Dataspaces: Co-Existence with Heterogeneity

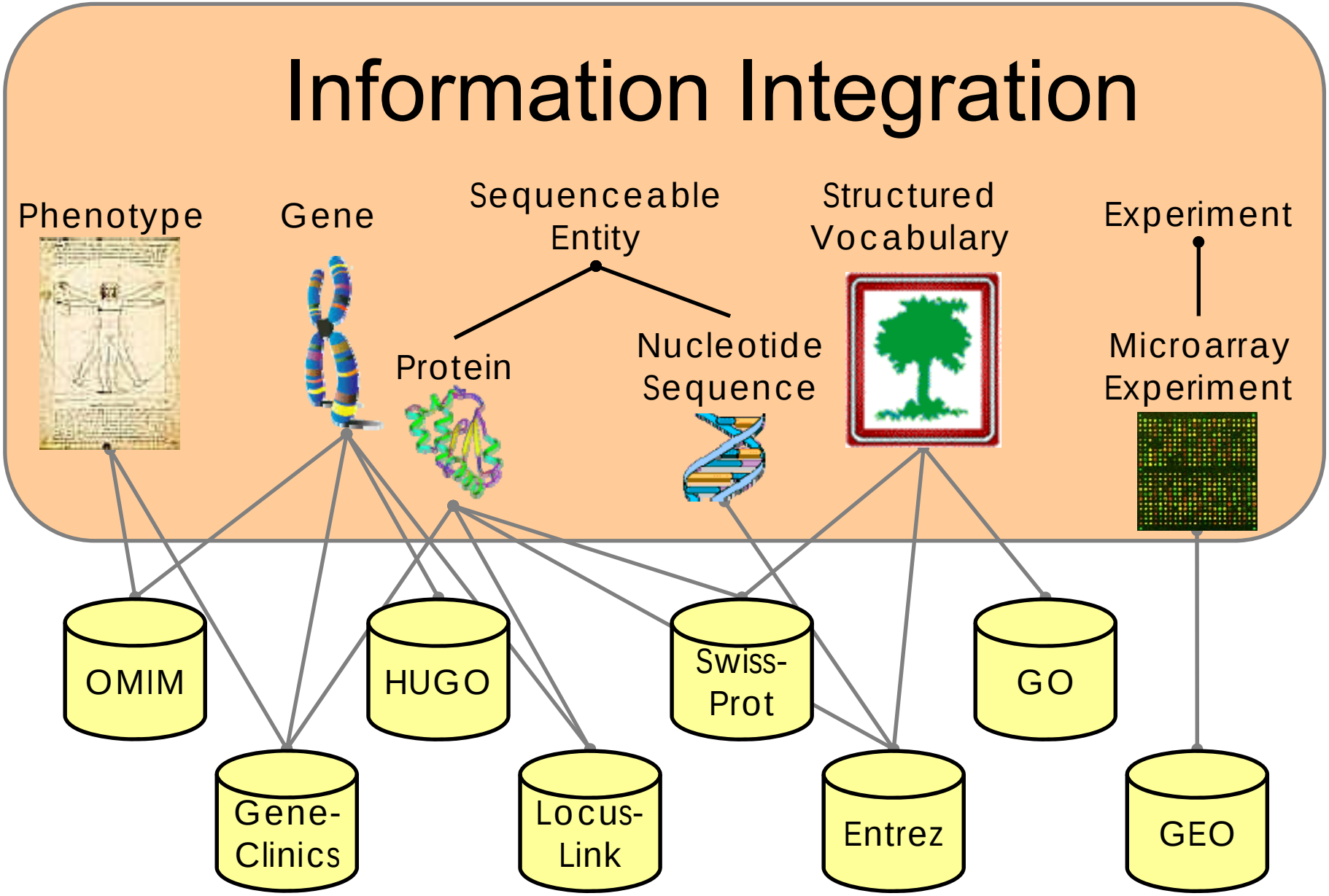
Alon Halevy

Google

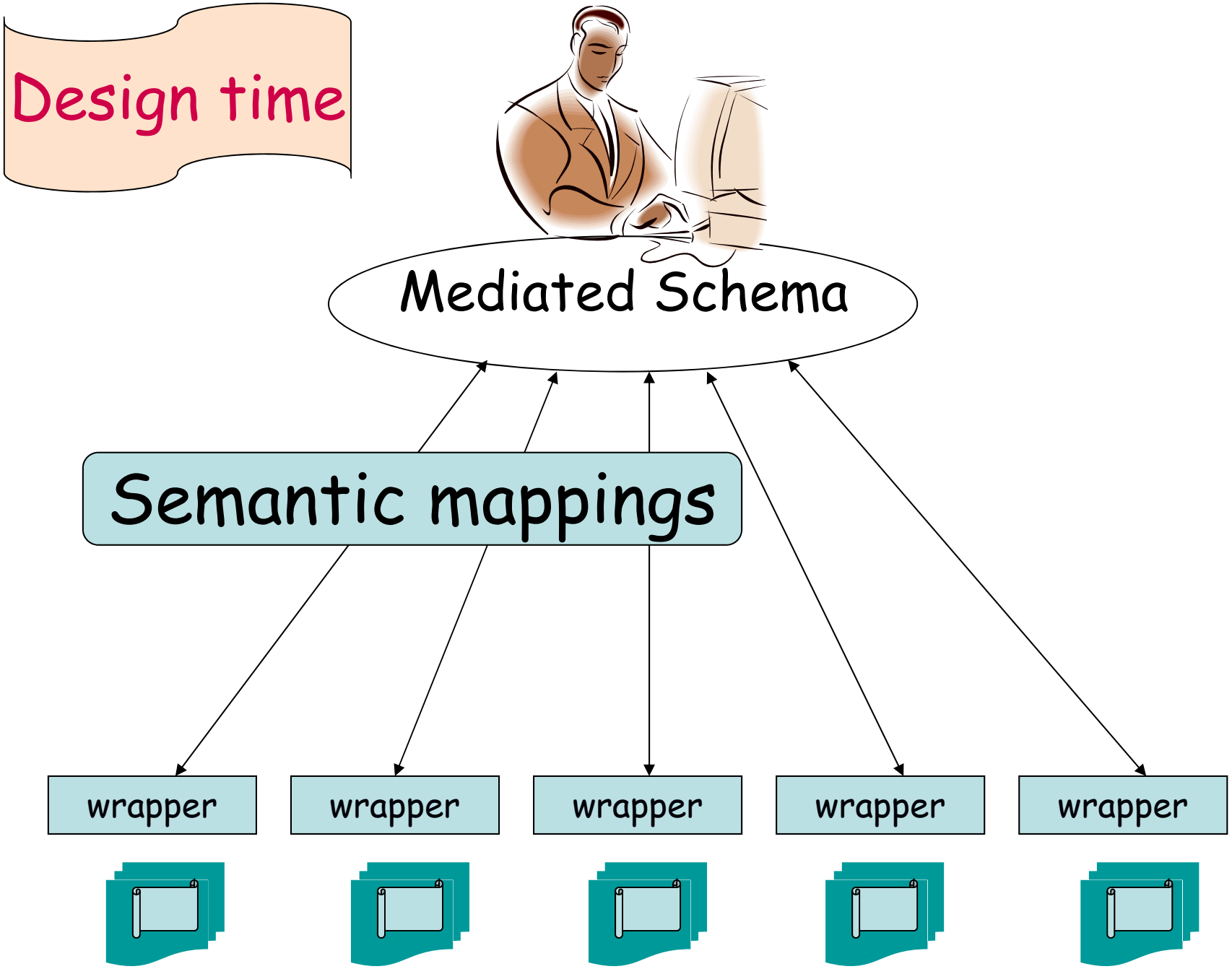
Agenda

- Basic assumption: ***KR = "fancy" DB***
- DB trends – what do they mean for KR?
  - ✓From DB's to integrating heterogeneous data
  - ➢From integration to co-existence
- Dataspaces: [Franklin, Halevy, Maier]
  - – "pay-as-you-go" data management
  - – Dataspace querying, evolution and reflection
  - – Need for KR services

