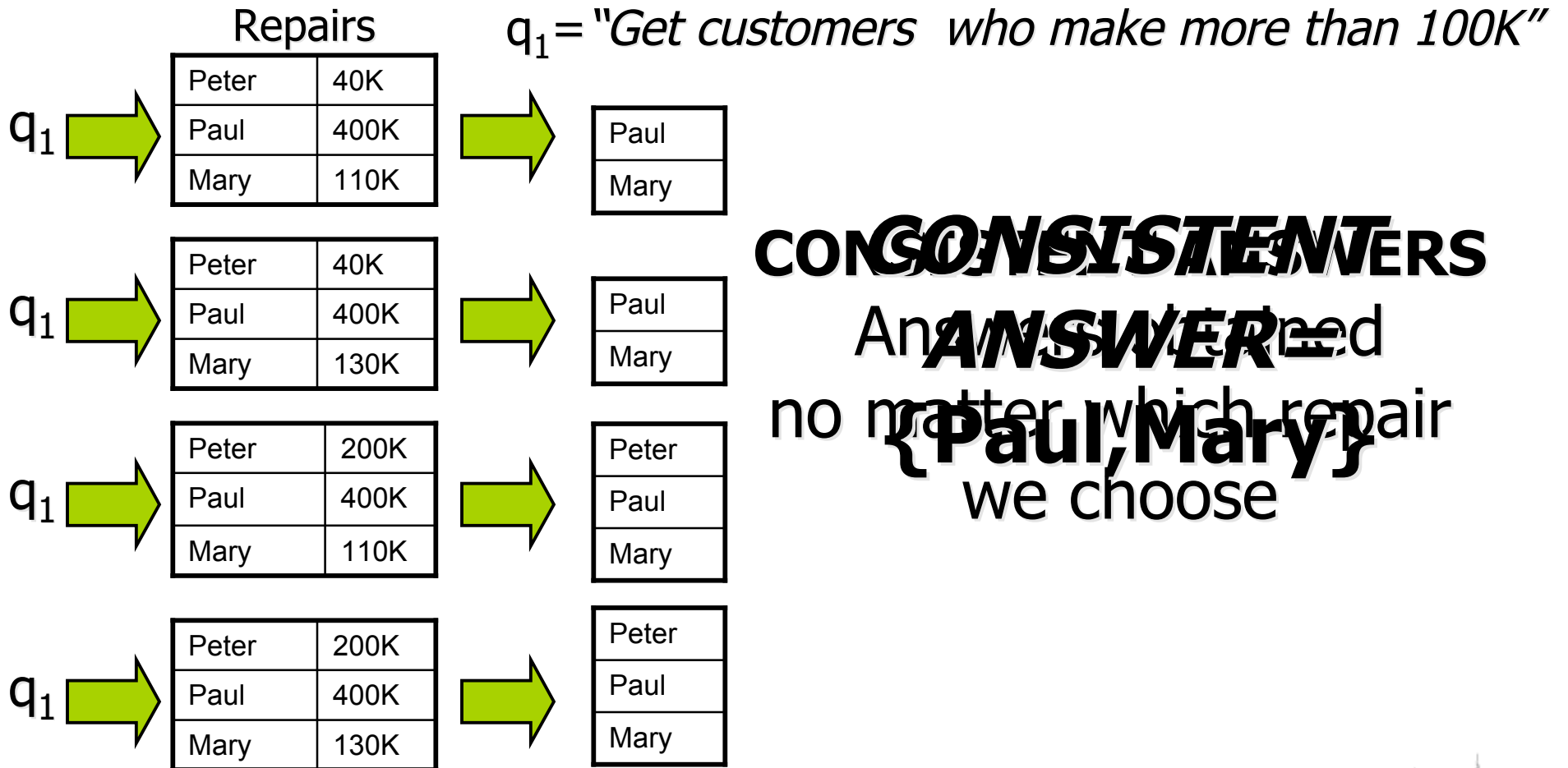
Peter	40K
Paul	400K
Mary	110K

Peter	40K
Paul	400K
Mary	130K

Peter	200K
Paul	400K
Mary	110K

Peter	200K
Paul	400K
Mary	130K

# Consistent Query Answers



# Problem

- Potentially **HUGE** number of repairs!
- Plenty of negative results
  - ▣ [Chomicki et al 02, Arenas et al. 01, Cali et al 04]
- Few tractability results
  - ▣ [Arenas et al. 99, Arenas et al. 01]
- Computation based on logic programming
  - ▣ [Bravo and Bertossi 03, Eiter et al. 03, see Rosati talk]
  - ▣ Consider rich sets of constraints
  - ▣ May be computationally expensive
  - ▣ Focus (mostly) small number conflicts

# Bridging Theory and Practice

Are these ideas applicable to realistic scenarios?

