

CSCI-548:

Information Integration on the Web

Spring 2015

Jose Luis Ambite
Craig Knoblock
University of Southern California



Course Overview

Information Integration Challenges

- Accessing the data
- Understanding the data
- Resolving differences
 - Schema-level
 - Data-level
- Efficiency
 - Query evaluation
 - Data transformation

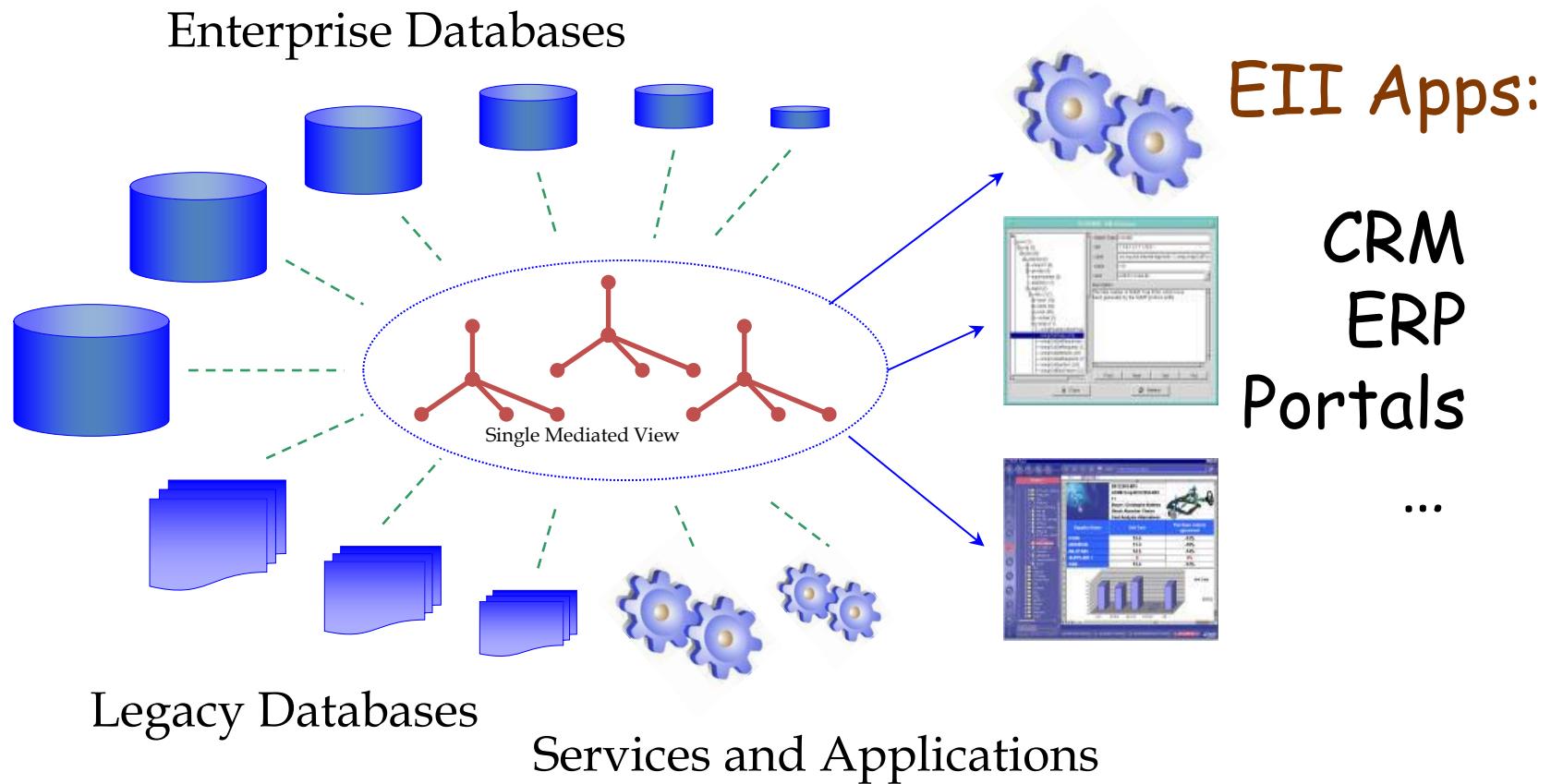


...on the Web

- The Web is an incredible source of data
- New challenges arise:
 - Need to turn web pages into structured data
 - Don't have control over the data
 - Sources have input/output constraints
 - Distributed nature of the web can make integration slow
 - How to relate information across sources