Information Integration



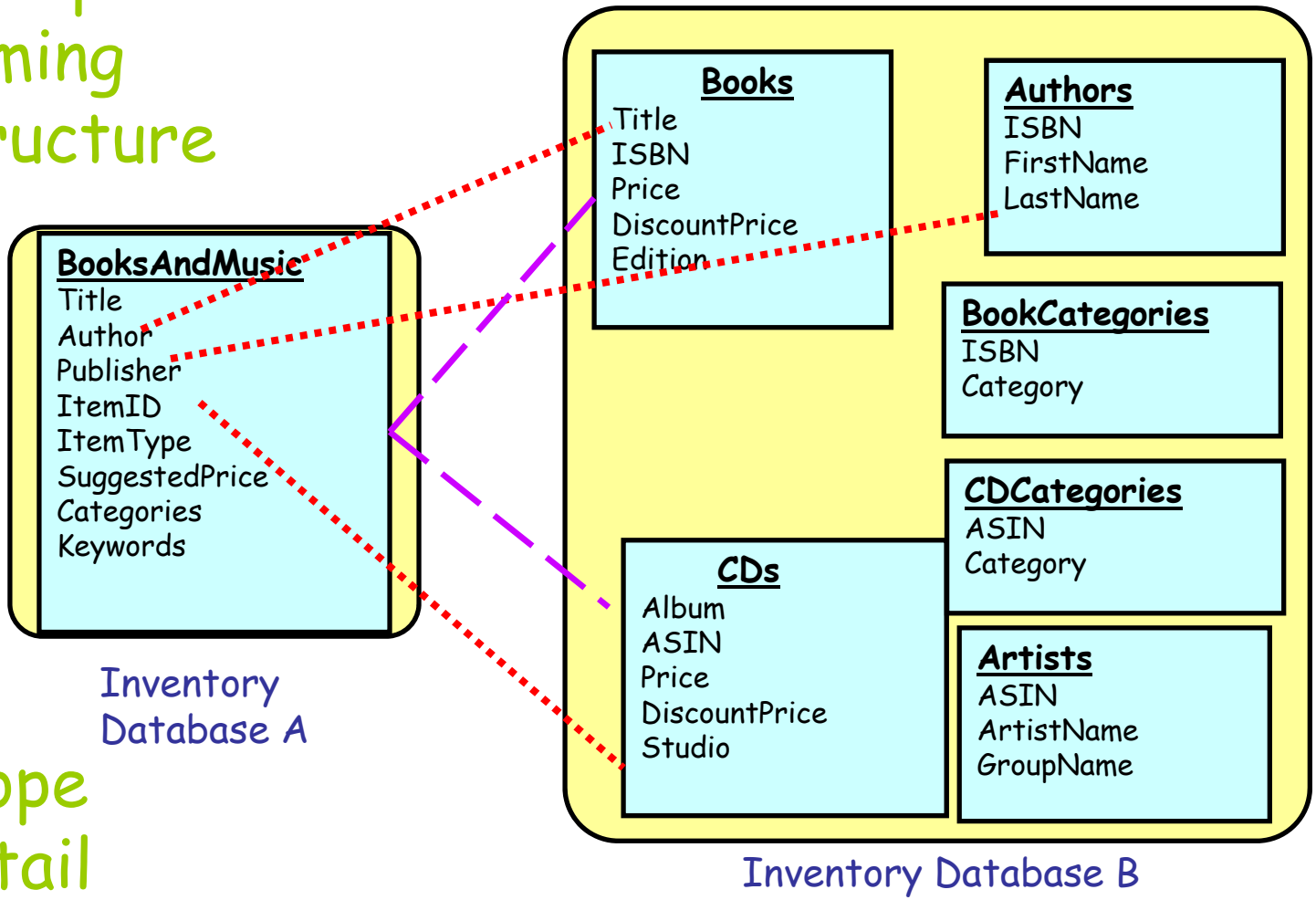
Query multiple sources with a single interface


Design time
Mediated Schema
Semantic mappings
wrapper
wrapper
wrapper
wrapper
wrapper

Semantic Mappings

Discrepancies in:

- naming
- structure



- scope
- detail

Mediation Languages

Requirements:
- expressive
- tractable
- easy to modify

Mediated Schema
CD: ASIN, Title, Genre,…
Artist: ASIN, name, …

logic

Lenzerini 02,
Halevy 01

CDs
Album
ASIN
Price
DiscountPrice
Studio

Books
Title
ISBN
Price
DiscountPrice
Edition

Authors
ISBN
FirstName
LastName

Artists
ASIN
ArtistName
GroupName

CDCategories
ASIN
Category

BookCategories
ISBN
Category

Q: reviews of books by Tom Friedman

Run time

Mediated Schema

Reformulation

Query processing (adaptive!)

The World is Flat
Lexus and the Olive Tree

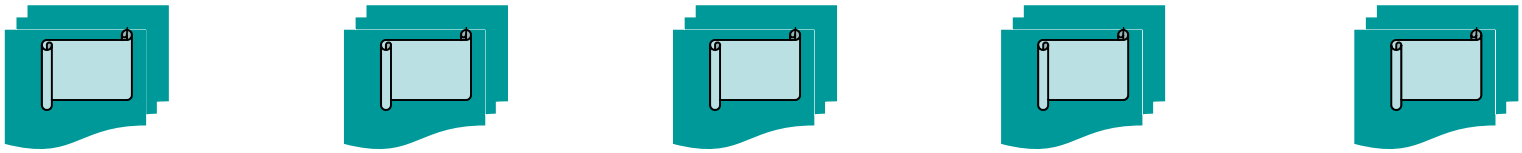
Great…
A bit repetitive…

Q2': TWiF
Q2": LatOT

Q1: books by T.F.

Q2: reviews of book

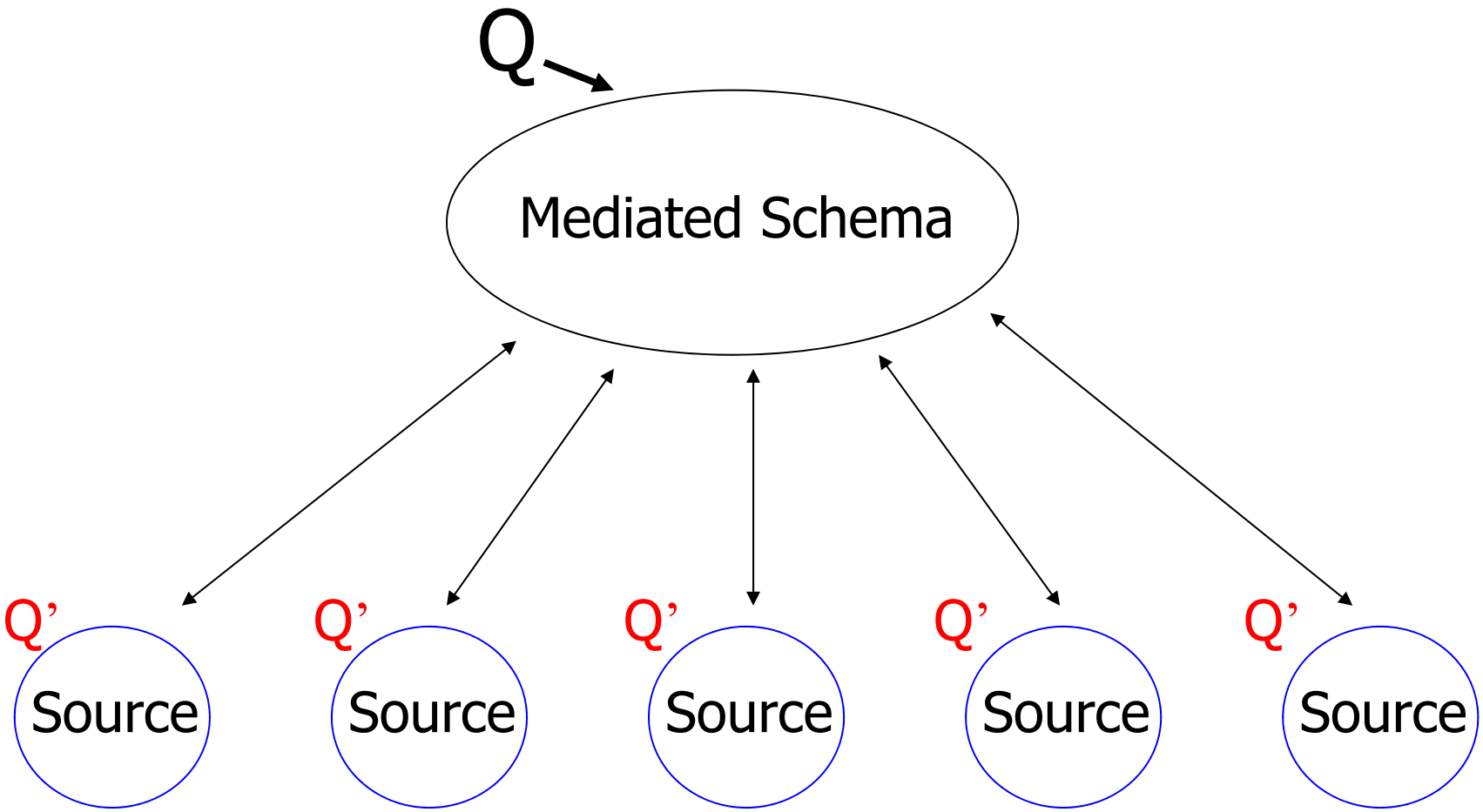
wrapper | wrapper | wrapper | wrapper | wrapper