**YES!**

Negative results do not rule out many realistic scenarios, such as:

- ▣ Key constraints
- ▣ Commonly used decision support queries

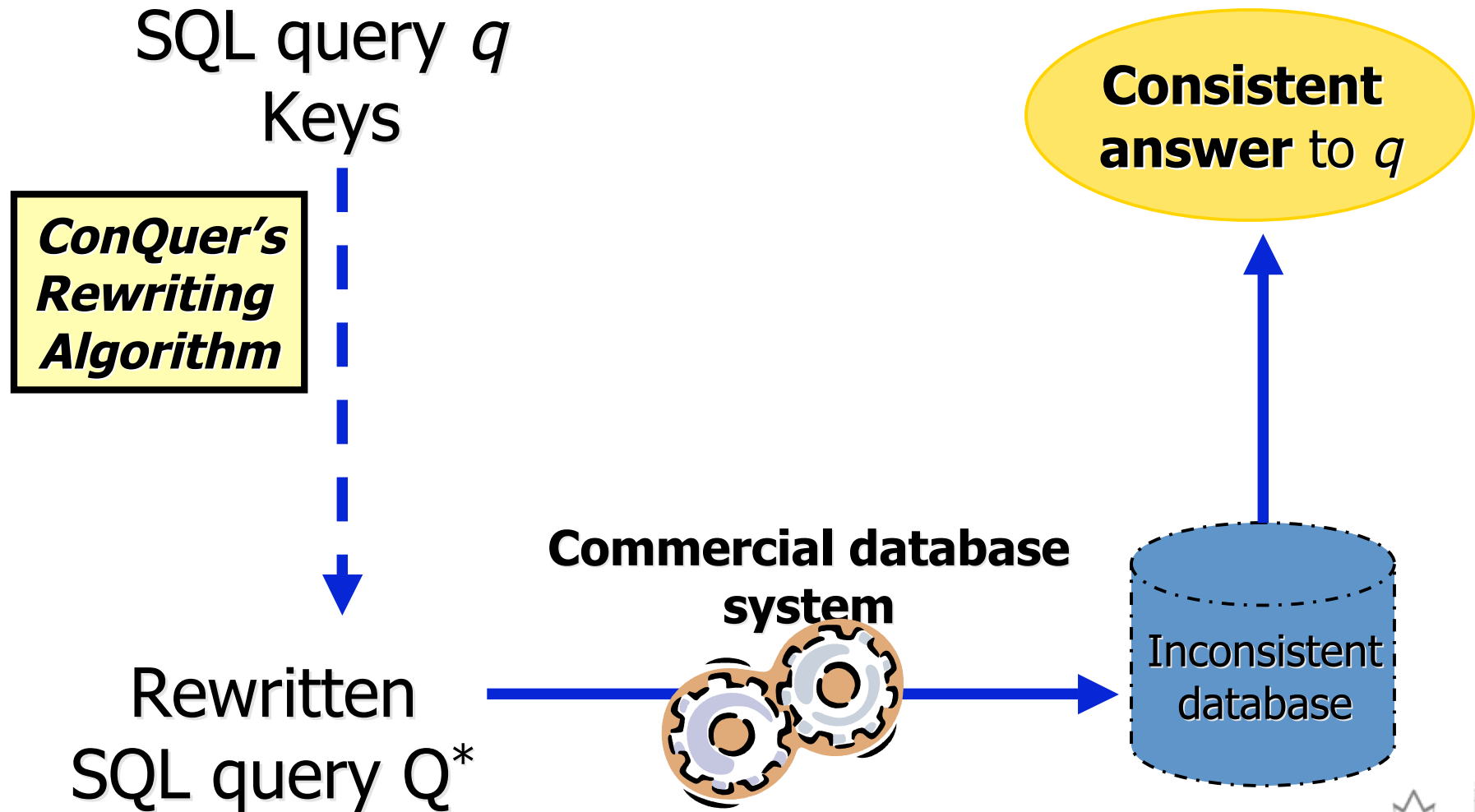
# Outline

- Data Integration Example
- Semantics
- **Contributions**
- Experimental Evaluation
- Future Challenges