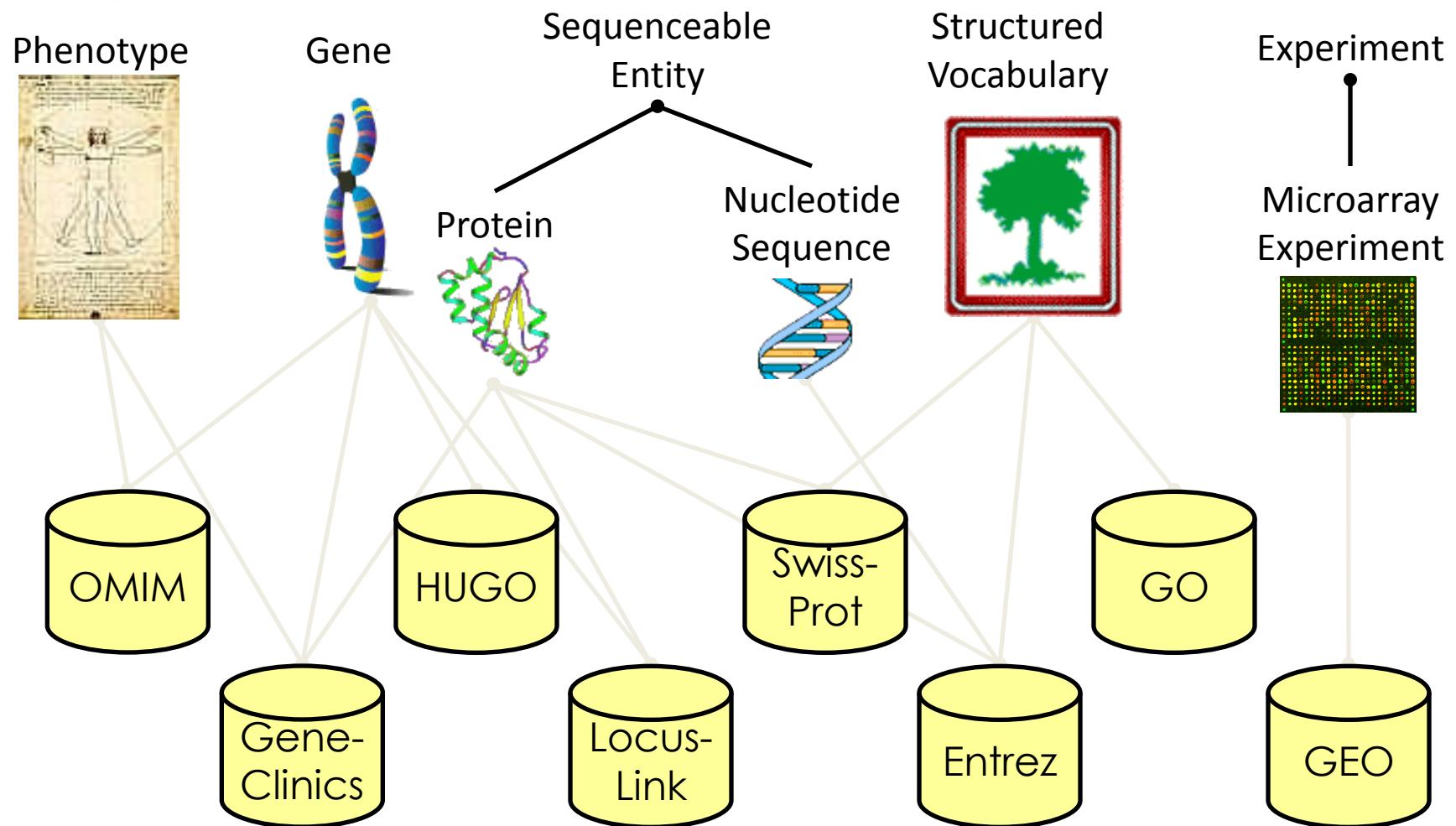


Application Area 1: Business



50% of all IT \$\$\$ spent here!

Application Area 2: Science



Hundreds of biomedical data sources available; growing rapidly!

Application Area 3: Web & Deep Web



Presenting the exciting world
ybw.com
books & charts

Quilt-Books.Com

KATSUKI
BOOKS 日本書店

Shelf
Books.com

FOREIGN
NATIONAL
BOOKS
International Books

COMPU-BOOKS



USBORNE
CHILDREN BOOKS

66books.com
God Gave His Word

Providence-Books
www.provdeuce-books.com



BJ
NGHOSSI
books



BORDERS.com
10 million books, CDs, and videos

CAECOTUS-BOOKS.COM
The Best Choice for Many
metro-Books

BARNES
& NOBLE
www.bn.com

StoneCreekBooks.com

ASTROLOGY-BOOKS.COM
Brought to you by SuperStar Books

SeeMe4Books

amazon.com.