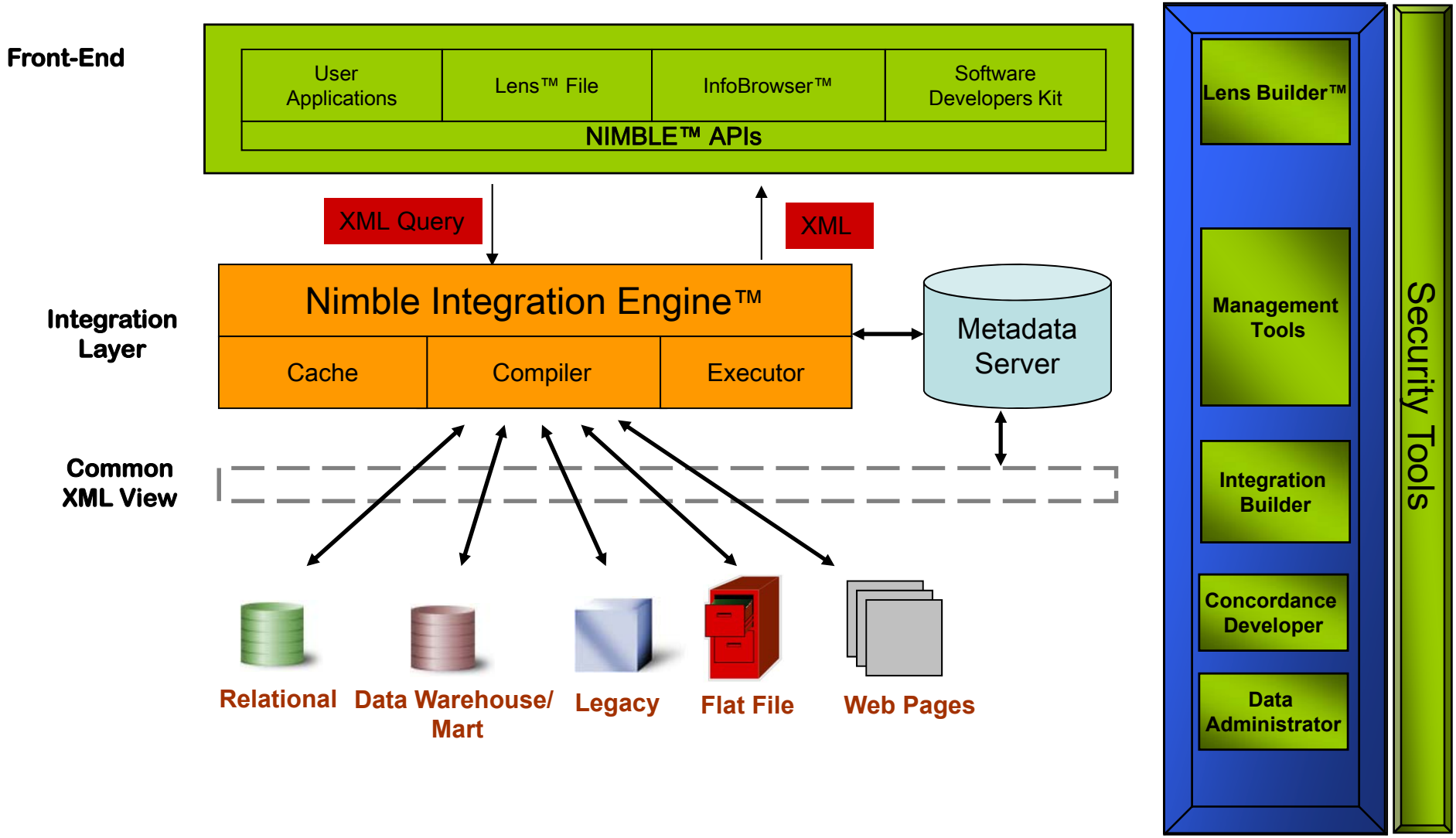
Enterprise Information Integration

- Late 90's -- anything goes:
  - Say "XML" -- get VC money.
  - A wave of startups:
    - Nimble, Enosys, MetaMatrix, Calixa, Cohera, …
  - Big guys made announcements (IBM, BEA).
  - [Delay] Big guys released products.
- Lessons:
  - Performance was fine. Need management tools.
  - Timing was less than optimal.

Data Integration: Before



Data Integration After $30M

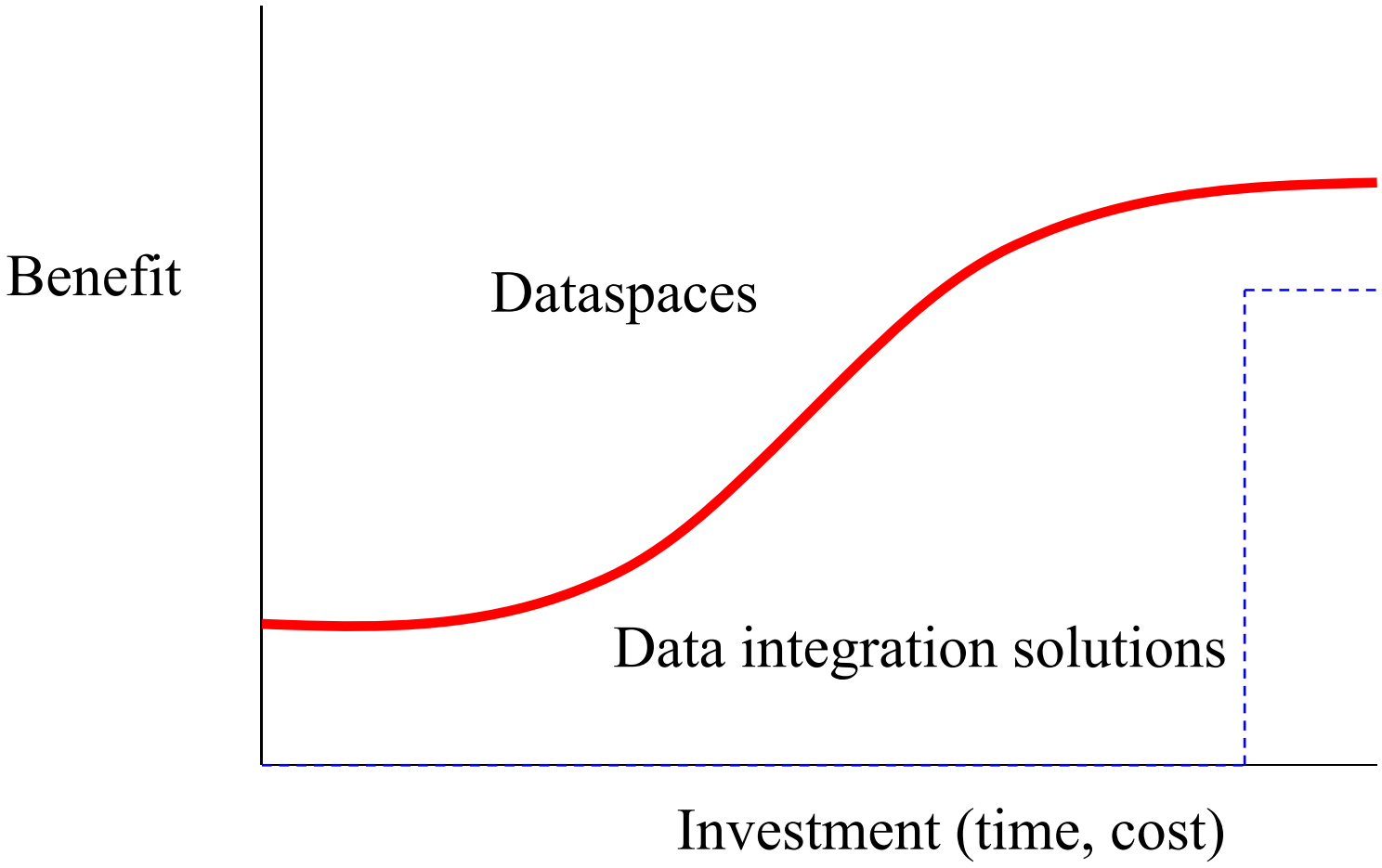


So What's Wrong?

