

Dataspaces: The Next Frontier for Data Integration

Alon Halevy



Shannon Lecture Series

April 26, 2007

Agenda

- Data integration:
 - Connecting disparate data sources
 - Great progress in last decade
- But we're still missing the point:
 - Dataspaces: a new abstraction
- A few connections to my Google work
- Predictions, subliminal messages (perhaps)



Abstractions 'R Us

- *Logical vs. Physical; What vs. How.*

Students:

SSN	Name	Category
123-45-6789	Charles	undergrad
234-56-7890	Dan	grad
...

Takes:

SSN	CID
123-45-6789	CSE444
123-45-6789	CSE444
234-56-7890	CSE142
...	...

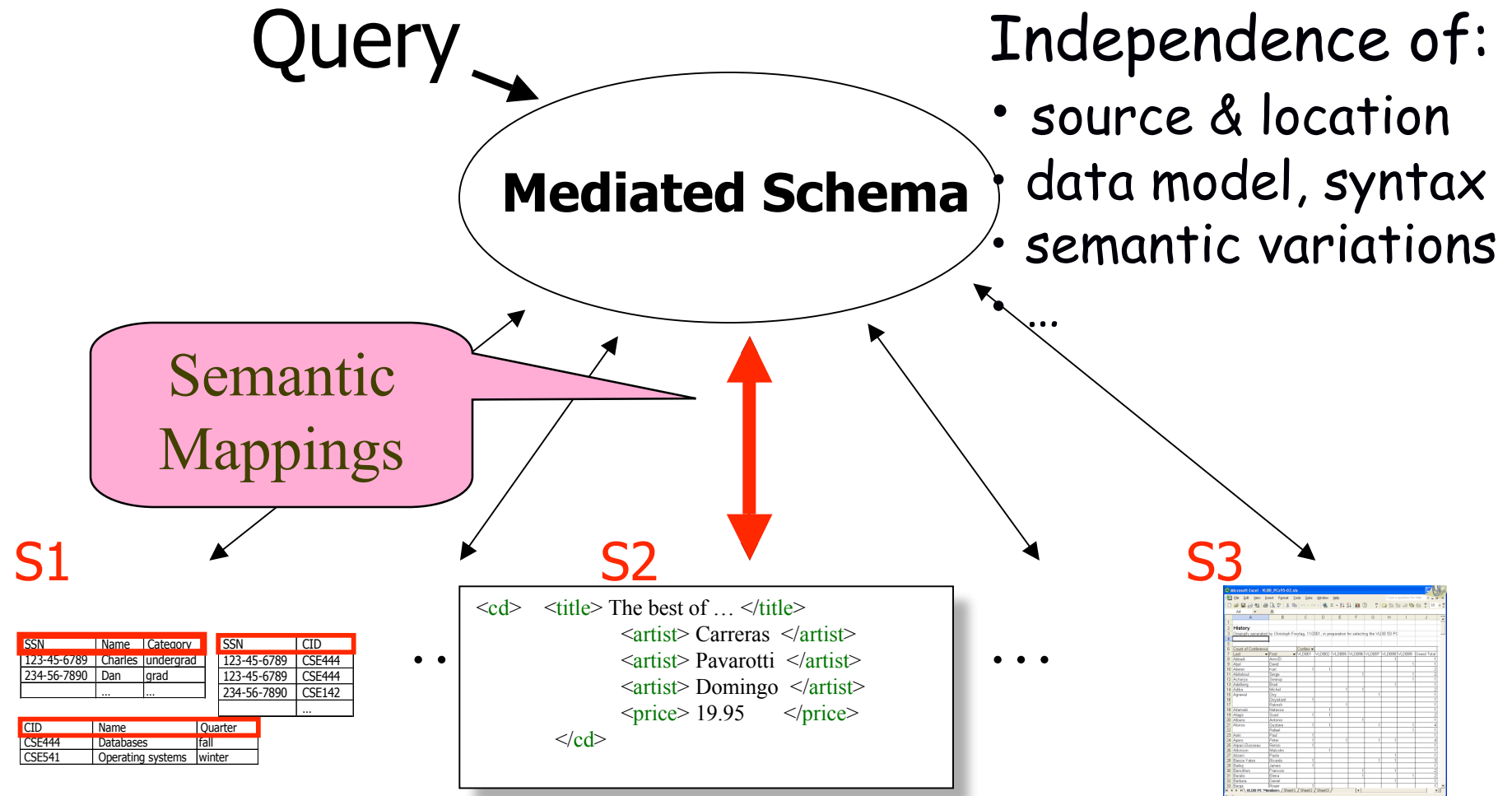
Courses:

CID	Name	Quarter
CSE444	Databases	fall
CSE541	Operating systems	winter

```
SELECT C.name
FROM Students S, Takes T, Courses C
WHERE S.name="Mary" and
      S.ssn = T.ssn and T.cid = C.cid
```



Data Integration: A Higher-level Abstraction



Data Integration

Kayak.com Search Results - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.kayak.com/s/flights.jsp?searchid=\$cnOJdldGolEqci6NLI

Travel Center Crossing the Structur... Schedule a Conferen... Main - Scuba Dashboard Sign out John Battelle's Searc...

Google sukka Search PageRank ABC Check AutoLink AutoFill

Best Fare Trend

90 Days Ago Today

Stops

nonstop 1 stop 2+ stops

Airlines

[select all](#) | [clear](#) **nonstop** **1+**

<input checked="" type="checkbox"/> Air France	\$10486
<input checked="" type="checkbox"/> All Nippon	\$1630
<input checked="" type="checkbox"/> American Airlines	\$2707
<input checked="" type="checkbox"/> Cathay Pacific	\$2845
<input checked="" type="checkbox"/> China Airlines	\$1483
<input checked="" type="checkbox"/> EVA Corporation	\$1453
<input checked="" type="checkbox"/> Multiple Airlines	\$2044 \$1423
<input checked="" type="checkbox"/> Singapore Airlines	\$1796
<input checked="" type="checkbox"/> United	\$2291 \$2170
<input checked="" type="checkbox"/> US Airways	\$2242

Fare	Route	Airline	Class	Time	Stops	Duration
\$1423	SFO > ICN	Asiana Airlines	1:33p	5:40p	0	(12h 07m)
	ICN > SFO	Multiple Airlines	5:30p	3:35p	1	(14h 05m)
						orbitz: \$1423 details email
\$1423	SFO > ICN	Asiana Airlines	1:33p	5:40p	0	(12h 07m)
	ICN > SFO	Multiple Airlines	5:30p	4:13p	1	(14h 43m)
						orbitz: \$1423 details email
\$1423	SFO > ICN	Asiana Airlines	1:33p	5:40p	0	(12h 07m)
	ICN > SFO	Multiple Airlines	5:30p	5:25p	1	(15h 55m)
						orbitz: \$1423 details email
\$1423	SFO > ICN	Asiana Airlines	1:33p	5:40p	0	(12h 07m)
	ICN > SFO	Multiple Airlines	5:30p	5:50p	1	(16h 20m)
						orbitz: \$1423 details email
\$1453	SFO > ICN	EVA Corporation	1:30a	6:10p	1	(24h 40m)
	ICN > SFO	EVA Corporation	7:15p	7:55p	1	(16h 40m)
						airfare: \$1453 details email
\$1483	SFO > ICN	China Airlines	1:05a	12:00p	1	(18h 55m)

Done Open Notebook

Phen

ment

array
ment

ons

Wikipedia:



A **mashup** is a website or [Web 2.0](#) application that uses content from more than one source to create a completely new service. This is akin to [transclusion](#).