HALF PRICE COMPUTER BOOKS
DANCE BOOKS

acma
BOOKS
.COM

NORTH
49
BOOKS

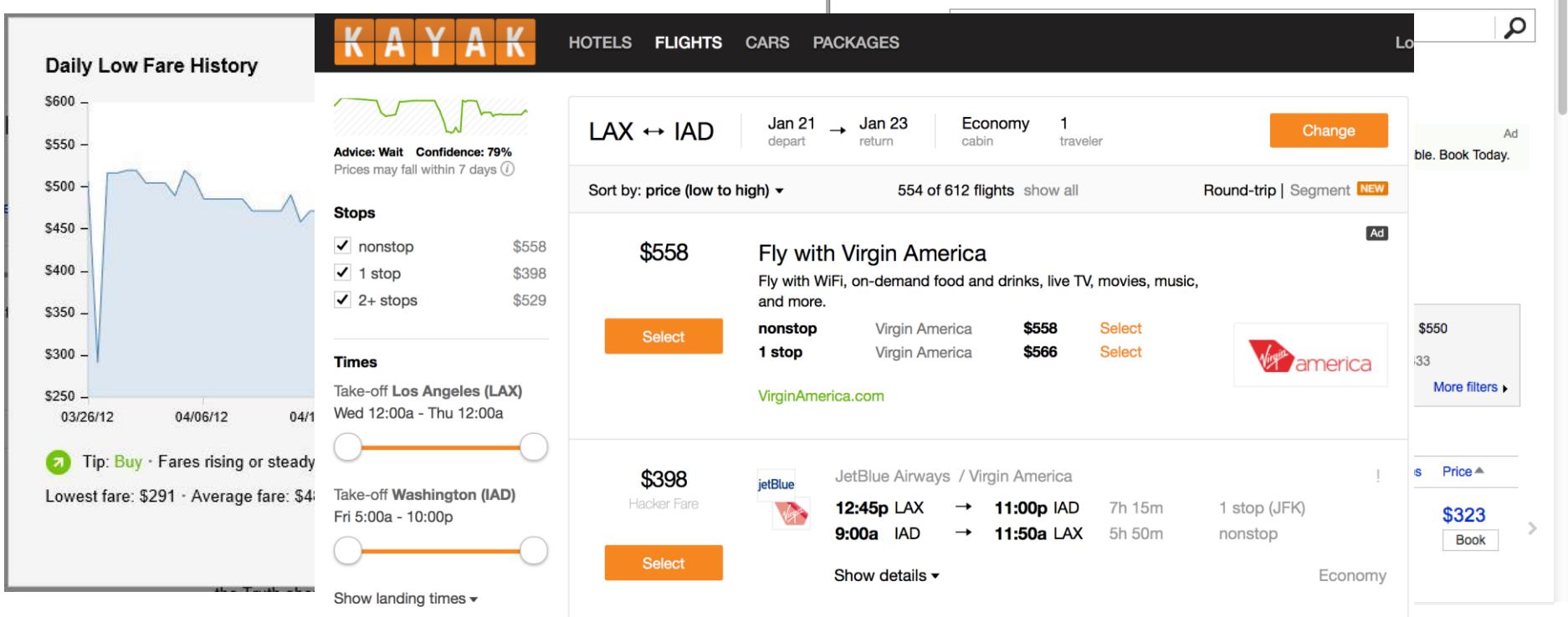
funny face
Bookstore
SE BARTON'S

Professional
Books Inc.

Just
Great
Books

Web Example: Airfare Prediction

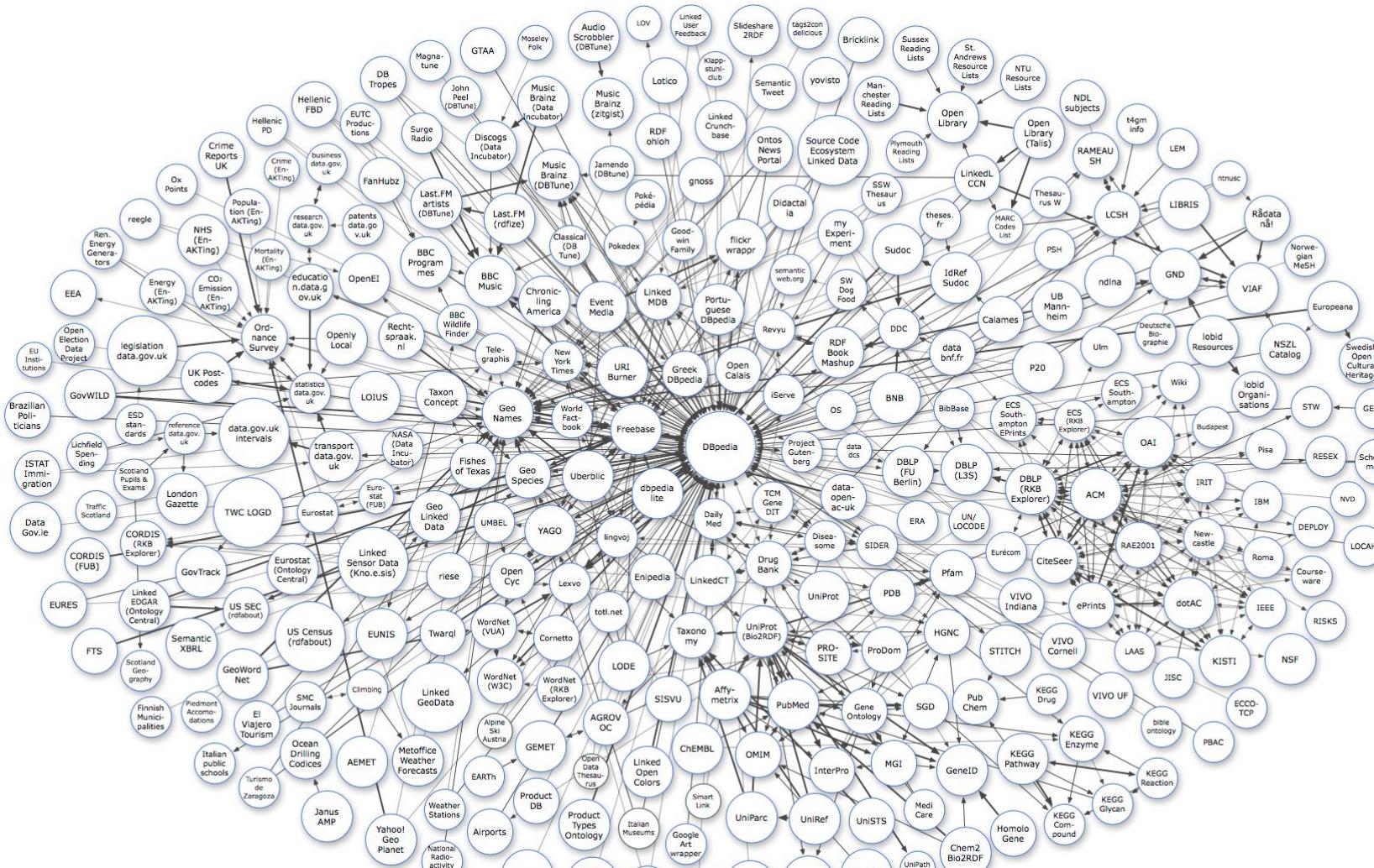
- Monitor, integrate flight prices across airlines
- Predict cheapest time to buy flight



- [Etzioni, Knoblock, et al. KDD 2003], lead to startup: Farecast, bought by Microsoft, incorporated in travel.bing.com, now in Kayak



Application Area 4: Semantic Web & Linked Data



Application Area 5: Big Data

- Dimension of Big Data:
 - Volume: size of data in bytes, files, or data objects.
 - Velocity: rate or frequency at which new data can be acquired or processed.
 - Variety: diversity in data sources, data models, data types: geospatial, genetic, imaging, relational, XML, web services, RDF, ...
 - Variability: diversity in data interpretation or meaning.
- This course address Variety and Variability
- Without semantics, big data = big mess

Overview of Lecture Topics

Semantic Web

- A machine-readable semantic layer on the web
- Research Topics
 - Organizing knowledge
 - Formal languages: XML, RDF, OWL, FOL
 - Reasoning and querying over semantic web data
 - Provenance, Trust
 - Linked Data



Can You Google the Answer to These Questions?