- We're still hung up on semantics:
  - No mapping, no service.
  - Too much upfront effort needed.

Dataspaces vs. Data Integration

Dataspaces are “pay as you go”

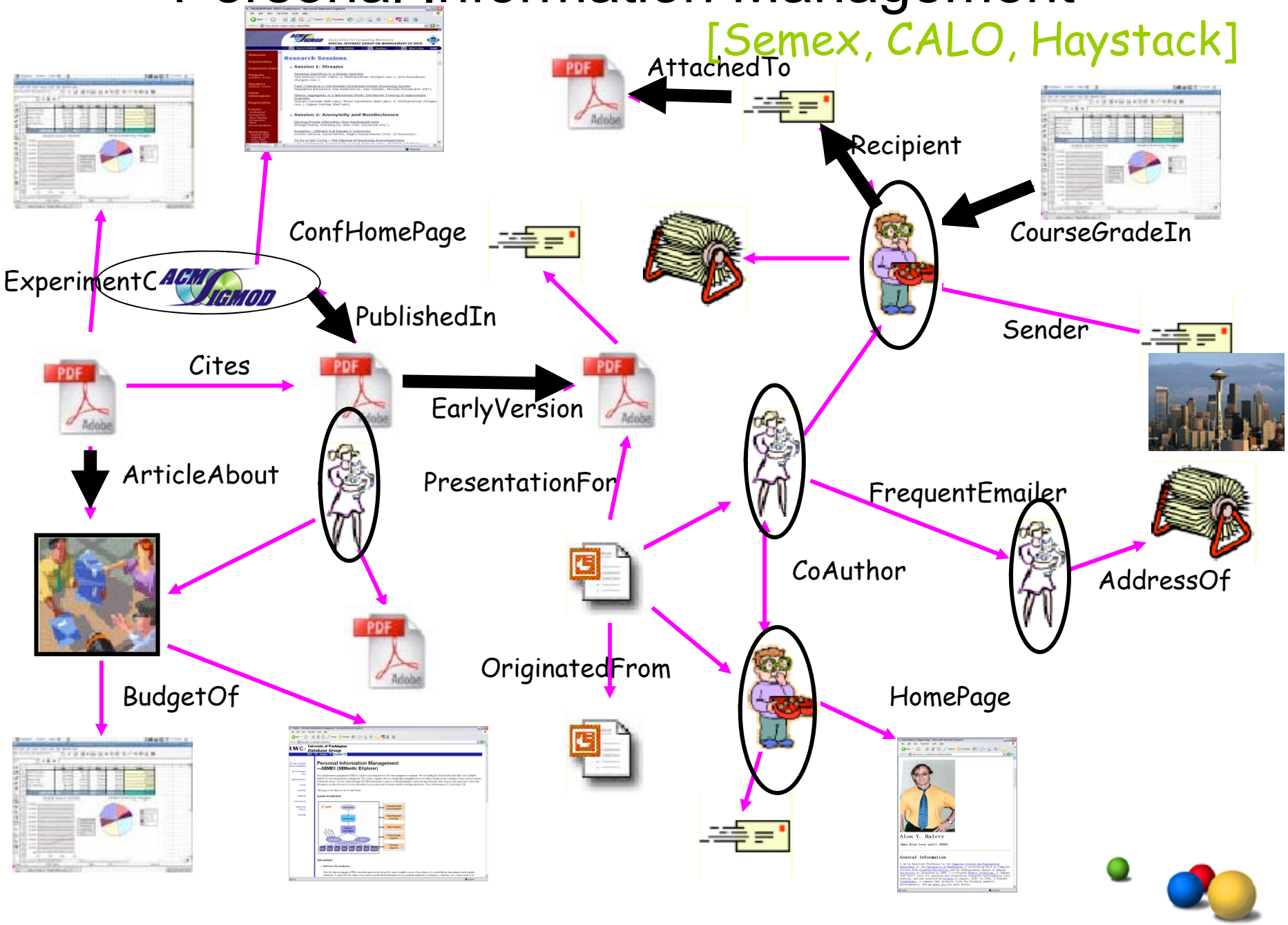


Shrapnels in Baghdad

Story courtesy of Phil Bernstein

Personal Information Management

[Semex, CALO, Haystack]



Google Base

cheese recipe - Google Search - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

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

Customize Links Free Hotmail Windows Marketplace Windows Media Windows

Google cheese recipe Search PageRank Check AutoLink AutoFill Options »

Google Base | Caribou - Inbox | Caribou - Inbox ... | cheese recip... | Applicant Tracki... | Madaba Vacatio...

Web Images Groups News Froogle Local Desktop Moma more »

Google cheese recipe Search Advanced Search Preferences

Web Results **1** - **10** of about **22,400,000** for **cheese recipe**. (**0.18** seconds)

Refine your search for **cheese recipe**

Main ingredient: cheese | Cuisine: all | Search recipes

GourmetSleuth - **Cheese** Making **Recipes**
Guide to **cheese** making **recipes** and resources for the home **cheese** maker. Large link section.
www.gourmetsleuth.com/**cheeserecipes**.htm - 71k - Cached - Similar pages

Tillamook **Cheese** - **Recipe** Index
Tillamook **Cheese** is a farmer owned cooperative inTillamook, Oregon that produces a premier line of dairy products including award-winning **cheese**, ice cream, ...
www.tillamook**cheese**.com/**recipes**/**recipe**sindex.html - 17k - Cached - Similar pages

Cheese Recipe Index
Cheese souffle, salad, fried goat **cheese** cigars, and rarebits with beer.
www.hugs.org/**cheese**dex.shtml - 6k - Cached - Similar pages

Sponsored Links

Cheese Recipe
Looking to find **cheese**?
Browse our **cheese** directory.
CheeseCatalog.net

Easy **Cheese** Recipes
Hundreds Of Free **Cheese** Recipes
From Better Homes and Gardens
www.BetterHomesandGardens.com

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

Cheese Recipe
Cheese recipe Online.
Shop Target.com
www.Target.com

Done

The Web is Getting Semantic

- Forms (millions)
- Vertical search engines (hundreds)
- Annotation schemes:
  - Flickr, ESP Game
- Google Base
- Google Coop

“A little semantics goes a long way”

“Data is the plural of anecdote”

Dataspace Characteristics

- Defined by boundaries (organizational, physical, logical)
  - ***Not*** by explicitly entering content
- Need to consider ***all*** the data in the space
- Must provide ***best-effort*** services:
  - Cannot wait for full integration
- Certainly cannot assume clean, schema conforming data