Zoomf.com property search

Review Link Rate Link Bookmark Link Remove Frame

order: highest price | < previous page | next page >

R55

show nearby

Results 1 to 10 of 56 - Property listings for sale - EC1

£475,000 [EC1](#)
 3 bedroom 1 bathroom Flat
 Very large and spacious 3 double bedroom apartment spread over the 1st to 3rd floors of a late Victorian property in the heart of Clerkenwell. Comprising of large op...
 Source: *Winkworth* [show on map](#)



£450,000 [EC1](#)
 2 bedroom 2 bathroom Flat
 A 893 sq ft, two bedroom, two bathroom, modern apartment, located on the first floor of Herbal Hill Gardens. The apartment also comprises large reception, modern fit...
 Source: *Winkworth* [show on map](#)



£319,950 [EC1](#)
 1 bedroom 1 bathroom Flat
 A well proportioned modern second floor one bedroom apartment located at 201 St. Johns Street. The one bedroom apartment comprises reception room, modern fitted kitc...
 Source: *Winkworth* [show on map](#)



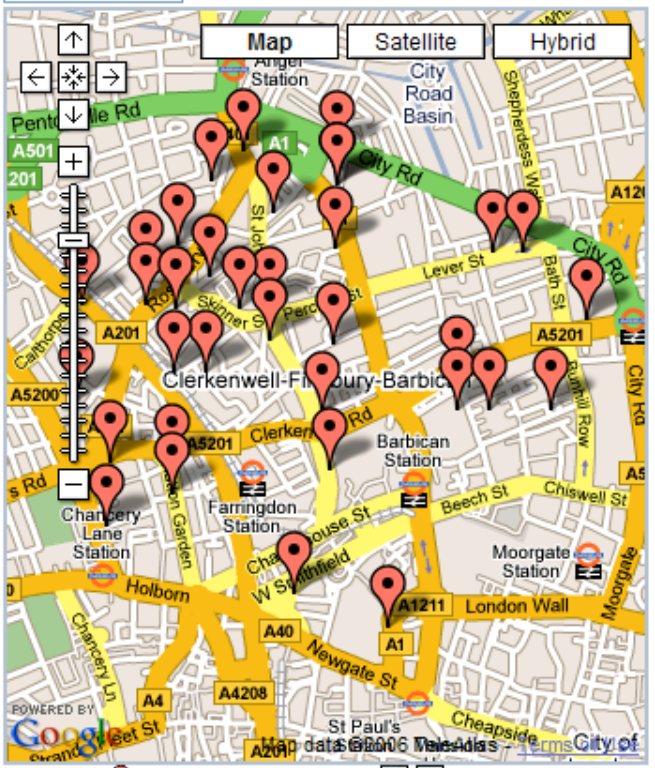
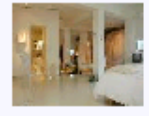
£309,000 [EC1](#)
 1 bedroom 1 bathroom Flat
 A large one bedroom, lower ground floor garden flat, in a Grade II listed, Georgian terrace on Myddelton Square. The property comprises large open plan reception wit...
 Source: *Winkworth* [show on map](#)



£250,000 [EC1](#)
 3 bedroom 1 bathroom Flat
 A large split level bedroom maisonette situated on the 4th floor of this low rise ex-local Authority estate just moments from Old Street and well located for both th...
 Source: *Winkworth* [show on map](#)



£995,000 [St John Street \(EC1\)](#)
 2 bedroom 1 bathroom Flat
 Incorporating striking interior design and quirky living and entertaining space, this utterly unique two bedroomed apartment provides the ultimate in style and reall...
 Source: *Foxtons* [show on map](#)



*Click on [pin] to see details and use [zoom in] [zoom out] to zoom in and out.

Why is it Hard?

- **Systems reasons:**
 - Managing different platforms
 - Query processing across multiple systems
- **Social reasons:**
 - Locating and capturing relevant data in the enterprise.
 - Convincing people to share (*data fiefdoms*)
 - Privacy and performance implications.
- **Logic reasons:**
 - Schema (and data) heterogeneity
 - *Challenge independent of integration architecture!*



Design time

Run time



Mediated Schema

mediation language

query reformulation

mapping tool

optimization & execution

XML

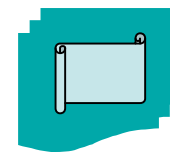
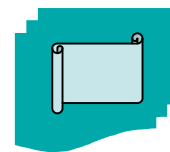
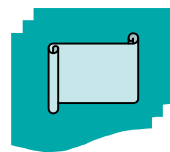
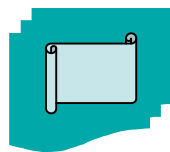
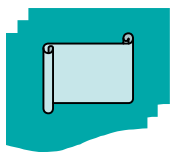
wrapper

wrapper

wrapper

wrapper

wrapper





Design time

Run time

Mediated Schema

mediation language

mapping tool

query reformulation

optimization & execution

XML

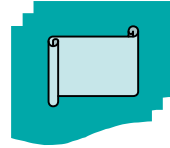
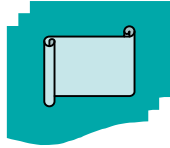
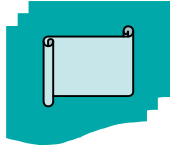
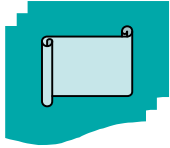
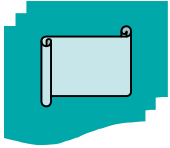
wrapper