# Contributions

- Formal characterization of a class of queries for which computing consistent answers is tractable **[ICDT05-JCSS07]**
- Algorithms for efficient computation of consistent answers using SQL rewriting **[SIGMOD05]**
- A probabilistic semantics that generalizes consistent answers **[ICDE06]**
- *ConQuer*, a system designed to compute consistent answers efficiently **[VLDB05 demo]**

# ConQuer



# A Large Class of Queries

- ConQuer supports SQL queries with
  - ▣ Joins between keys
  - ▣ Joins between non-keys and keys
  - ▣ Arbitrary number of relations
  - ▣ Set and bag semantics
  - ▣ Grouping and aggregation
- Showed (experimentally) applicable to
  - ▣ Large instances (50G)
  - ▣ Large number of conflicts (50% dirty data)
- Restrictions
  - ▣ No joins between non-keys (unless in SELECT)
  - ▣ Acyclic non-key to key structure (no self joins)
  - ▣ Subqueries must be decorrelated



# Uncertain Data

PROVENANCE INFORMATION  
(e.g., source reputation)

**0.3**

*Web*

<i>custid</i>	...	<i>income</i>
Peter	...	40K
Paul	...	400K
Mary	...	110K

*Integrated Database*

<i>custid</i>	...	<i>income</i>
Peter	...	40K
Peter	...	200K
Paul	...	400K
Mary	...	110K
Mary	...	130K

**0.3**

**0.7**

**1**

**0.3**

**0.7**

**0.7**

*Sales*

<i>custid</i>	...	<i>income</i>
Peter	...	200K
Paul	...	400K
Mary	...	130K

# Probabilistic Semantics

- Probabilistic model based on the basic assumptions of repairs
  - ▣ Tuples that share the same key value do not appear in the same repair (mutually exclusive)
  - ▣ Tuples that do not share a key value are chosen independently
- Probability distribution over repairs
- Query results are assigned a “confidence” of being obtained from the clean database

# Uncertain Inconsistent DB

- An uncertain inconsistent DB assumes
  - ▣ Alternatives for a key value are mutually exclusive
  - ▣ Tuples for different key values are independent
- Presented rewriting for computing consistent answers with probabilities [ICDE 06]
  - ▣ Again focus was on scalability/execution of rewriting
- Since our work such databases have also been considered by Dalvi and Suciu [PODS 07] invited talk
  - ▣ “Disjoint-independent” databases
  - ▣ Presents classification of queries that extends ours

# Outline

- Data Integration Example
- Semantics
- Contributions
- **Experimental Evaluation**
- Future Challenges

# Overview of Experimental Results

- Showed scalability of rewritings on
  - ▣ TPC-H queries
  - ▣ Up to 50G with varying degree and type of inconsistency
  - ▣ Overhead
    - Getting consistent and possible answers for most queries takes 2-3 times as long as getting possible answers alone

# Outline

- Data Integration Example
- Semantics
- Contributions
- Experimental Evaluation
- **Future Challenges**

# Future Challenge 0

- Our approach does not work for ALL queries
  - ▣ Our rewriting is sound but it may “miss” some consistent answers for queries outside our class (that is for “hard” queries)
  - ▣ Hard queries are only hard for some instances
  - ▣ Can we approximate consistent answers for all queries? (Apply another Rosenthal decomposition...)