Other Dataspace Characteristics

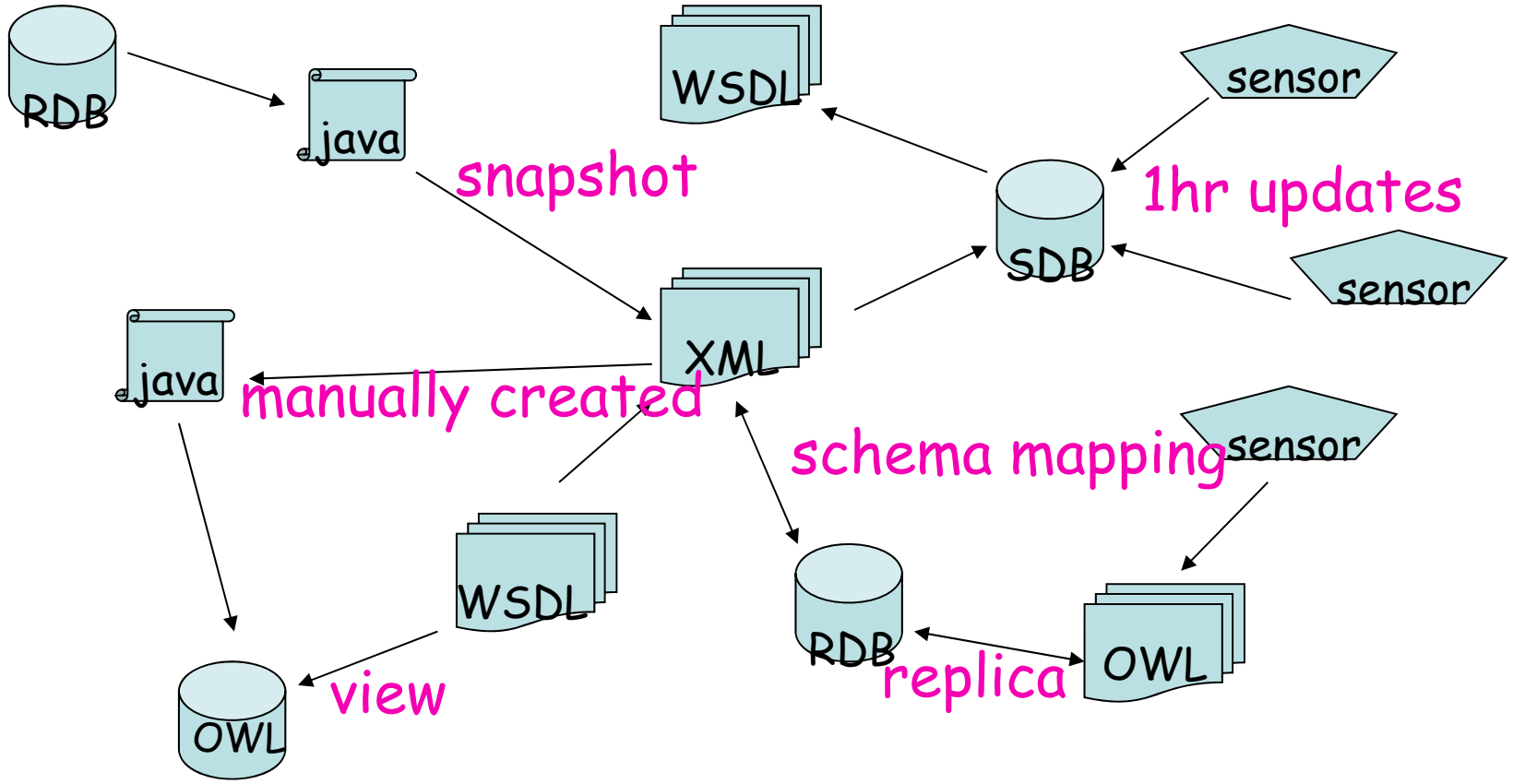
- All dataspaces contain >20% porn.

- The rest has >50% spam.

Outline

- Logical model for dataspaces
  - Participants and relationships
  - Dataspace support platforms (DSSPs)
- Querying dataspaces
- Dataspace evolution
  - Generating semantic mappings
- Dataspace reflection

Logical Model: Participants and Relationships



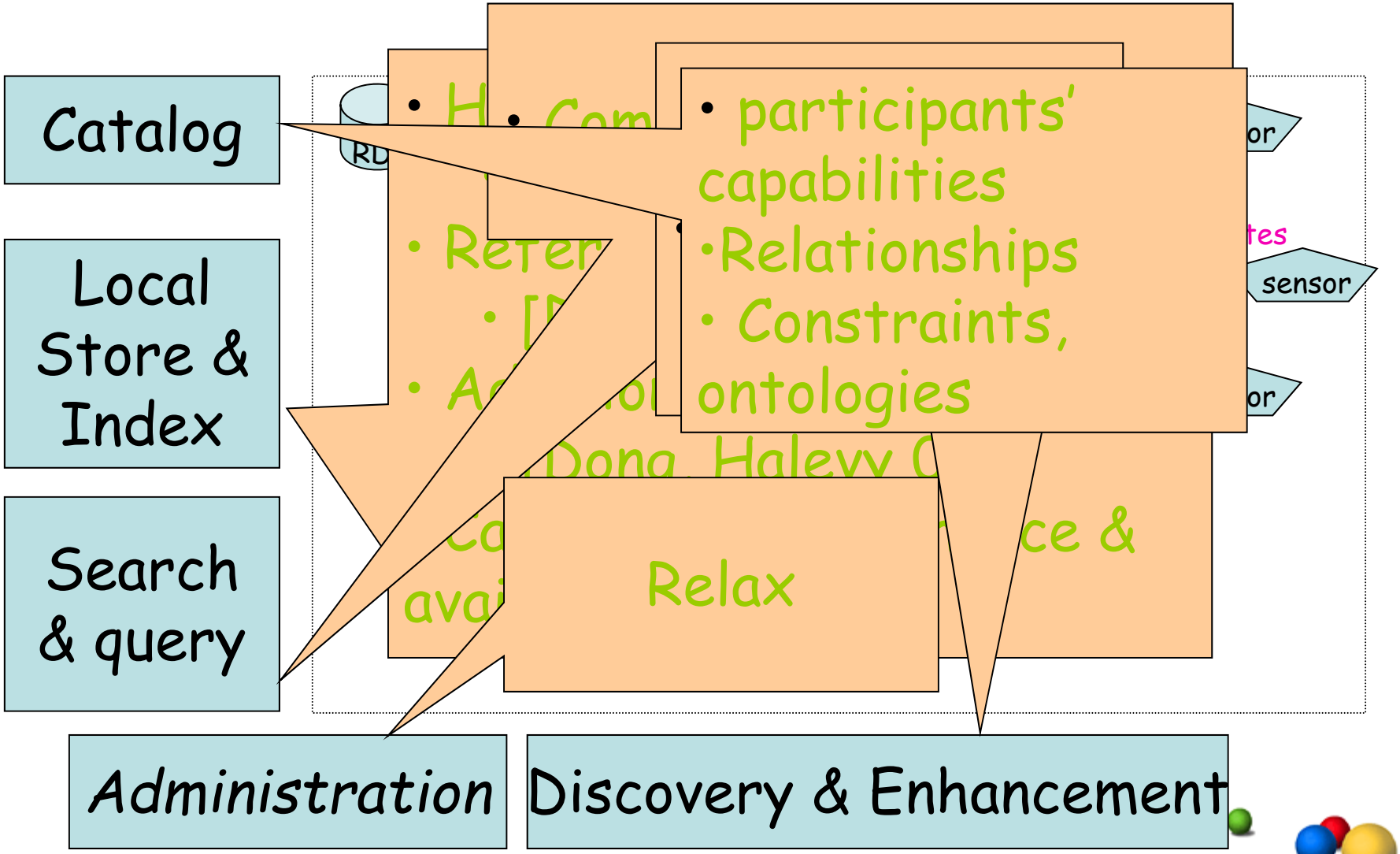
Relationships

- General form:

$$(Obj_1, Rel, Obj_2, p)$$

- Obj_1, Obj_2: instances, sources, fragments,…
- *Rel*: relationship – mapping, co-reference,..
- *p*: degree of certainty about the relationship

Dataspace Support Platforms (DSSP)



Outline

- ✓ Logical model for dataspaces
- ➢ Querying dataspaces
  - Queries
  - The semantics of answers
  - Answering queries
- Dataspace evolution
  - Generating semantic mappings
- Dataspace reflection

Dataspace Queries

- Keyword queries as starting point
  - Later may be refined to add structure
  - Formulated in terms of user's "ontology"
- Mostly of the form
  - *Instance*:*
    - *"britany spears"*
  - *P (instance)*
    - *"lake windermere weather"*

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=

Go weather seattle

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

Google weather seattle Search PageRank Check AutoLink AutoFill Options

Refrigerator Shabbos Times for 5766 | weather seattle - Google Search

Sign in

Web Images Groups News Froogle Maps Desktop Moma more »

Google weather seattle Search Advanced Search Preferences

Weather Seattle

Web ...0 of about **71,300,000** for **weather seattle**. (**0.10** seconds)

Try your sea... ...s, Technorati, Feedster, Wikipedia, Bloglines, Altavista

992 results stored on your computer - Hide - About

My Yahoo! - **Weather** Rehovot 69.86 F Palo Alto, CA 50.72 F **Seattle**, WA 48

Weather for **Seattle, WA**

57°F
Overcast
Wind: W at 6 mph
Humidity: 67%

Thu	Fri	Sat	Sun
62° | 50°	58° | 50°	58° | 51°	64° | 50°

Find more forecasts at Yahoo, Ask, Netscape, CNN, USA Today, Ameriwx, Weather Underground, Weather.com, AccuWeather

NWsource: **Weather**: **Seattle**, Washington
Browse **weather**. **Seattle** forecast · **Weather** maps · Snow report · Web cams · Tide tables · **Weather** links. tools & services. Your account · E-mail newsletters ...
www.nwsource.com/**weather**/scr/ - 28k - May 23, 2006 -
Cached - Similar pages - Filter

Sponsored Links

Seattle Weather
KING5 **Seattle** area **weather**, forecasts, radar images, and more.
www.KING5.com

Weather Seattle
Scenic Byways, Quaint Towns & More! Travel through Washington & SayWA.
www.ExperienceWA.com
Washington

Weather Seattle
Here are the top 8 sites on **Weather Seattle**
www.8bestsites.com/**Seattle**