- Which US cities have more than 500,000 people?
- Which congress members voted “No” on pro environmental legislation in the past four years, with high-pollution industry in their congressional districts?
- How many Twitter followers does Obama have in each state?
- Find all stores that sell a Canon 7D for less than \$1,700



SATURDAY, MARCH 15, 2003

GOOGLE TO HARVEST RDF-A

Search giant goes semantic

Norris Center for the Performing Arts, 27570 Crossfield Drive in RHE, on Saturday, March 20 at 2 and 4 p.m. Tickets are \$38 for adults and \$18 for students for the evening performance and \$30 for adults and \$15 for students at the matinee performance. For reservations, call 544-0403.

* UPCOMING — The Palos Verdes Peninsula Unified School District and Friends of School Music host the 15th Palos Verdes Elementary Choral Festival on March 23, 24 and 25 at the Norris Center for the Performing Arts, 27570 Crossfield Drive in RHE. All shows begin at 7:30 p.m. For tickets, call the Norris box office at 544-0403.

* ONGOING — The Distinctive Edge, 29050 S. Western Ave., Suite 113 in RHE, continues "Third Time's a Charm," an exhibit of 3-D collages by artist Steve Jacobsen, through March 30. For gallery hours, call 833-3613.

* ONGOING — "Natural Treasures" exhibition contin-

Two years ago, U.S. Navy personnel and their families assigned to the Atsugi Navy base, home of the U.S.S. Kittyhawk, were treated to a rare experience when Terry Fleming and his local British-American band, Innisfree, travelled to the base to entertain them on St. Patrick's Day. Fleming and the other five members of Innisfree were delighted and honored to be able to go to Japan and lift the spirits, if only for a few hours, of the Navy personnel and their families. For the third year in a row, Fleming — a local insurance broker in Rolling Hills by day and an entertainer by night — and the band travel to entertain the Navy men, women and families at various bases throughout Japan.

Fleming, the leader of the band on accordion and harmonica, actually is the only member of the band from Ireland. Other members include lead singer Julie Delaney, a civil engineer in Newport Beach; Terry Doyle, gitarist, a news director with CBS news; Dennis Doyle, Celtic harpist, a professor at Glendale College; Kevin Woods, keyboards and bagpipes, music

teacher and assistant director of the Orange County Symphony; and Mike Tiffany, bass, a computer engineer. The band has been playing the length and breadth of California for the past 25 years. They have played at graduations, weddings, birthdays and on occasions where there was little excuse for throwing a party.

Fleming says it was by coincidence the band got the opportunity to travel to Japan. Another band was unable to travel at the last minute and so he and his band were offered the opportunity to go in their place.

With some trepidation they made their first trip and with the overwhelming response they received at Atsugi, any fears they had were quickly allayed. On a damp St. Patrick's Day, hundreds of families, clad in many shades of green, whooped it up, sang their hearts out and danced up a storm. As the evening wore on, many in the audience were emboldened to try their hand or foot at the Irish gig, with much encouragement from the band.

Even though far from home, the Atsugi base — situated a few hours

south of Tokyo — felt like home away from home, with its lush green rolling landscape and its multitude of cherry blossom trees. "Yes," Fleming says,

"we were struck by the commitment and dedication of our men and women in uniform as they played their part in protecting and serving in

an ever-challenging and hostile

world."

and entertained the locals for a few days-filled hours. It turned out that it was just one of many establishments in the city.

A special bond developed between the band members and these families and already exchange visits have occurred when the same families were on leave in the United States.

For more information about the band, log on to www.bringingbackmusic.com.





New York Times Index

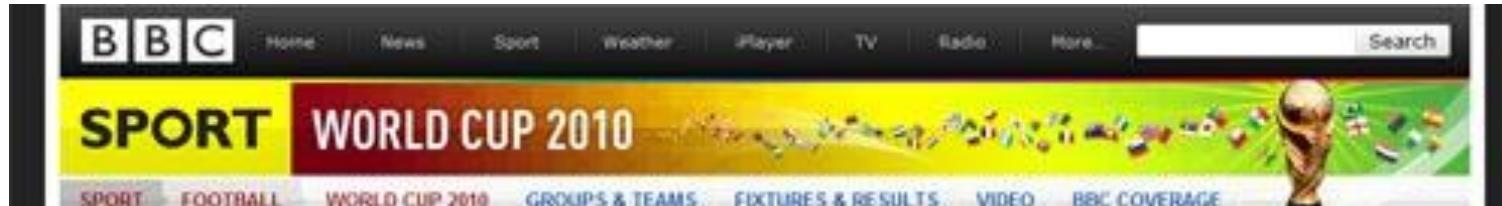
- Started in 1851,
cross ref'd every NYT article,
published since 1913
- 10k tags on all kinds of topics
- 3500 in very frequent use
- Open License, usable as
reference ontology for anybody
- BTW: strong internal use of semantic
technologies for ... targeting advertising



BBC World Cup Web site

BBC Home News Sport Weather Player TV Radio More Search

SPORT WORLD CUP 2010



SPORT FOOTBALL WORLD CUP 2010 GROUPS & TEAMS FIXTURES & RESULTS VIDEO BBC COVERAGE

Blogs

Phil McNulty


Cool Rooney key to England hopes
Manchester United striker needs to curb temper in World Cup

David Bond


Paying the penalty
David Bond on new spot-kick rules and the threat of match-fixing

Paul Fletcher


Like father, like son
The US family plotting England's World Cup downfall

No fears over Rooney - Ferdinand



Klopp Ferdinand says his England and Man Utd team-mate Wayne Rooney will have no disciplinary issues in South Africa.