wrapper

wrapper

wrapper

wrapper



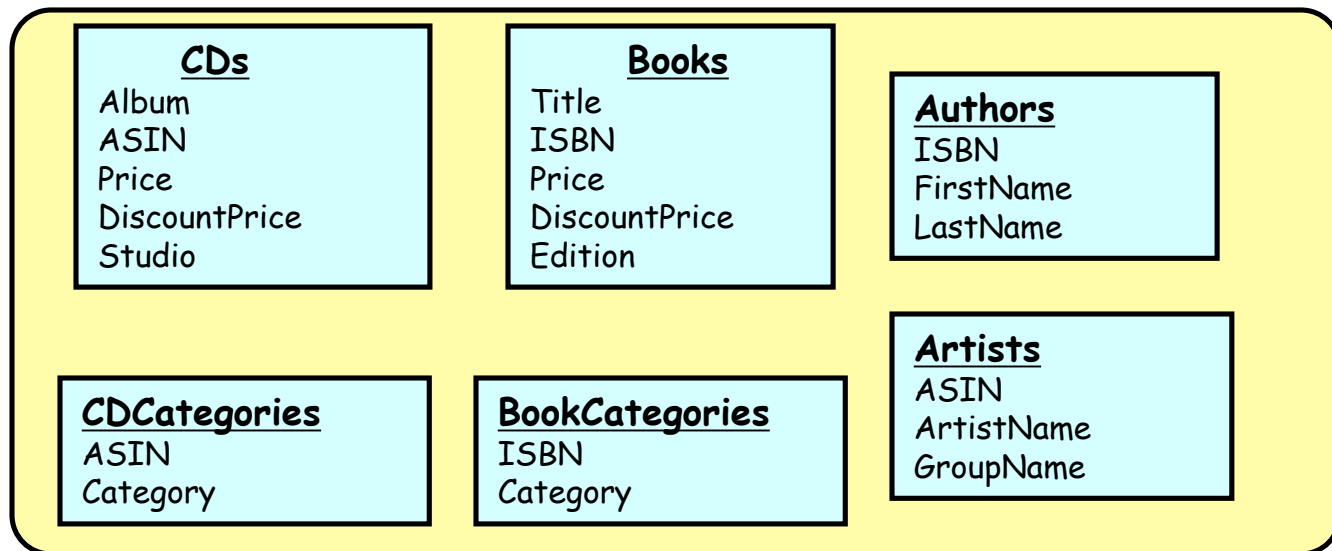
Mediation Languages

Mediated Schema

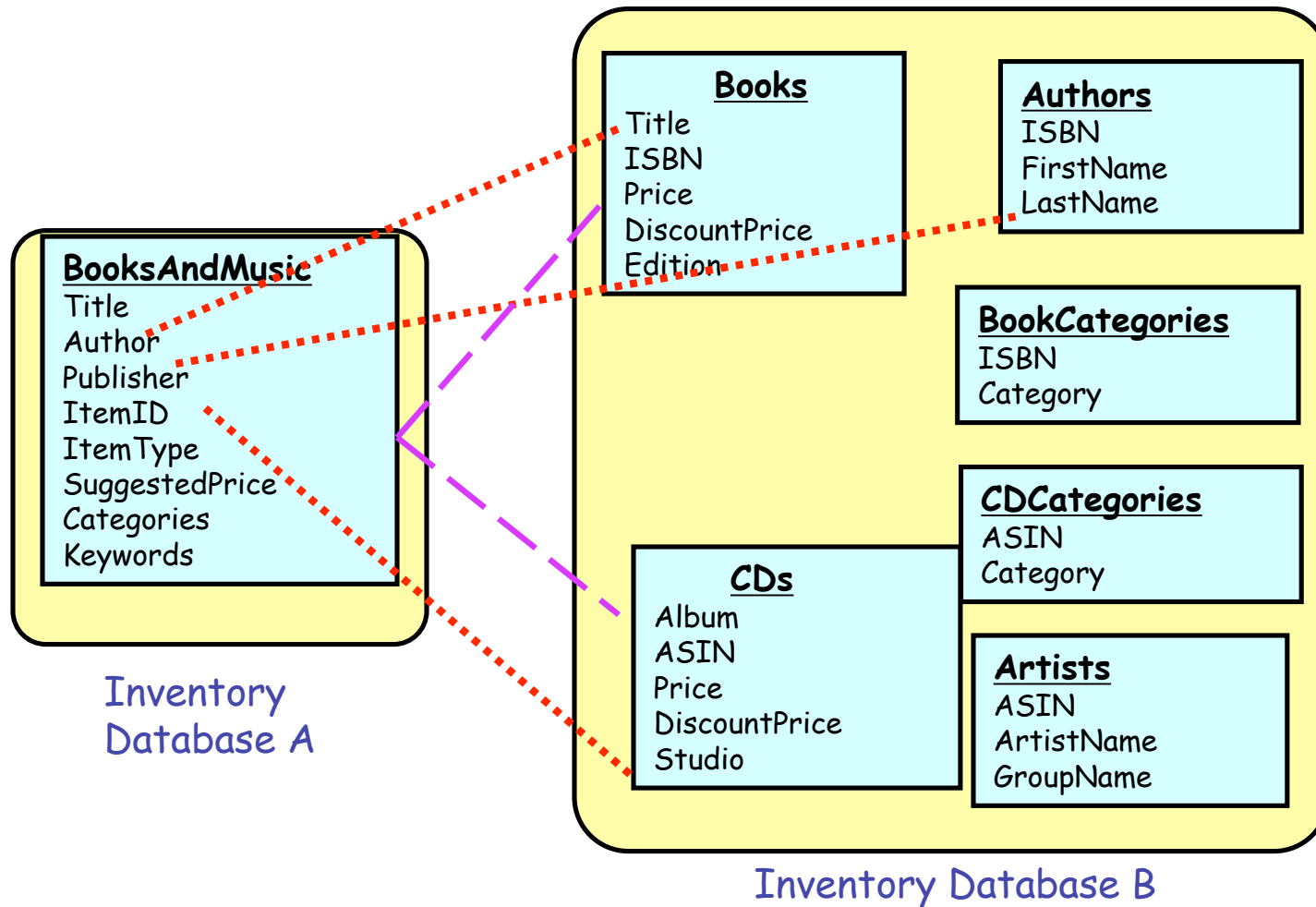
CD: ASIN, Title, Genre,...

Artist: ASIN, name, ...

logic



Semantic Mappings



“Standards are great, but there are too many of them.”



Techniques for Schema Mapping

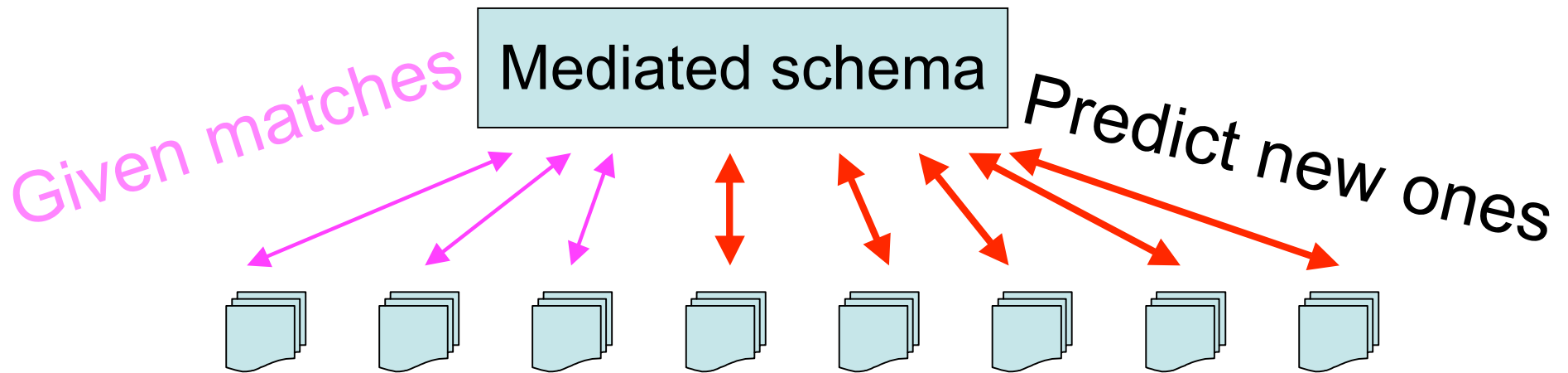
[Survey by Rahm and Bernstein, VLDBJ 2001]

- Compare schema elements based on:
 - Names (or n-grams)
 - Data types and instances
 - Text descriptions, integrity constraints
- Combine multiple techniques:
 - [Momis, Cupid, LSD, Coma]
- Create mappings from matches
 - [Clio @ IBM + Miller]



A Machine Learning Approach

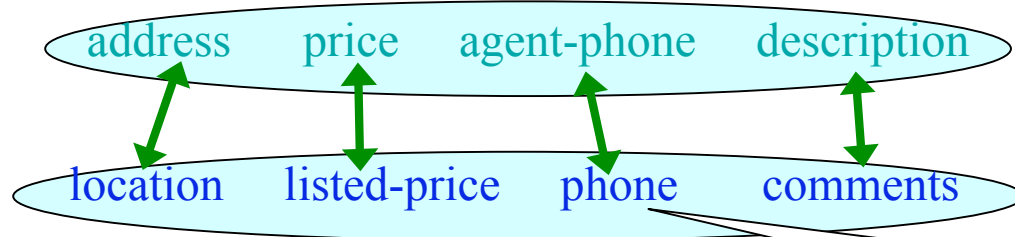
[Doan et al., 2001, ACM Distinguished Dissertation 2003]




- Many mapping tasks are repetitive
- Learn from previous experience:
 - Build a classifier for every element of the mediated schema.
 - Many kinds of cues → multi-strategy learning

Matching Real-Estate Sources

Mediated schema



Schema of realestate.com

realestate.com → 

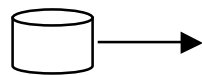
location	listed-price	phone	comments
Miami, FL	\$250,000	(305) 729 0831	Fantastic house
Boston, MA	\$110,000	(617) 253 1429	Great location
...

Learned hypotheses

If "phone" occurs in the name => agent-phone

If "fantastic" & "great" occur frequently in data values => description

homes.com



price	contact-phone	extra-info
\$550,000	(278) 345 7215	Beautiful yard
\$320,000	(617) 335 2315	Great beach
...

Used by Transformic to create Everyclassified.com

Reference Reconciliation

To Join or not to Join?

- Many ways to refer to the same object in the world:
 - “IBM”, “International Business Machines”
 - Alon Levy, Alon Halevy
- Automated methods are necessity
 - Can’t go through all the data manually
- Very active area in ML, KDD, DB, UAI,
...





Design time

Run time

Mediated Schema

mediation language

mapping tool

query reformulation

optimization & execution

XML

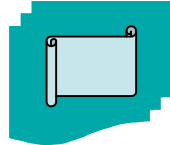
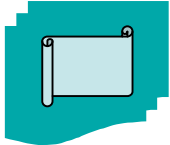
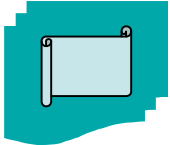
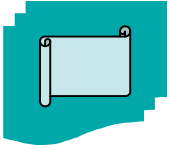
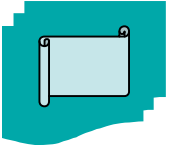
wrapper

wrapper

wrapper

wrapper

wrapper



Query Processing

To Plan or to Execute?