Done

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

Acura Integra Palo alto

acura integra palo alto - Google Search - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.google.com/search?q=acura+integra+palo+alto&start=0&ie=utf-8&oe=utf-8&clier

Customize Links | Free Hotmail | Windows Marketplace | Windows Media | Windows

acura integra palo alto - Google... | SIGMOD/PODS 2006 Conference

Sign in

Web

acura integra palo alto

Web Results **1** - **10** of about **85,500** for **acura integra palo alto**. (**0.14** seconds)

Sunnyvale **Acura** — Sponsored Link
www.sunnyvaleacura.com The Bay Area's **Acura** Dealer, Large Inventory, Instant Price Quotes

3 results stored on your computer - Hide - About
Framework for structured.. - **Palo Alto**. The advantage of such summary - Nov 23
Google Base - Report bad item WHITE 1999 **ACURA INTEGRA**-FOR SALE IN SANTA - Nov 14

Refine your search for **acura integra palo alto**

Location	Make	Model	
palo alto	acura	integra	Search vehicles

Remember this location

Compare Prices and Read Reviews on 1996 **Acura Integra** Coupe at ...
I own a 1996 **Acura Integra** RS coupe that I've had for three years now. ... Location: **Palo Alto**, CA Reviews written: 11 Trusted by: 294 members ...
www.epinions.com/auto-review-56BB-4C1D058-37C89626-bd4 - 58k - Cached - Similar pages

Sponsored Links

Mtn. View **Acura** Repair
Larry's AutoWorks
When you want it right.
www.autoworks.com
San Francisco-Oakland-San Jose, CA

Palo Alto Acura
Search Local Dealer Inventory & Request a Free Price Quote Today!
AutoWestDirect.com

Honda **Palo Alto**
Low prices, no hassle, great deals at Anderson Honda in **Palo Alto**, CA.
www.AndersonDirect.com/
San Francisco-Oakland-San Jose, CA

Done

volvo palo alto - Google Search - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.google.com/search?q=volvo+palo+alto&start=0&ie=utf-8&oe=utf-8&client=fire

Go volvo palo alto

Customize Links Free Hotmail Windows Marketplace Windows Media Windows

Google volvo palo alto Search PageRank Check AutoLink AutoFill Options volvo palo alto

volvo palo alto - Google Search | Caribou - Inbox | Caribou - Inbox | SIGMOD/PODS 2006 Conference

Sign in

Web Images News Froogle Local Desktop Moma more »

Google Search Advanced Search Preferences

Volvo Palo alto

Web Results **1** - **10** of about **255,000** for **volvo** **palo alto**. (**0.42** seconds)

Sponsored Links

Carlsen **Volvo**
CarlsenVolvo.com Your SF Bay Area Discount Dealer! Serving Northern California

Volvo Auto Repair
www.autoworks.com **Volvo** experts in Mountain View, CA Larry's AutoWorks

Local results for **volvo** near **Palo Alto, CA**

Volvo Parts-Carlsen **Volvo** - 2.4 miles SE - 4190 El Camino Real, Palo Alto, 94306 - (650) 493-1515
BMW-Mercedes-**Volvo** Service & Repair - 2.4 miles SE - 830 E Charleston Rd, Palo Alto, 94303 - (650) 857-1240
Smythe **Volvo**: New & Used Car Sales - 12 miles SE - 4500 Stevens Creek Blvd, San Jose, 95129 - (408) 983-2400

Palo Alto Auto Parts
4190 El Camino Real, **Palo Alto**, CA 94306. Car & Auto Dealers: New & Used Cars • Car & Auto Dealers: **Volvo** • Car & Auto Parts & Accessories • Car & Auto ...
local.sanfrancisco.com/Palo+Alto/Auto+Parts.zq.html - 53k - Cached - Similar pages

Sponsored Links

Volvo at Carmax
Actual Prices & Photos of Over 20,000 New & Used Vehicles Online
www.carmax.com

Volvo Palo Alto
Use Our Handy Map to Locate a **Volvo** Dealer in Your Area.
www.BayArea-VolvoDealers.com

Volvo Research
Get Unbiased **Volvo** Info
Free Price Quotes at edmunds.com!
www.Volvo.edmunds.com

Done

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

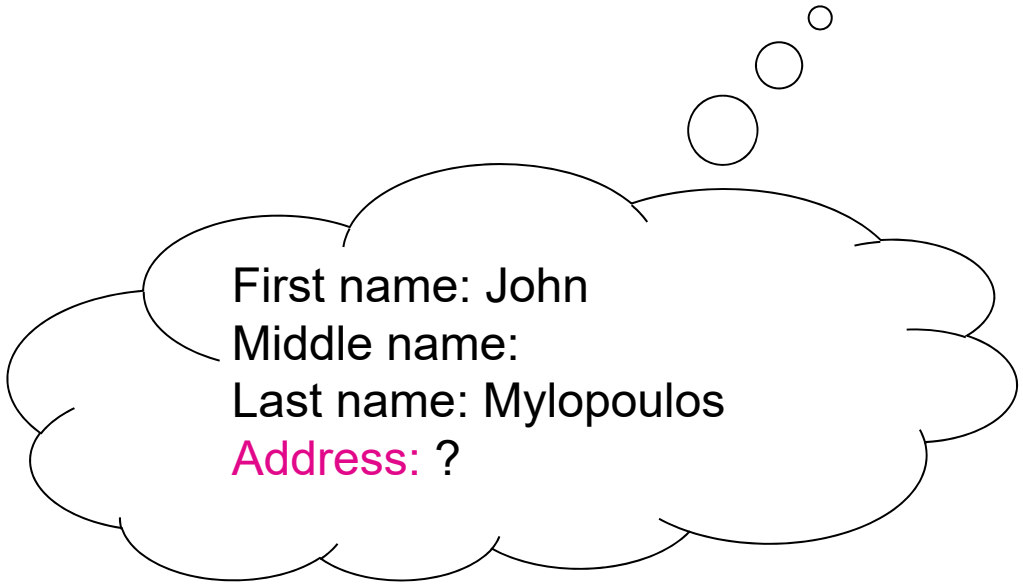
Related or Partial Answers

- In which country was John Mylopoulos born?
  - Athens
- Latest edition of software X:
  - 2004 edition
- Is the space needle higher than the Eiffel Tower?
  - Height of Space Needle, height of Eiffel Tower