- I heard knee snap - Ferdinand
- Ferdinand staying upbeat
- Ref warns Rooney over temper
- Máner suffers high temperature
- Rooney satisfied with win
- Winning all-important - Delfoe
- Terry admits to altitude concerns
- Barry to miss England opener

Portugal's Nani out of World Cup



Tuesday's World Cup 2010 round-up



- Terry admits altitude concerns
- World Cup nerves for South Africa
- England call too late - Scholes
- Robben confident of quick return
- Demerit warns US of Rooney
- Nigerians defend venue choice
- Ivory Coast hopeful over Drogba

Friday's matches

South Africa v Mexico	15:00
Uruguay v France	19:30

Group tables

GROUP A TEAMS	W	D	L	GO	PTS
France	0	0	0	0	0
Mexico	0	0	0	0	0
South Africa	0	0	0	0	0
Uruguay	0	0	0	0	0

TEAM TRACKER
Follow your team and make predictions here



Video choice

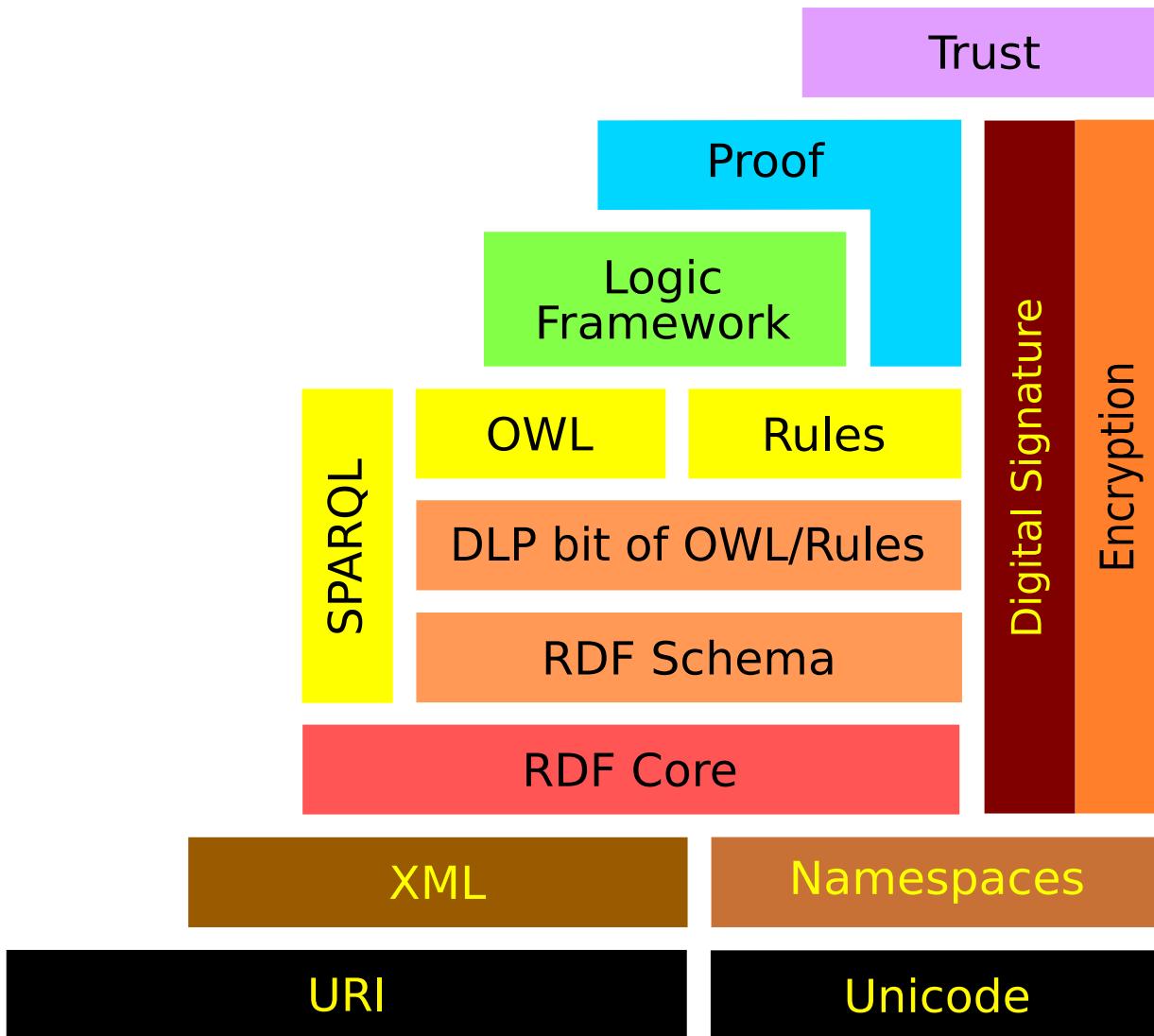
World Cup legends - Pele  Watch

World Cup legends - Zidane  Watch

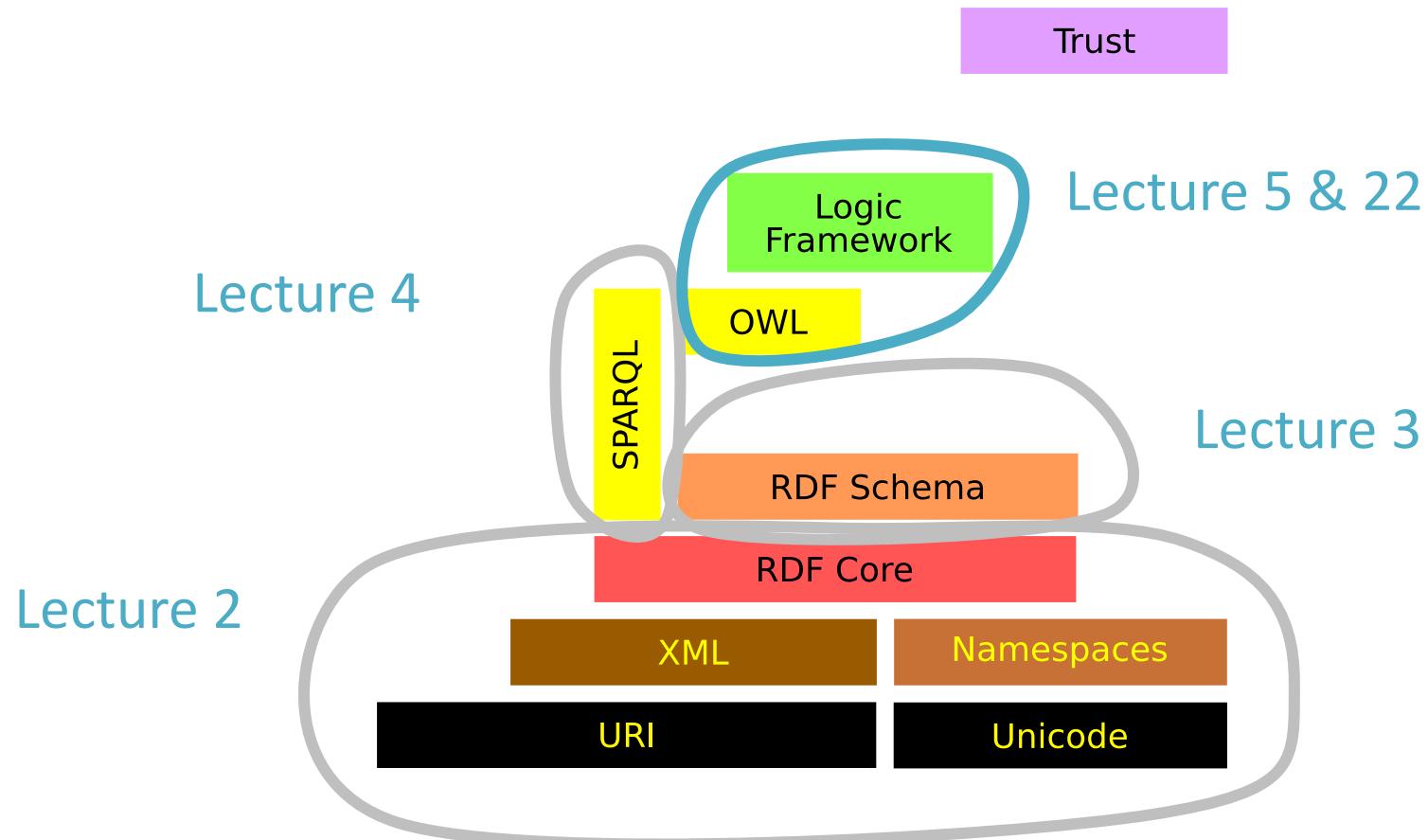
2006 World Cup - Top 10 goals  Watch