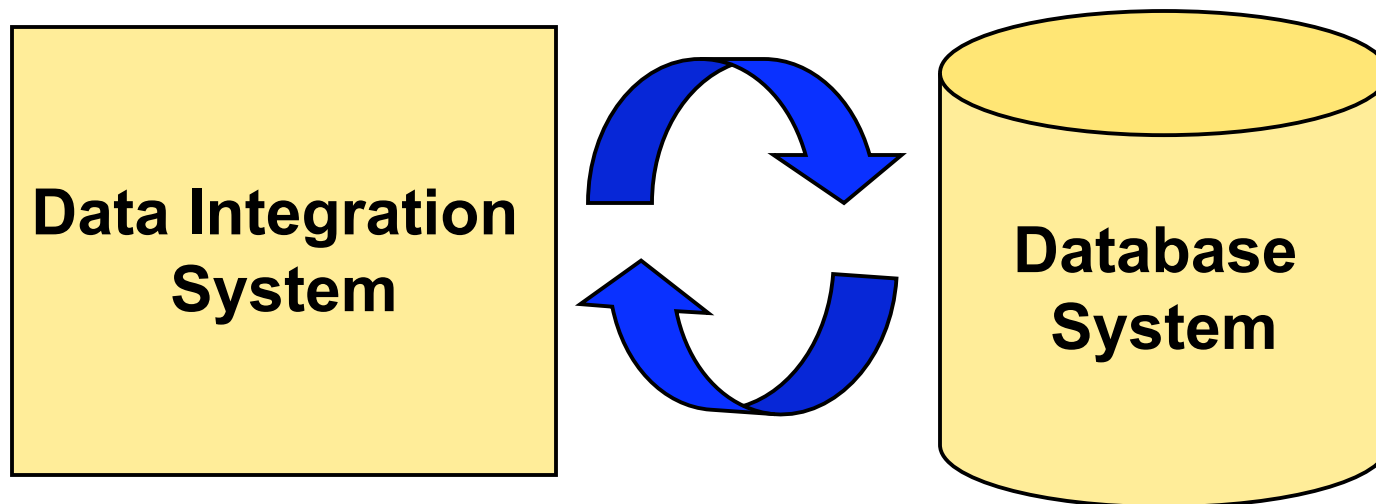
# Future Challenge 1

- How do we meaningfully assign probabilities to alternative, inconsistent tuples?
  - ▣ Is it enough to use output of a similarity measure (or approximate join technique)?
  - ▣ What probability assignments lead to the most meaningful query answers?



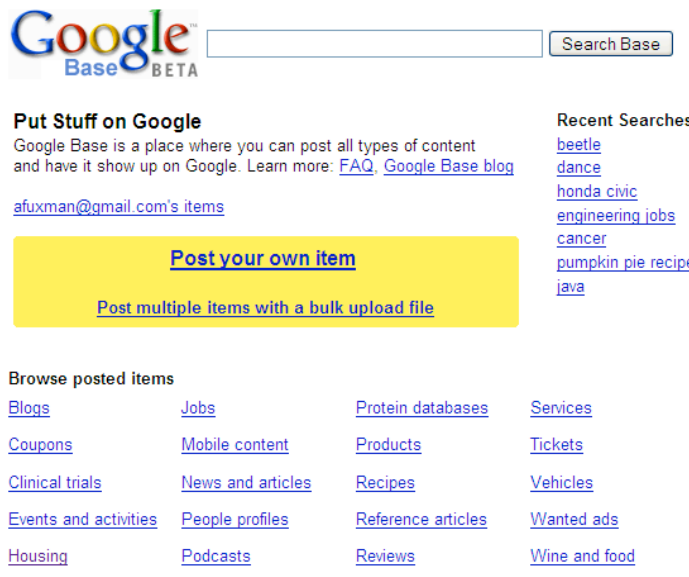
# Future Challenge 2

- Promote querying dirty data as an integral part of mainstream data integration and database systems



# Future Challenge 3

- Bring the paradigm of querying dirty data to new domains, including Web applications



**Google Base BETA**

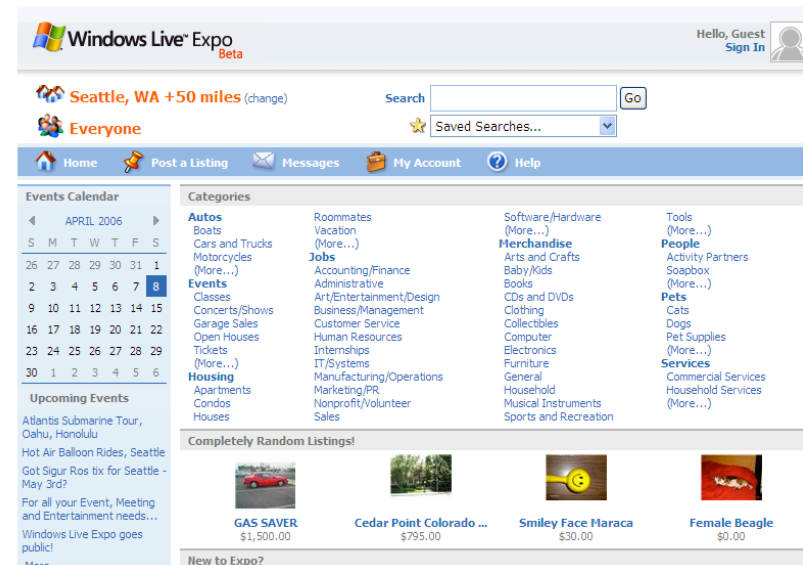
**Put Stuff on Google**  
Google Base is a place where you can post all types of content and have it show up on Google. Learn more: [FAQ](#), [Google Base blog](#)  
[afuxman@gmail.com's items](#)