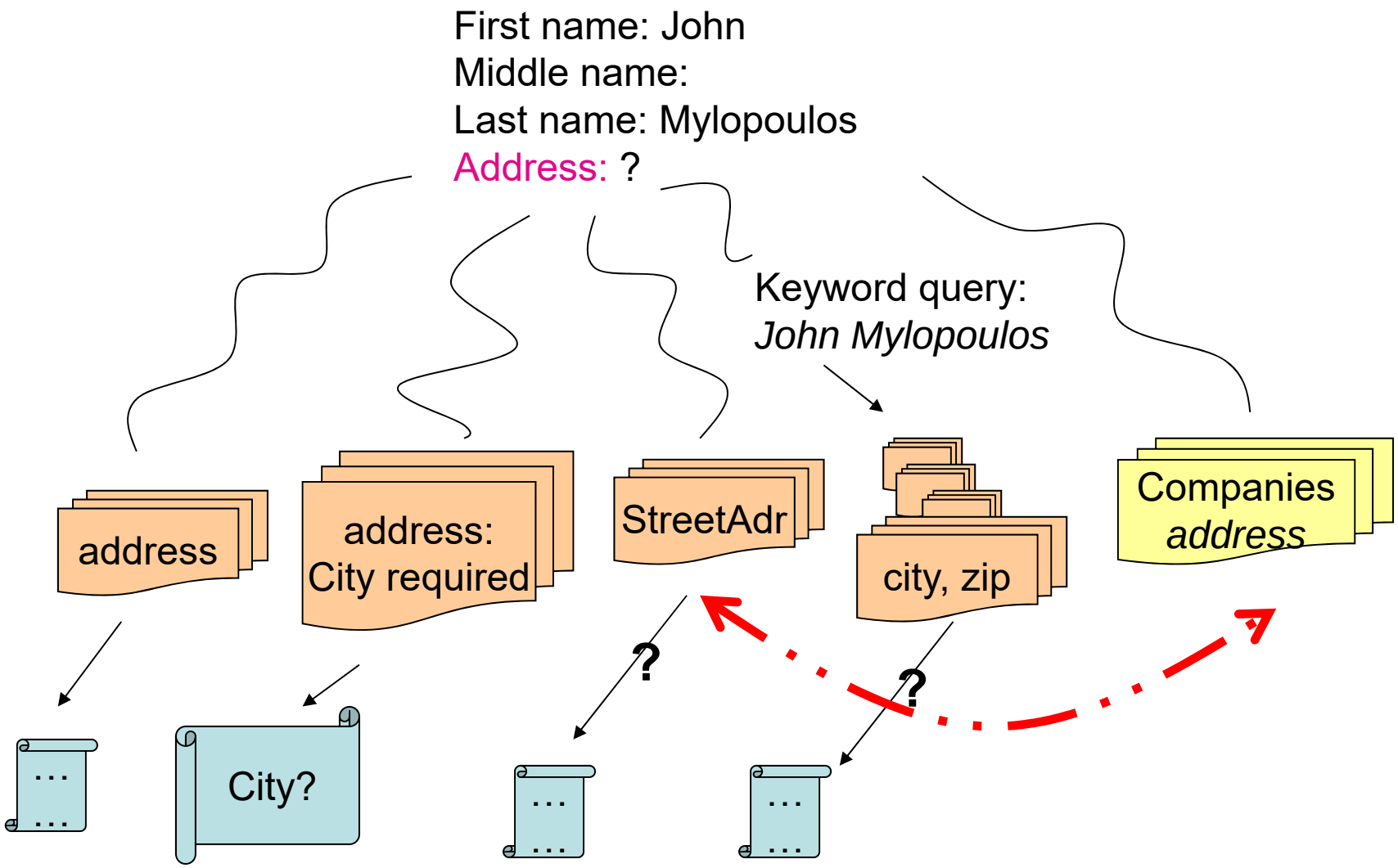
➢ Ranking answers of all types

Query Processing: DSSPs

Query: John Mylopoulos address



Query Processing: Evidence Gathering



Issues: uncertainty, belief revision, truth maintenance, …

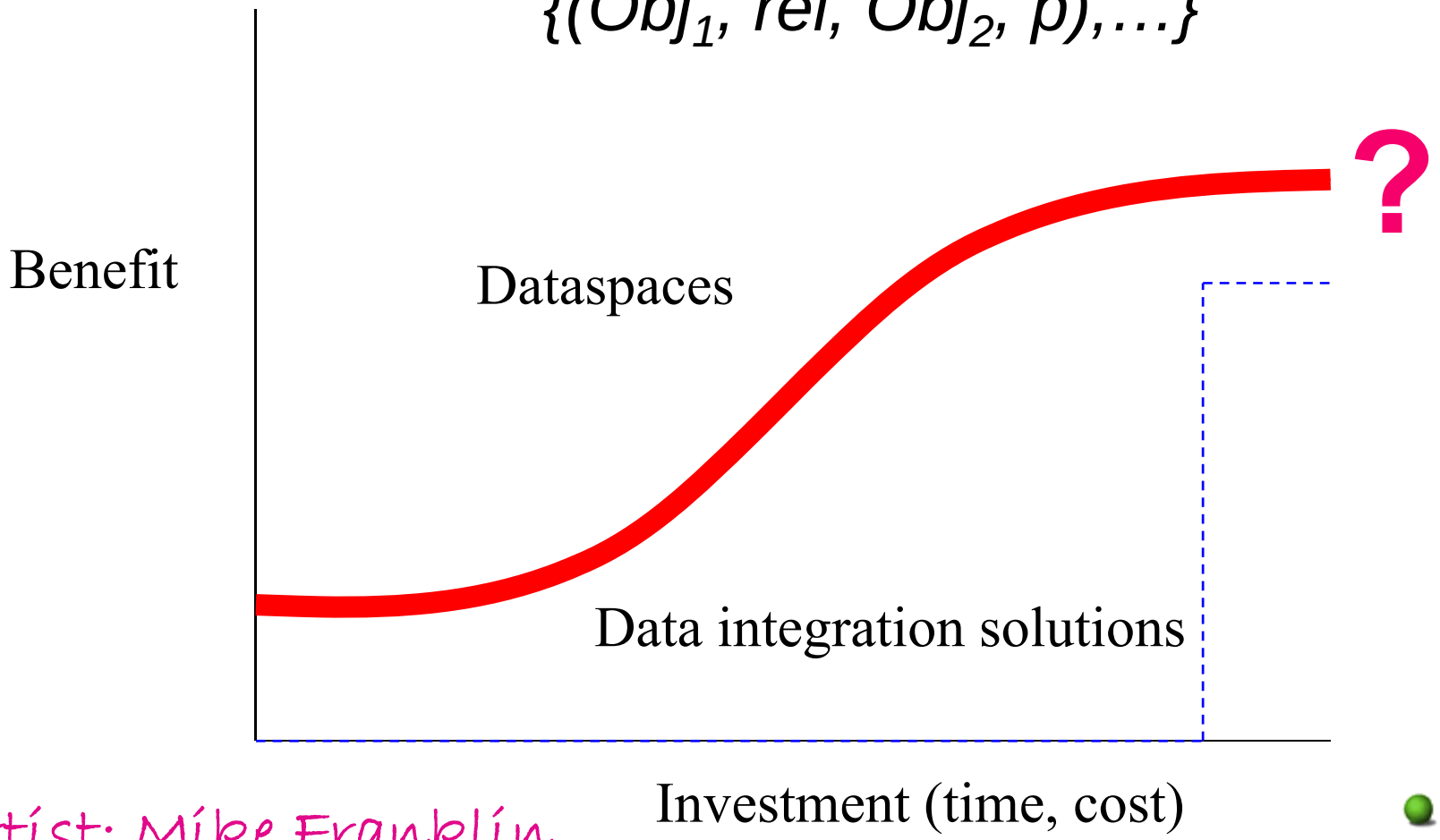
Outline

- ✓ Logical model for dataspaces
- ✓ Querying dataspaces
- ➢ Dataspace evolution
  - – Reusing human attention
  - – Corpus-based ontology matching
  - – Other examples of the reuse principle
- Dataspace reflection

The Cost of Semantics

Semantic integration modeled by:

{(Obj$_1$, rel, Obj$_2$, p),…}



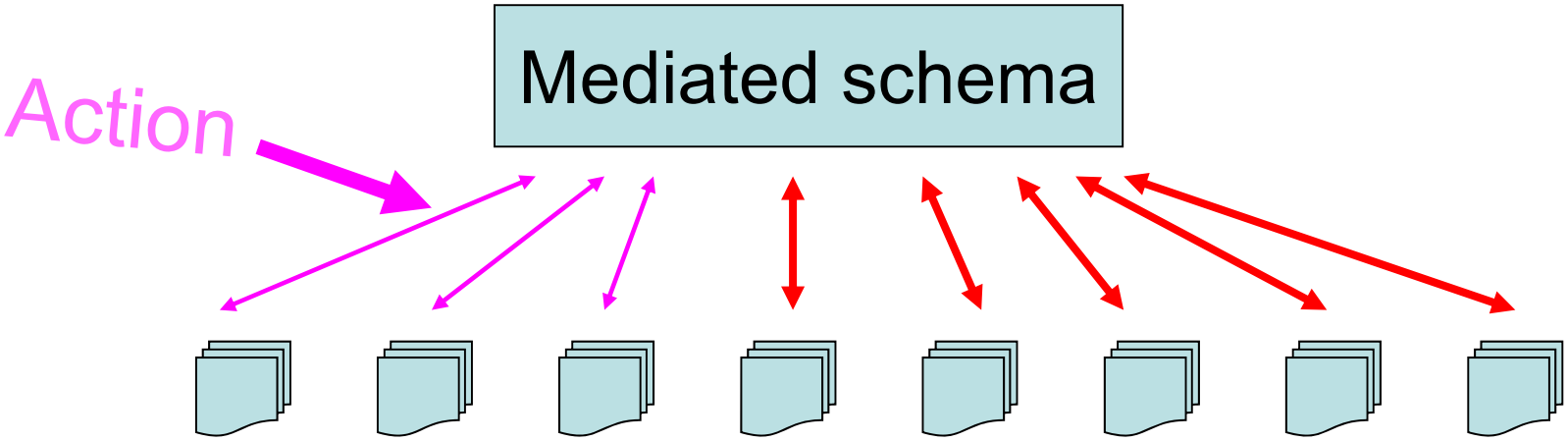
Artist: Mike Franklin

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 content of sources, 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

Learning Schema Mappings

[Doan et al., ACM Thesis Award, Transformic]



- Learn classifiers for elements of the mediated schema

➢ *Thousands of web forms mapped in little time*

❖ [Madhavan et al.]: infer mappings for any schemas in the domain

Corpus-based Matching

Product

productID	name	price	salePrice
0X7630AB12	The Concert in Central Park	$13.99	$11.99