slide from Frank van Harmelen

Semantic Web Layer Cake



Semantic Web Layer Cake



Semantic Web: Description Logics, Ontology Web Language (OWL2)

formal underpinning of the semantic web

Constructor	DL Syntax	Example	FOL Syntax
intersectionOf	$C_1 \sqcap \dots \sqcap C_n$	Human \sqcap Male	$C_1(x) \wedge \dots \wedge C_n(x)$
unionOf	$C_1 \sqcup \dots \sqcup C_n$	Doctor \sqcup Lawyer	$C_1(x) \vee \dots \vee C_n(x)$
complementOf	$\neg C$	\neg Male	$\neg C(x)$
oneOf	$\{x_1\} \sqcup \dots \sqcup \{x_n\}$	{john} \sqcup {mary}	$x = x_1 \vee \dots \vee x = x_n$
allValuesFrom	$\forall P.C$	\forall hasChild.Doctor	$\forall y.P(x,y) \rightarrow C(y)$
someValuesFrom	$\exists P.C$	\exists hasChild.Lawyer	$\exists y.P(x,y) \wedge C(y)$
maxCardinality	$\leq n P$	≤ 1 hasChild	$\exists^{\leq n} y.P(x,y)$
minCardinality	$\geq n P$	≥ 2 hasChild	$\exists^{\geq n} y.P(x,y)$

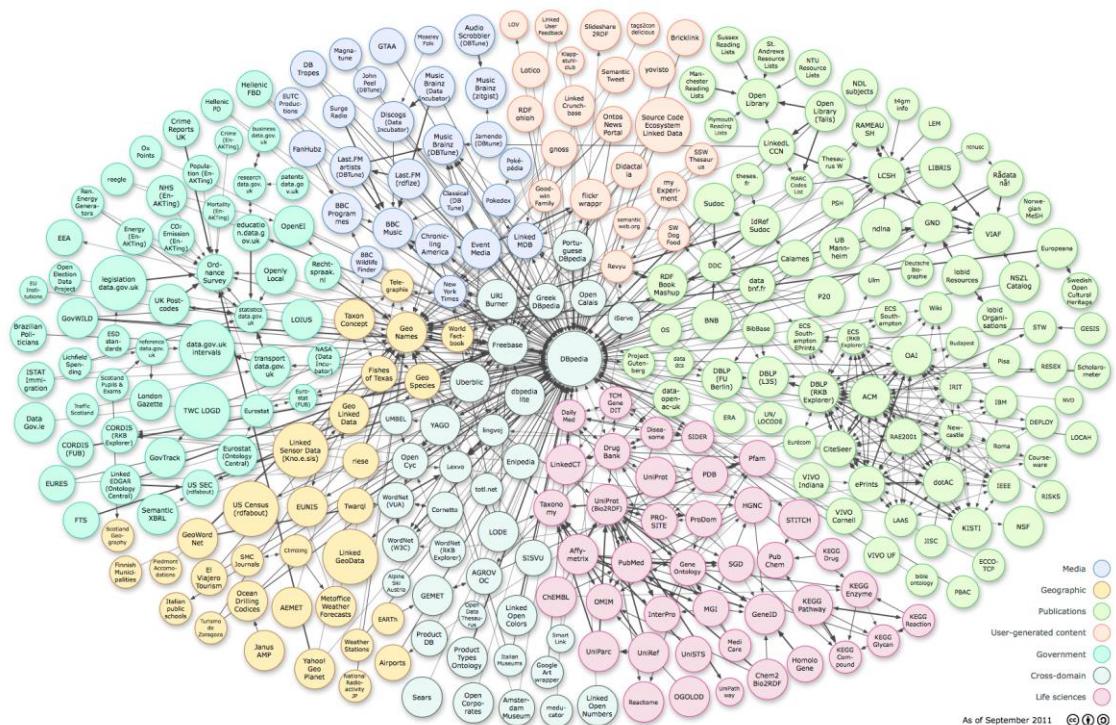
HappyParent = Person \wedge \forall hasChild.(Doctor \vee Rich)