- In addition to distributed query processing issues:
 - Few statistics, if any.
 - Network behavior issues: latency, burstiness,...
 - Garlic @IBM
- “Adaptive query processing”:
 - Stonebraker saw it coming in Ingres.
 - Revivals by Graefe (1993) and DeWitt (1998).
 - Query scrambling [Urhan & Franklin]
 - Eddies [Avnur & Hellerstein]
 - Convergent query processing [Ives et al.]



XML Query Processing

- XML = “data integration appetizer”.
- Industry went ahead of research:
 - Nimble, Enosys, XQRL
 - Inspiration from Tukwila, MIX, Strudel/Agora
- (some) Issues:
 - Designing the internal algebra
 - Dealing with evolving XQuery standard
- The database community has served an impressive smorgasbord of XML techniques.



Other Major Developments

- Peer-to-peer data integration (e.g., Piazza @ UW)
- Data exchange systems (@IBM)
- Model management (Bernstein @ MSR)

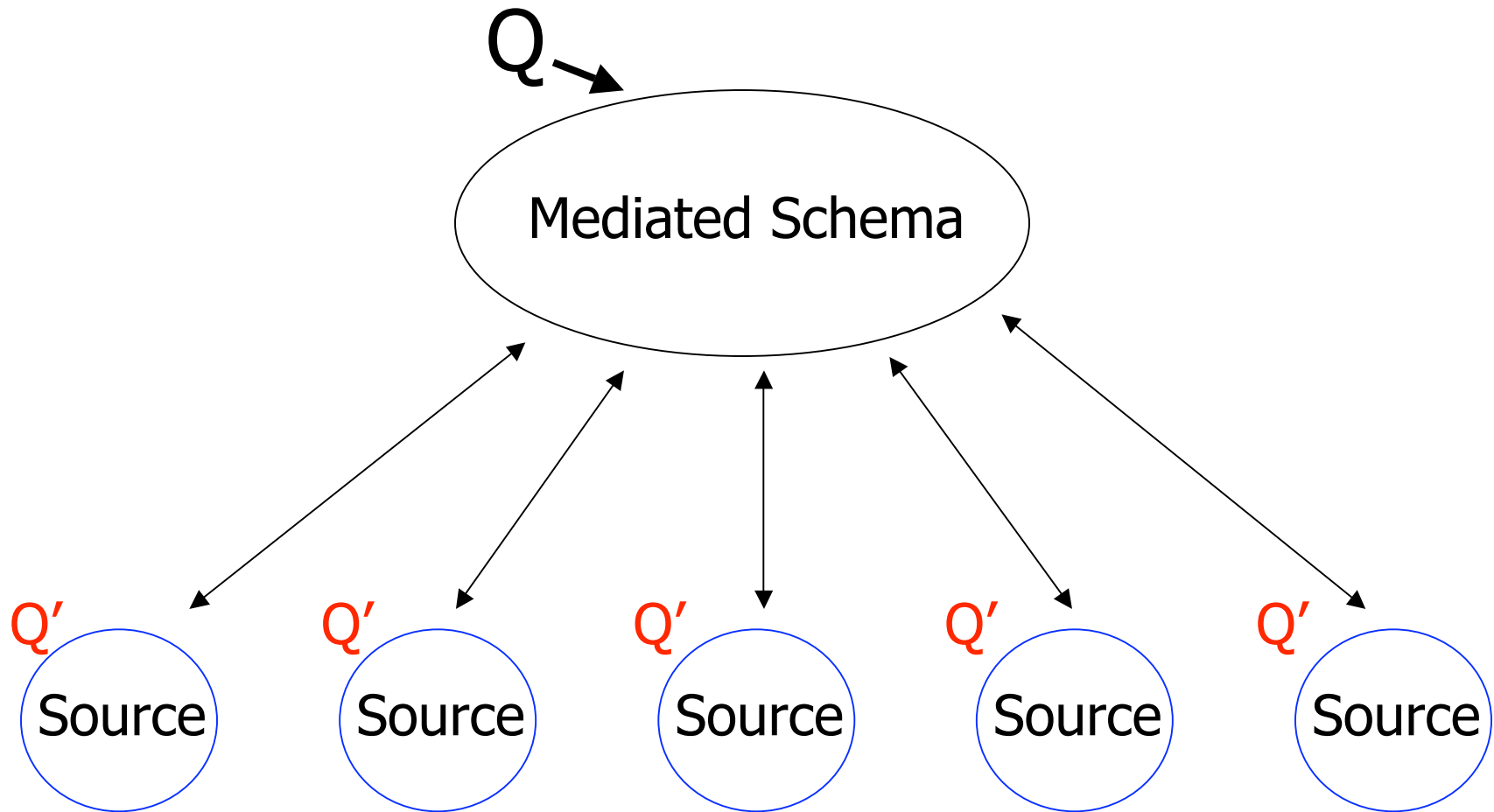


A Few Comments about Commerce

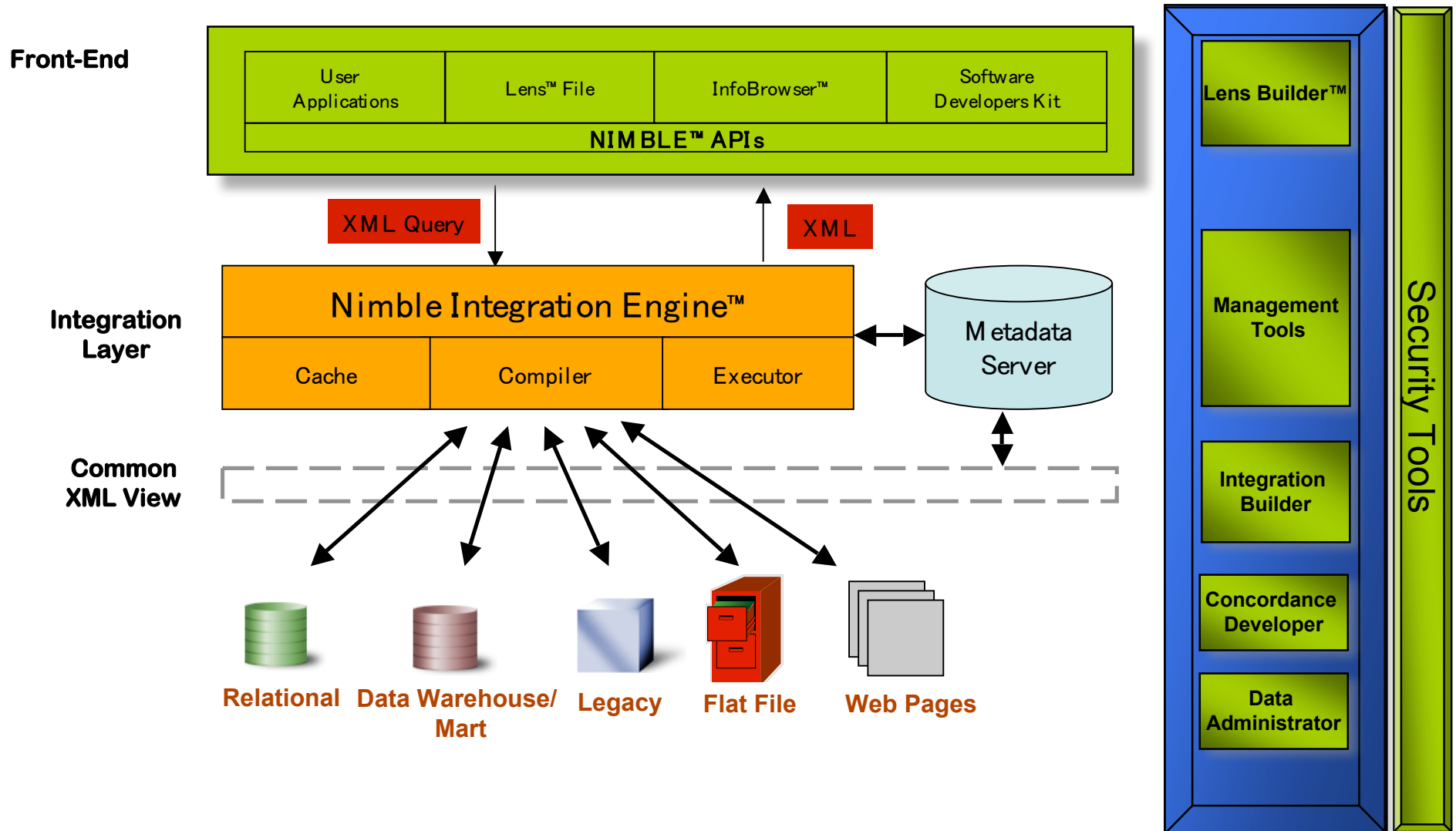
- Until a few years ago:
 - Data integration = Data warehousing.
- Since then:
 - A wave of startups:
 - Nimble, Enosys, MetaMatrix, Calixa, Composite
 - Big guys made announcements (IBM, BEA)
- Success! analysts invent new buzzword – EII
- Lessons:
 - Performance was fine. Need management tools.
 - Timing, timing, timing