Music

ASIN	title	artists	recordLabel	discountPrice
				(no tuples)

[Madhavan et al., 2005]

Obtaining More Evidence

Product CD

productID prodID	name albumName	price	salePrice
0X7630AB12	The Concert in Central Park	$13.99	$11.99

Corpus

MusicCD

ASIN	album	artistName	price	discountPrice
4Y3026DF23	The Best of the Doors	The Doors	$16.99	$12.99

CD

prodID	albumName	artists	recordCompany	price	salePrice
9R4374FG56	Saturday Night Fever	The Bee Gees	Columbia	$14.99	$9.99

Comparing with More Evidence

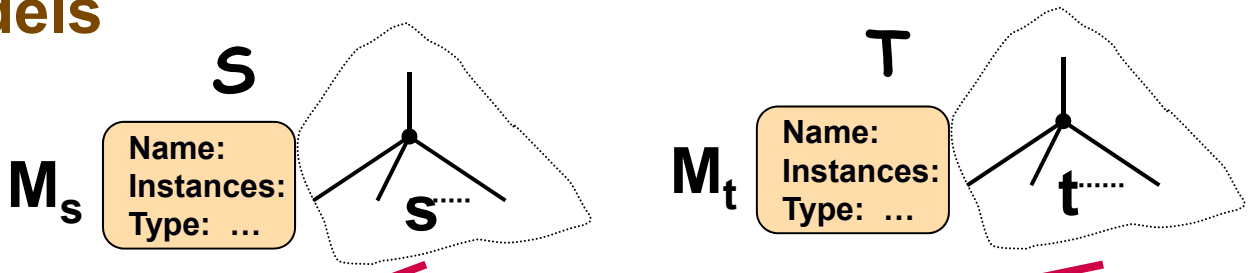
Product CD

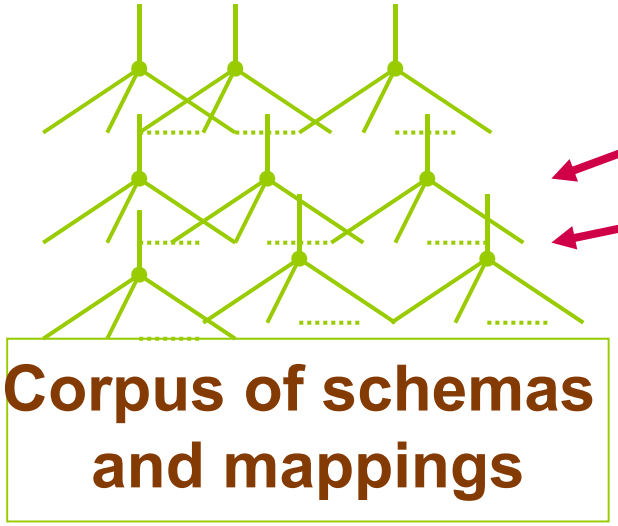
productID prodID	name albumName	price	salePrice
0X7630AB12	The Concert in Central Park	$13.99	$11.99

Music MusicCD

ASIN	Title album	artists artistName	recordLabel recordCompany	discount price
4Y6DF23	The Best of the Doors	The Doors	Columbia	$12.99

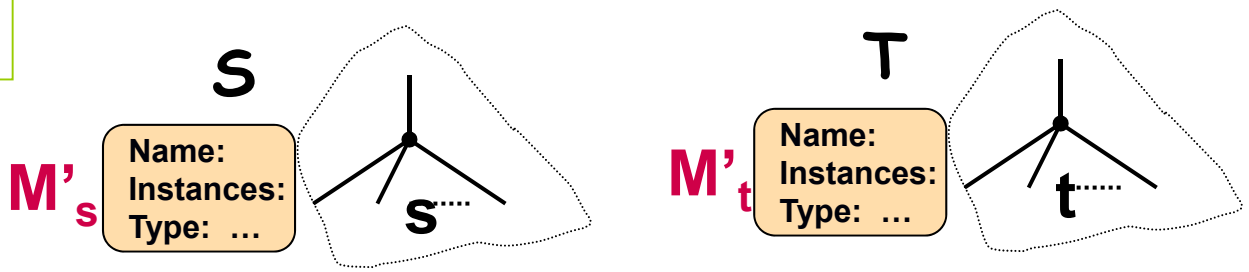
1. Build initial models





2. Find similar elements in corpus

3. Build augmented models



4. Match using augmented models

5. Use additional statistics (IC's) to refine match

Learning from Query Logs

- Action: posing a query
- What's interesting about it?
  - values used in the query
  - joins across sources (even implicit ones)
- We can derive:
  - What is a data source about
  - Constraints on its contents
  - Relationships between sources

Dimensions of Reuse

- Actions
- Generalization mechanism
- Learning from:
  - past actions & existing structure:
    - [Dong et al., 2004, 2005], [He & Chang, 2003]
  - Current actions
  - Requesting new actions:
    - ESP [von Ahn], mass collaboration [Doan et al], active learning [Sarawagi et al.], CbioC (ASU.edu)

Outline

- ✓ Logical model for dataspaces
- ✓ Querying dataspaces
- ✓ Dataspace evolution
- ➢ Dataspace reflection
  - – Uncertainty, lineage, inconsistency, …

Dataspace Reflection

- Life is uncertain with dataspaces:
  - Answers are derived with imprecision
  - Semantic relationships are uncertain
  - Data sources may be imprecise
  - Data extraction (structuring) often imprecise
- Data will often be inconsistent
  - No way to enforce integrity constraints
- Answers meaningless without the “how”

The Main Challenge
(KR to the Rescue)

- Need a single formalism for modeling:
  - uncertainty,
  - inconsistency, and
  - lineage (how a tuple/answer was derived)
  - Incompleteness
- DB community starting to think about combining these formalisms.

Israel population

population israel - Google Search - Mozilla Firefox

el&start= Go population israel

dows Crossing the Structur...

eck AutoLink AutoFill Options

engEDU Computer ScienceTraining Prog... | KLM Royal Dutch Airlines Comprehensiv... | population israel - Google Search

Israel — **Population**: 6,276,883
According to http://www.cia.gov/cia/publications/factbook/rankorder/2119rank.html - More sources »

News results for **population israel** - View today's top stories

Israel fumes as Sweden grants visa - Jerusalem Post - 1 hour ago
Olmert's government takes office with mandate to set **Israel's ...** - Canada.com - 7 hours ago
Israel marks 58th Independence Day with countrywide celebrations - Ha'aretz - May 3, 2006

CIA - The World Factbook -- **Israel**
Population:. Definition · Field Listing · Rank Order 6352117 note: includes about 187000 **Israeli** settlers in the West Bank, about 20000 in the **...**
www.cia.gov/cia/publications/factbook/geos/is.html - 112k - May 3, 2006 -
Cached - Similar pages - Filter

Israeli Population Statistics
Israeli Population Statistics. Year. 2004. 1999. 1996. 1995. 1990 **...** 122.5, 123.8, 125.1, 115.5, 89.7, 77.1. Source: **Israeli** Central Bureau of Statistics.
www.jewishvirtuallibrary.org/jsource/Society_&_Culture/demographics.html - 32k -
Cached - Similar pages - Filter

Population in **Israel**
The estimates are based on registration of **population** movements received by the end of **...**
Copyright © 1997- The State of **Israel**. All rights reserved. See **...**
www.cbs.gov.il/**population**/popul_eng.htm - 8k - Cached - Similar pages - Filter

http://www.cia.gov/cia/publications/factbook/rankorder/2119rank.html

Uncertainty and Inconsistency

- Inconsistency = uncertainty about the truth
  - Salary (John Doe, $120,000)
  - Salary (John Doe, $135,000)
  - ➢ Salary (John Doe, $120,000 | $135,000)

- Orchestra @ U. Penn:
  - Allow inconsistent data, but ensure that lineage is tracked.

Conclusion and Outlook

- Data management moving to consumer market
  - But it's messy, and we need to live with it
- Dataspace framework offers:
  - Pay as you go data management
  - Evolution by reusing human attention
- The role of KR:
  - Fanciness needed to navigate messy spaces
  - Reflection: certainty, belief revision, data gaps

Some References

- SIGMOD Record, December 2005:
  - Original dataspace vision paper
- PODS 2006:
  - Specific technical challenges for dataspace research
- Semex: an example dataspace system
  - [Dong et al., 05, 06]
- Teaching integration to undergraduates:
  - SIGMOD Record, September, 2003.