Linked Data

a giant Web-scale graph of data about everything

Wikipedia: "...
recommended
best practice for
exposing, sharing,
and connecting
pieces of data,
information, and
knowledge on the
Semantic Web
using URIs and
RDF."



31+ billion assertions



Linked Data

- Why Linked Data
- The Linked Data principles
- Web of data
- How to Publish Linked Data on the Web
- Tools
- Applications



Services Over Linked Data

<http://www.programmableweb.com>

The screenshot shows the homepage of ProgrammableWeb. At the top, there are logos for programmableweb, Site24X7.com, and Web Performance. Below the header, a navigation bar includes links for Home, API News, API Directory, Mashups, Community, and How-to. A banner below the navigation bar reads "Keeping you up to date with APIs, mashups and the Web as platform. Learn more »". On the left, a search bar says "Find APIs, mashups, code and developers" with a "Search" button. Below it, "Popular searches" include photo, google, flash, mapping, enterprise, and sms. In the center, a large box displays statistics: 8393 APIs and 6897 Mashups. On the right, sections for "New APIs" and "Mashup of the Day" are shown, along with a list of "New Mashups". A footer at the bottom includes a Creative Commons license logo (CC BY NC SA) and credits to Ambite, Knoblock, and Szekely.

programmableweb

Site24X7.com

Web Perform

Supports: Web Apps, Server
40+ loca

Hot APIs » Twitter YouTube Facebook Google Maps Flickr LinkedIn More »

Latest news 3D System

Home API News API Directory Mashups Community How-to

Keeping you up to date with APIs, mashups and the Web as platform. [Learn more »](#)

Find APIs, mashups, code and developers [Search](#)

Popular searches: [photo](#) [google](#) [flash](#) [mapping](#) [enterprise](#) [sms](#)

8393 APIs

6897 Mashups

New APIs

- ▶ [MapAlerter](#)
- ▶ [SMTP.com](#)
- ▶ [Charlotte City Club Photo Album](#)
- ▶ [Toronto Pearson Connecting Guide](#)
- ▶ [USGS Contour Service](#)
- ▶ [TEXTKING](#)
- ▶ [See more APIs](#)

Mashup of the Day

GeoBus

▶ See previous winners

New Mashups

- ▶ [Local Geonius](#)
- ▶ [National VIP](#)
- ▶ [TripNotice](#)
- ▶ [The Global Map of Musicians](#)
- ▶ [This is Now!](#)
- ▶ [The Beat](#)
- ▶ [See more mashups](#)

CC BY NC SA

University of Southern California

Ambite, Knoblock and Szekely

Data Cleaning

- Types of dirty data
- Challenges
 - Detection
 - Fixing
 - Large data sets
- Techniques
- Research problems



Data Transformation By Example

content_type

application/pdf

application/pdf

text/plain

text/plain

text/plain

text/html; charset=Windows-1252

text/html; charset=Windows-1252

application/zip

image/jpeg

image/jpeg

Transform

Examples You Entered:

application/pdf

text/html; charset=Windows-1252

Recommended for Examining:

application/zip

All Records:

pdf

html

zip

pdf

pdf

plain

plain

html

html

zip

jpeg

jpeg

Cancel

Submit

The screenshot shows a data transformation interface. On the left, a vertical list of 'content_type' values is displayed. In the center, a 'Transform' dialog box is open. The 'Examples You Entered:' section contains 'application/pdf' and 'text/html; charset=Windows-1252'. The 'Recommended for Examining:' section contains 'application/zip'. The 'All Records:' section lists all the content_type values from the left list. At the bottom right of the dialog are 'Cancel' and 'Submit' buttons.



Database theory basics: queries, query containment, Datalog

- Conjunctive queries
- Query containment
 - $q1(x) :- r(x,x), r(x,y1), r(y1, y2), r(y2, y3), r(y3,x)$
 - $q2(x) :- r(x,x)$ [select * from R(a,b) where R.a=R.b]
 - $q1 = q2 ???$
- Recursive logic programs: Datalog
 - $\text{path}(X, Y) :- \text{arc}(X, Y)$
 - $\text{path}(X, Y) :- \text{path}(X, Z), \text{path}(Z, Y).$

Logical Data Integration

- Warehouses: Extract-transform-load (ETL)
- Virtual Integration: mediators
 - Automatically select and compose information across sources
 - Research Topics
 - Modeling language: Relational, XML, Description Logics (OWL 2 & Profiles)
 - Schema mappings: Logical formulas that relate source and global schemas
 - Global-as-View (GAV), Local-as-View (LAV) integration, more general formulas (GLAV)
 - Scalability: efficient query rewriting, query optimization
- Sample Tools
 - Research: BIRN mediator
 - Commercial: IBM InfoSphere Federation Server, Oracle Data Integrator, Talend, Informatica, Expressor, Infobright, ...