Data Integration: Before



After \$30M



2007 Status Report

[Enterprise Angle]

- Enterprise Information Integration is established:
 - IBM, BEA, Oracle, MetaMatrix (soon RedHat), Composite, Actuate, ...
- Impact on design tools:
 - IBM Rational Data Architect
 - ADO .NET v. 3



Forrester Says...

"Enterprises are facing the **growing challenges of using disparate sources of data** managed by different applications, including problems with data integration, security, performance, availability and quality.... New technology is emerging that Forrester has coined **"information fabric,"** a term defined as a **virtualized data layer** that integrates heterogeneous data and content repositories in real time.... The potential benefits of this technology are so great that enterprises should develop a strategy to **leverage** information fabric technology as it becomes more widely available."



2007 Status Report

[Web Angle]

- Vertical search engines: *one* domain
- At scale: need even better source descriptions
 - deep web can be surfaced
- Terminology: *Data integration = mashups!*



Agenda

- Data integration:
 - Connecting disparate data sources
 - Great progress in last decade
- **But we're still missing the point:**
 - **Dataspaces: a new abstraction**
- A few connections to my Google work
- Predictions, subliminal messages (perhaps)



Shrapnel in Baghdad



Story courtesy of Phil Bernstein



The Web is Getting Semantic

cheese recipe - Google Search - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.google.com/search?q=cheese+recipe&start=0&ie=utf-8& Go cheese recipe

Customize Links Free Hotmail Windows Marketplace Windows Media Windows Crossing the Structur...

Google cheese recipe Search PageRank ABC Check AutoLink AutoFill Sign in

Web Images Groups News Froogle Maps Desktop Moma more »

Google cheese recipe Search Advanced Search Preferences

Web Results 1 - 10 of about 20,700,000 for **cheese recipe**. (0.60 seconds)

Try your search on [Yahoo](#), [Ask](#), [AllTheWeb](#), [Teoma](#), [MSN](#), [Lycos](#), [Technorati](#), [Feedster](#), [Wikipedia](#), [Bloglines](#), [Altavista](#)

Refine your search for **cheese recipe**

Main ingredient Cuisine

GourmetSleuth - Cheese Making Recipes
Guide to **cheese** making **recipes** and resources for the home **cheese** maker. Large link section.
www.gourmetsleuth.com/cheeserecipes.htm - 74k - [Cached](#) - [Similar pages](#) - [Filter](#)

Cheese Recipes from Hugs's Homeearth
Cheese souffle, salad, fried goat **cheese** cigars, and rarebits with beer.
www.hugs.org/cheesedex.shtml - 6k - [Cached](#) - [Similar pages](#) - [Filter](#)

Goat Milk Cheese Recipe
This page gives some great **recipes** for using raw goat milk.
www.utterlydivine.com/Recipes.htm - 11k - [Cached](#) - [Similar pages](#) - [Filter](#)

Sponsored Links

Danlac Dairy ingredients
Make healthy dairy products such as **Cheese**, Yogurt, Kefir, Sour Cream .
www.danlac.com/store

Cheese Recipe
Whatever you're looking for you can get it on eBay.
www.eBay.com

Home Cheesemaking
Articles on home cheesemaking, movies, recipes, and a forum area
www.rickandlynne.com

Cheese Recipe
Cheese recipe Online

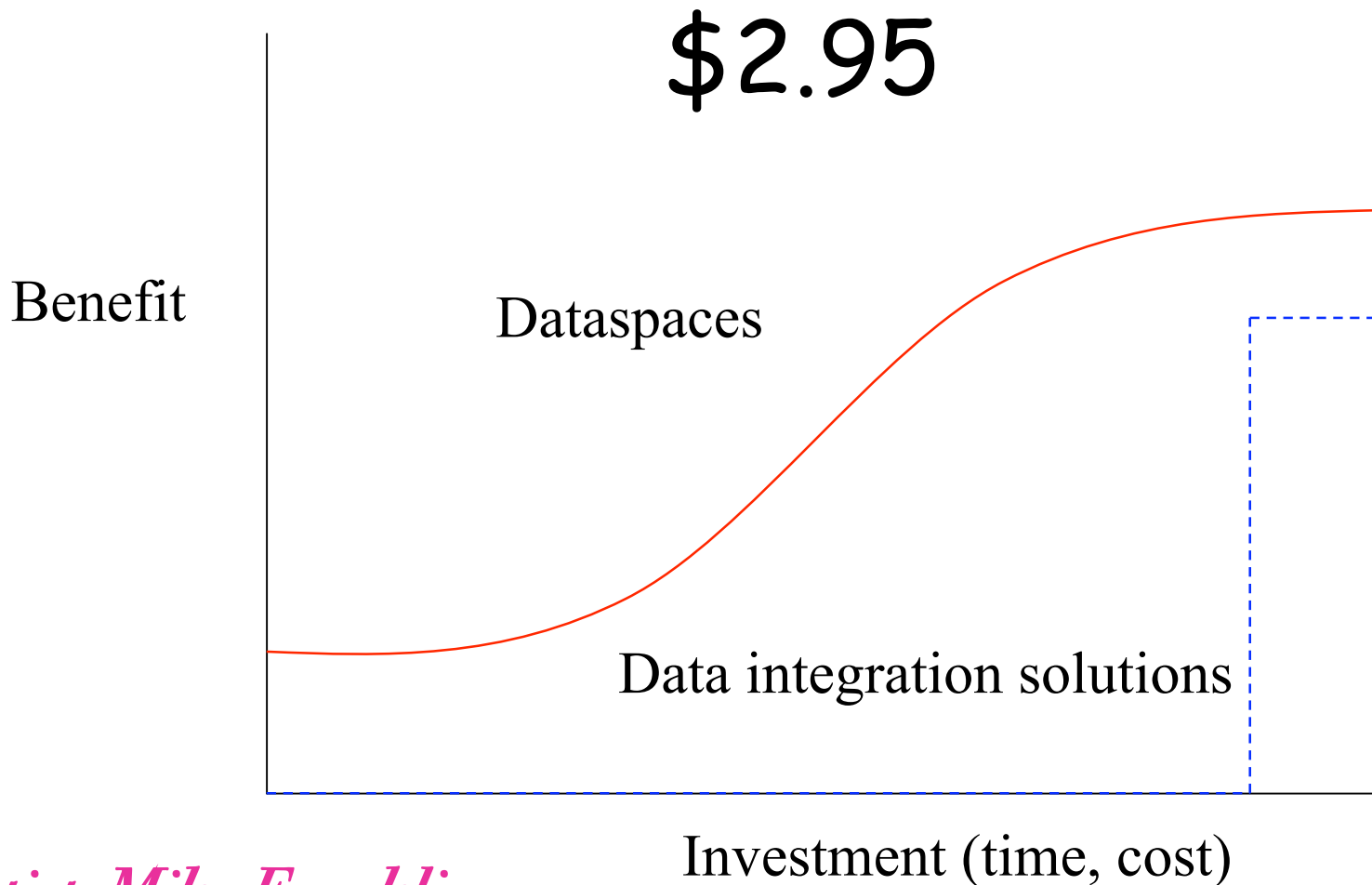
Done Open Notebook

“Data is the plural of anecdote”