**Recent Searches**  
[beetle](#)  
[dance](#)  
[honda civic](#)  
[engineering jobs](#)  
[cancer](#)  
[pumpkin pie recipe](#)  
[java](#)

**Post your own item**  
[Post multiple items with a bulk upload file](#)

**Browse posted items**

<a href="#">Blogs</a>	<a href="#">Jobs</a>	<a href="#">Protein databases</a>	<a href="#">Services</a>
<a href="#">Coupons</a>	<a href="#">Mobile content</a>	<a href="#">Products</a>	<a href="#">Tickets</a>
<a href="#">Clinical trials</a>	<a href="#">News and articles</a>	<a href="#">Recipes</a>	<a href="#">Vehicles</a>
<a href="#">Events and activities</a>	<a href="#">People profiles</a>	<a href="#">Reference articles</a>	<a href="#">Wanted ads</a>
<a href="#">Housing</a>	<a href="#">Podcasts</a>	<a href="#">Reviews</a>	<a href="#">Wine and food</a>



**Windows Live Expo Beta** Hello, Guest

Seattle, WA +50 miles (change)    
Everyone




[Home](#) [Post a Listing](#) [Messages](#) [My Account](#) [Help](#)

**Events Calendar**  
4 APRIL 2006  
S M T W T F S  
26 27 28 29 30 31 1  
2 3 4 5 6 7 8  
9 10 11 12 13 14 15  
16 17 18 19 20 21 22  
23 24 25 26 27 28 29  
30 1 2 3 4 5 6  
Upcoming Events  
Atlantis Submarine Tour, Oahu, Honolulu  
Hot Air Balloon Rides, Seattle  
Got Sigur Ros tix for Seattle - May 3rd?  
For all your Event, Meeting and Entertainment needs...  
Windows Live Expo goes public!  
More.....

**Categories**

<b>Autos</b> Boats Cars and Trucks Motorcycles (More...)	<b>Jobs</b> Accounting/Finance Administrative Art/Entertainment/Design Business/Management Customer Service Human Resources Internships IT/Systems Manufacturing/Operations Marketing/PR Nonprofit/Volunteer Sales	Roommates Vacation (More...) <b>Merchandise</b> Arts and Crafts Baby/Kids Books CDs and DVDs Clothing Collectibles Computer Electronics Furniture Manufacturing/Operations Musical Instruments Sports and Recreation	Tools (More...) <b>People</b> Activity Partners Soapbox (More...) <b>Pets</b> Cats Dogs Pet Supplies (More...) <b>Services</b> Commercial Services Household Services (More...)
---	--	--	---

**Completely Random Listings!**





 <b>GAS SAVER</b> \$1,500.00	 <b>Cedar Point Colorado ...</b> \$795.00	 <b>Smiley Face Maraca</b> \$30.00	 <b>Female Beagle</b> \$0.00
---	--	---	---

Now to Expo?



# Future Challenge 3

The screenshot shows a Windows Live Expo interface. At the top, the logo reads "Windows Live™ Expo Beta". Below this is a search bar containing the text "SAN FRANCISCO, ARGENTINA" in red, with a "Go" button to its right. A navigation bar includes links for "Home", "Post a Listing", "Messages (0)", "My Account", and "Help". The listing details are as follows:

Date listed: 4/9/2006	 Map Listing	 Username: abcd Member since: 3/25/2006 <a href="#">View profile</a>	 <a href="#">Edit Listing</a>  <a href="#">Close Listing</a>
Expiration: 5/7/2006			
Price: <b>Unknown</b>			

Category: [Housing](#) > [Houses](#) > **FOR RENT**

# Questions?

---

Web Page:

[www.cs.toronto.edu/~miller](http://www.cs.toronto.edu/~miller)