- Personal information management
- Digital libraries, enterprises, “smart homes”
- Circle of Blue
 - Data about the world’s water sources
- The Boeing 777
 - [Hanrahan @ Stanford]

Pay-as-you-go Data Management

Dataspaces: Franklin, Halevy, Maier [see PODS 2006]



Artist: Mike Franklin

Other Dataspace Characteristics

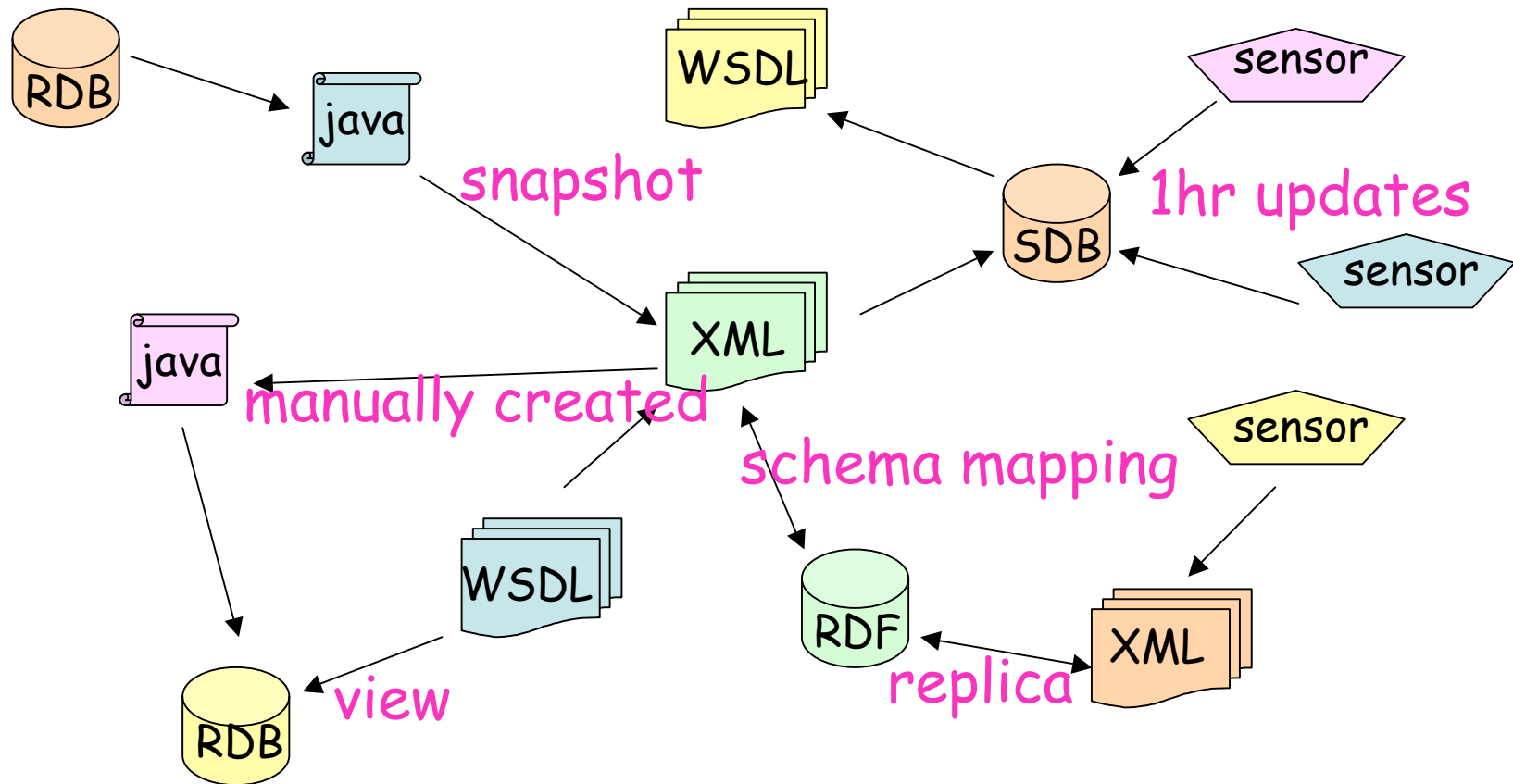
- All dataspaces contain $>20\%$ porn.



- The rest has $>50\%$ spam.



Participants and Relationships



Dataspace Support Platforms (DSSP)

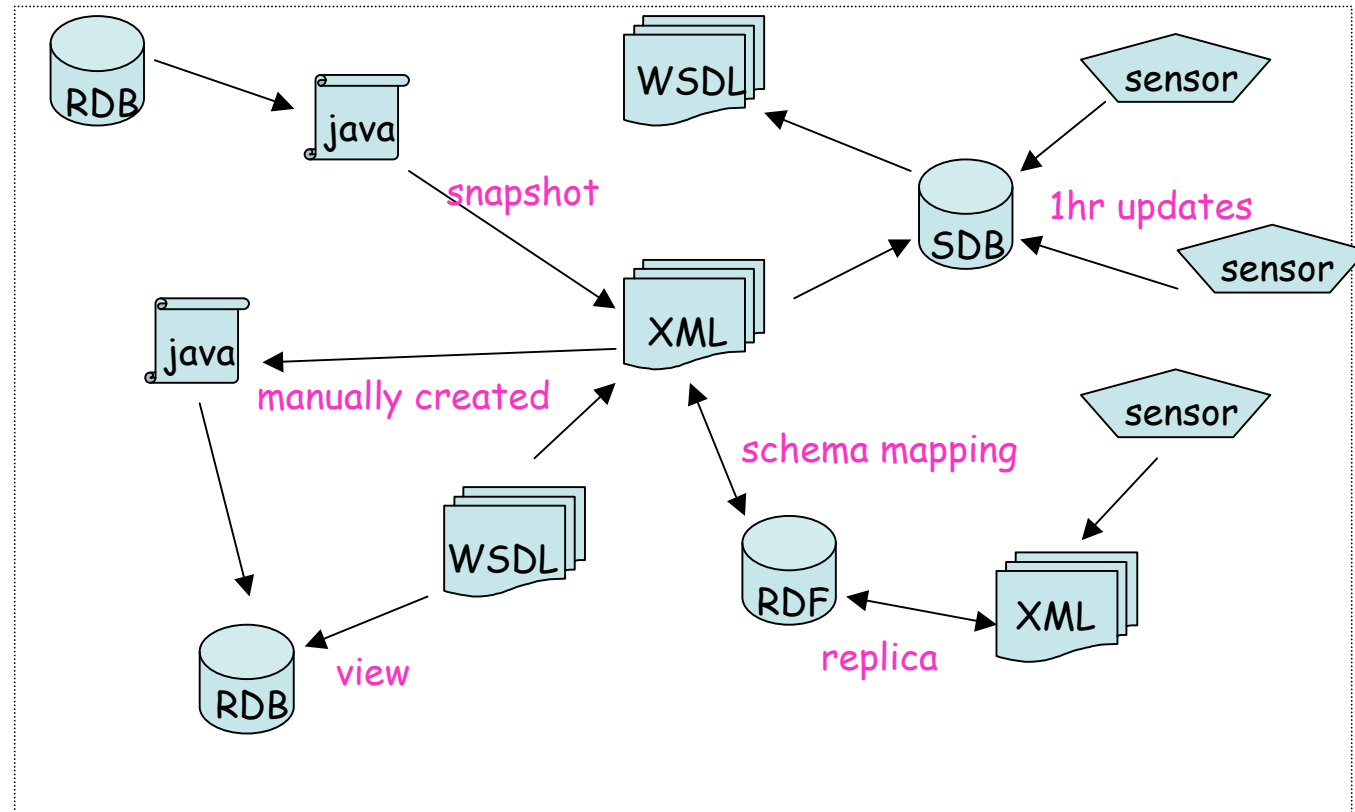
Discover & Enhance

Catalog

Local Store & Index

Search & query

Administration



Dataspaces: Main Points

- New types of queries & answers
- Reusing human attention for evolving a dataspace
- Visualization at the forefront.



Dataspace Queries

- Keyword queries as starting point
 - Later may be refined to add structure
 - Formulated in terms of user's "schema"
- Mostly of the form
 - *Instance**:
 - "britany spears"
 - *P (instance)*
 - "palo alto weather"
 - "PC chair SIGMOD"



Semantics of Answers

1. The actual answers:
 - $P(instance)$, $P^*(instance)$



weather seattle - Google Search - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.google.com/search?q=weather+seattle&start=0&ie weather seattle

Customize Links Free Hotmail Windows Marketplace Windows Media Windows Crossing the Structur...

Google weather seattle Search PageRank ABC Check AutoLink >>

Sign in



Google Weather Seattle

Web Results 1 - 10 of about 63,000,000 for [weather seattle](#). (0.11 seconds)

Try your search on [Yahoo](#), [Ask](#), [AllTheWeb](#), [Teoma](#), [MSN](#), [Lycos](#), [Technorati](#), [Feedster](#), [Wikipedia](#), [Bloglines](#), [Altavista](#)

Weather for Seattle, WA

61°F
Mostly Cloudy
Wind: N at 5 mph
Humidity: 59%

Mon	Tue	Wed	Thu
 67° 50°	 69° 51°	 72° 52°	 73° 51°

Find more forecasts at [Yahoo](#), [Ask](#), [Netscape](#), [CNN](#), [USA Today](#), [Ameriwx](#), [Weather Underground](#), [Weather.com](#), [AccuWeather](#)

Sponsored Links

[Weather Seattle](#)
Scenic Byways, Quaint Towns & More!
Travel through Washington & SayWA.
www.ExperienceWA.com
Washington

NWsource: [Weather: Seattle, Washington](#)
The **Seattle** Times Company, jobs ... Browse **weather**. **Seattle** forecast - **Weather** maps ...
Seattle, WA (98101) CURRENT CONDITIONS ...
www.nwsources.com/weather/scr/ - 27k - Jun 18, 2006 - [Cached](#) - [Similar pages](#) - [Filter](#)

[Seattle, Washington \(98101\) Conditions & Forecast: Weather ...](#)
Want to contribute your **weather** data? The **Weather** Underground has teamed up with
Ambient ... Historical Data & Charts — The **Flees Weather Station - Seattle**, ...
www.wunderground.com/US/WA/Seattle.html - 191k - Jun 18, 2006

http://www.google.com/webhp?hl=en

Open Notebook



Semantics of Answers

1. The actual answers:
 - $P(instance)$, $P^*(instance)$
2. Sources where answer can be found:
 - Partially specify the query to the source
 - Help the user *clean* the query



toyota corolla palo alto - Google Search - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.google.com/search?q=toyota+corolla+ yota corolla palo alto

Customize Links Free Hotmail Windows Marketplace Windows Crossing the Structur...

Google toyota corolla palo alto

Toyota Corolla Palo alto

Web Results 1 - 10 of about 70,200 for [toyota corolla palo alto](#). (0.64 seconds)

Try your search on [Yahoo](#), [Ask](#), [AllTheWeb](#), [Teoma](#), [MSN](#), [Lycos](#), [Technorati](#), [Feedster](#), [Wikipedia](#), [Bloglines](#), [Altavista](#)

Refine your search for [toyota corolla palo alto](#)

Location	Make	Model
<input type="text" value="palo alto"/>	<input type="text" value="toyota"/>	<input type="text" value="corolla"/>

Remember this location

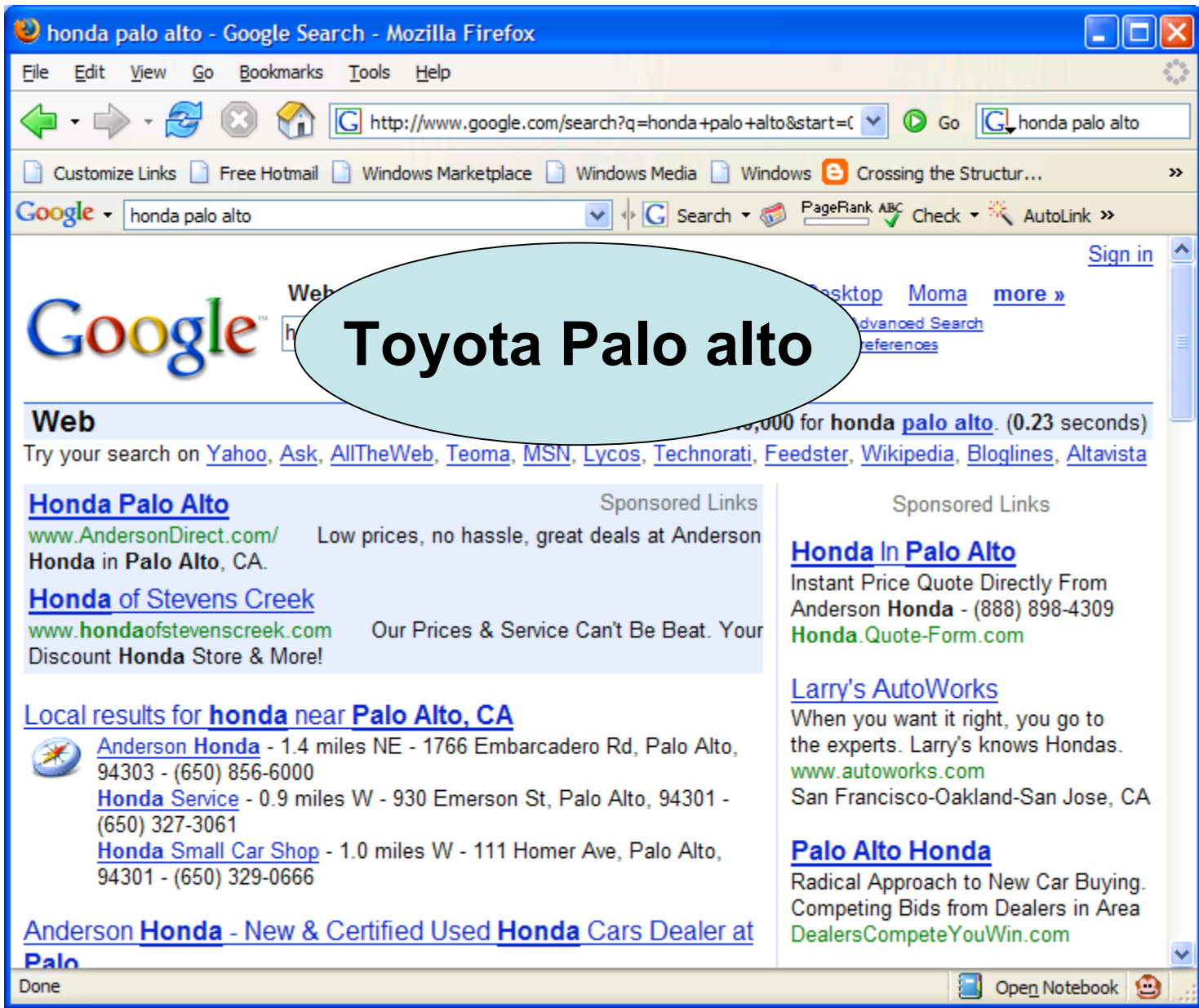
Fogster | [Toyota 1996 Corolla - \\$3900 \(Palo Alto\)](#)
Taking the power of newspaper classifieds and the reach of the Internet, Fogster.com is a new approach to classifieds both online and in print.
www.fogster.com/listing.php?id=79414 - 10k -
[Cached](#) - [Similar pages](#) - [Filter](#)

Sponsored Links

[Toyota Menlo](#)
Toyota Models on Sale!
Find The Latest Prices, Reviews
www.Toyota.edmunds.com

[Palo alto toyota](#)
Local dealer in Sunnyvale. We offer online specials, ask for a quote.
www.Toyota-Sunnyvale.com
San Francisco-Oakland-San Jose, CA

Done



Toyota Palo alto

honda palo alto - Google Search - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.google.com/search?q=honda+palo+alto&start=c Go honda palo alto

Customize Links Free Hotmail Windows Marketplace Windows Media Windows Crossing the Structur...

Google Search PageRank ABC Check AutoLink

Google Web Search Desktop Moma more Sign in

Web 10,000 for honda palo alto. (0.23 seconds) Try your search on Yahoo, Ask, AllTheWeb, Teoma, MSN, Lycos, Technorati, Feedster, Wikipedia, Bloglines, Altavista

Honda Palo Alto Sponsored Links www.AndersonDirect.com/ Low prices, no hassle, great deals at Anderson Honda in Palo Alto, CA. Honda of Stevens Creek www.hondaofstevenscreek.com Our Prices & Service Can't Be Beat. Your Discount Honda Store & More!

Honda In Palo Alto Sponsored Links Instant Price Quote Directly From Anderson Honda - (888) 898-4309 Honda.Quote-Form.com

Local results for honda near Palo Alto, CA Anderson Honda - 1.4 miles NE - 1766 Embarcadero Rd, Palo Alto, 94303 - (650) 856-6000 Honda Service - 0.9 miles W - 930 Emerson St, Palo Alto, 94301 - (650) 327-3061 Honda Small Car Shop - 1.0 miles W - 111 Homer Ave, Palo Alto, 94301 - (650) 329-0666

Larry's AutoWorks When you want it right, you go to the experts. Larry's knows Hondas. www.autoworks.com San Francisco-Oakland-San Jose, CA

Anderson Honda - New & Certified Used Honda Cars Dealer at Palo

Palo Alto Honda Radical Approach to New Car Buying. Competing Bids from Dealers in Area DealersCompeteYouWin.com

Done Open Notebook

Semantics of Answers

1. The actual answers:
 - $P(instance)$, $P^*(instance)$
2. Sources where answer can be found:
 - Partially specify the query to the source
 - Help the user *clean* the query
3. Supporting facts or sources:
 - Facts that can be used to derive $P(instance)$
 - Rest of derivation may be obvious to user



Related or Partial Answers

- In which country was Alon Halevy born?
 - *Rehovot*
- Latest edition of software X:
 - *2004 edition*
- Is the Space Needle higher than the Eiffel Tower?

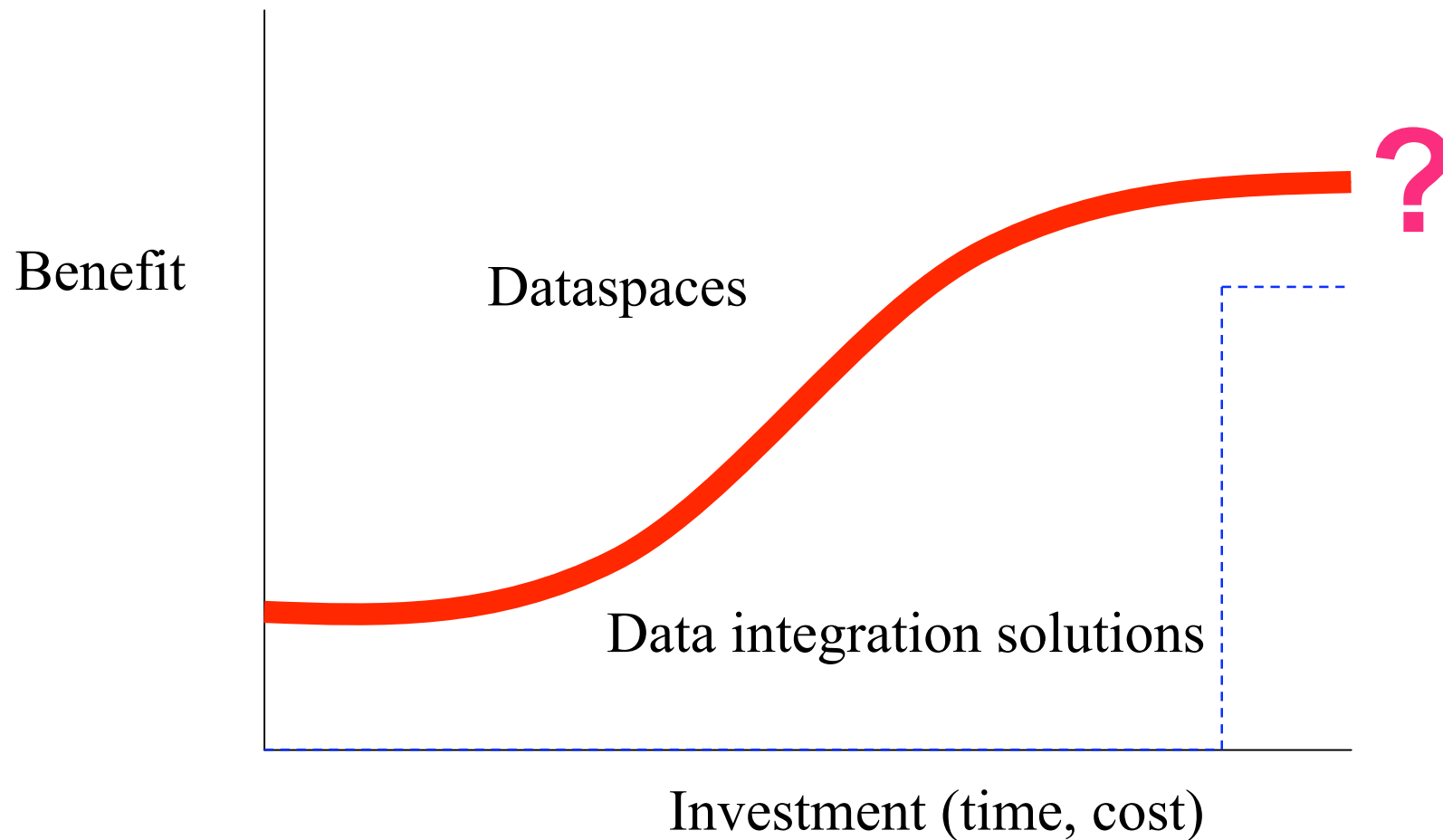
184m – *Height of Seattle Space Needle*

324m – *Height of Eiffel Tower*

Rank all types of answers



The Cost of Semantics



Reusing Human Attention

- Principle:
 - *User action = statement of semantic relationship*
 - *Leverage actions to infer other semantic relationships*
- Examples
 - Providing a semantic mapping
 - Infer other mappings
 - Writing a query
 - Infer source contents, relationships between sources
 - Creating a “digital workspace”
 - Infer “relatedness” of documents/sources
 - Infer co-reference between objects in the dataspace
 - Annotating, cutting & pasting, browsing among docs



Examples of Reuse

- Leverage past actions & existing structure:
 - [Dong et al., 2004, 2005], [He & Chang, 2003]
- Generalize from current actions
 - Queries, schema mappings
- Beg for extra attention:
 - ESP [von Ahn], mass collaboration [Doan+], active learning [Sarawagi et al.]



Conclusions

- Data integration is now real.
- Next step for data management:
 - Consumer facing interfaces (data management for the masses)
- Dataspaces: a key abstraction for the new agenda
 - Principle: reuse human attention



Some References

- Dataspaces:
 - Original vision: SIGMOD Record, December 2005
 - Technical challenges: PODS 2006
- Data Integration:
 - The Teenage Years: VLDB 2006
 - EII experiences: SIGMOD 2005
 - The book: in progress...
- Teaching integration to undergraduates:
 - SIGMOD Record, September, 2003.



Some References

- www.cs.washington.edu/homes/alon
- Piazza: ICDE03, WWW03, VLDB-03, SIGMOD-04
- SSS: [Madhavan, forthcoming], VLDB-04.
- Semex: IIWeb-04
- Surveys on schema matching languages:
 - Halevy, VLDB Journal 01
 - Lenzerini, PODS 2002
- Teaching integration to undergraduates:
 - SIGMOD Record, September, 2003.