FBIRN Data Integration

The screenshot shows the HID XNAT Integration Portal interface. At the top, there are logos for BIRN, USC, and UCIrvine. Below the header, there are tabs for Subjects, Experiments, Subject Assessments, and Investigators. The main area is titled "Experiment Characteristics" and includes dropdown menus for Scan Type (T1, T2, Structural MRI, Functional MRI, Any), Scanner (GE, Philips, Siemens, Any), and Time Interval. Below this, there are sections for Subject Characteristics, including Race (White, Black, Asian, Native Hawaiian, Native American, Any) and Ethnicity (Hispanic, Not Hispanic, Any). There is also a "Submit Query" button. At the bottom, a table displays results for subjects 000934691111, including Handedness_left_writing (0), Handedness_right_writing (2), Mother's Education (null), and Number of Children (null).

Human Imaging Database(s)
Oracle DB

[Front. NeuroScience 2010]

Results

Source	Subject ID	Age	Gender	Scan Type
XNAT	OAS1_0266_MR1	51	M	t1
XNAT	OAS1_0034_MR1	51	M	t1
XNAT	OAS1_0207_MR1	51	M	t1
HID	000998262706	51	M	t1
HID	000998262706	52	M	t1
HID	000998041611	53	M	t1
HID	000913186207	53	M	t1
HID	000960528669	53	M	t1
HID	000913186207	54	M	t1
HID	000947193547	55	M	t1
XNAT	OAS1_0389_MR1	55	M	t1

result
(XML)

```
<cell>76</cell>
<cell>active</cell>
</row>
- <row>
  <cell>OAS1_0016_MR1</cell>
  <cell><CENTRAL_OASIS_CS></cell>
  <cell>82</cell>
  <cell>active</cell>
</row>
```



EXtensible Neuroimaging Archive Toolkit
Web service API

Schema Mappings

BIRN mediator uses logical schema mappings

- Source-to-target tuple-generating dependencies (st-tgds), aka global/local as view (GLAV) mappings
$$\forall \vec{X}, \vec{Y}, \Phi_S(\vec{X}, \vec{Y}) \rightarrow \exists \vec{Z}, \Psi_G(\vec{X}, \vec{Z})$$
- Intuitively: query over source predicates (antecedent Φ_S), whose answers populate the domain/target predicates (consequent Ψ_G)

S1_EXPSEGMENT(eID, projectID, subjectID, prtclID, date ...) ^

S2_PROTOCOL(prtclID, protocolDescription, ...) →

Scan(eID, subjectID, prtclID, date ...) ^ hasDisease(subjectID, 'Bipolar')

Subject(subjectID, age, 'Male', ...) ^ T1(prtclID) ^ age >= 40

- Mappings used to answer queries or to generate ETL plans

Query Rewriting

- Given Query Q for T1 scans of Male subjects over 50 (over domain schema)

$Q(\text{subjectID}, \text{age}, \text{diseaseID}, \text{eID}) \leftarrow$

$\text{Subject}(\text{subjectID}, \text{age}, \text{'Male'}, \dots) \wedge \text{hasDisease}(\text{subjectID}, \text{diseaseID}) \wedge$
 $\text{Scan}(\text{eID}, \text{subjectID}, \text{prtclID}, \text{date} \dots) \wedge \text{T1}(\text{prtclID}) \wedge \text{age} > 50$

- and Schema Mappings m1 and m2:

m1: $S1_EXPSEGMENT(\text{eID}, \text{projectID}, \text{subjectID}, \text{prtclID}, \text{date} \dots) \wedge$

$S2_PROTOCOL(\text{prtclID}, \text{protocolDescription}, \dots) \rightarrow$
 $\text{Scan}(\text{eID}, \text{subjectID}, \text{prtclID}, \text{date} \dots) \wedge \text{hasDisease}(\text{subjectID}, \text{'Bipolar'})$
 $\text{Subject}(\text{subjectID}, \text{age}, \text{'Male'}, \dots) \wedge \text{T1}(\text{prtclID}) \wedge \text{age} \geq 40$

m2: $S3_SUBJECT(\text{subjectID}, \text{age}, \text{gender}, \dots) \rightarrow \text{Subject}(\text{subjectID}, \text{age}, \text{gender}, \dots)$

- System rewrites query Q to an executable query Q' (over source schemas)

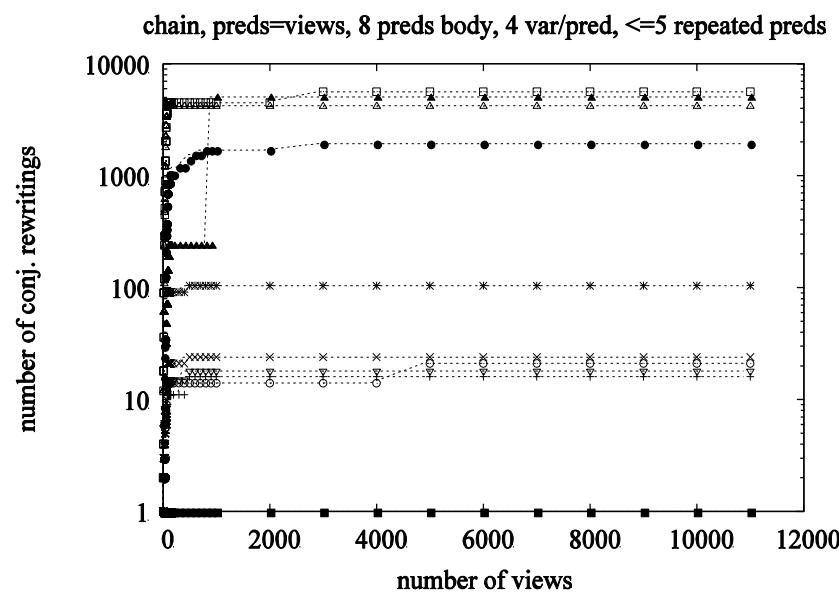
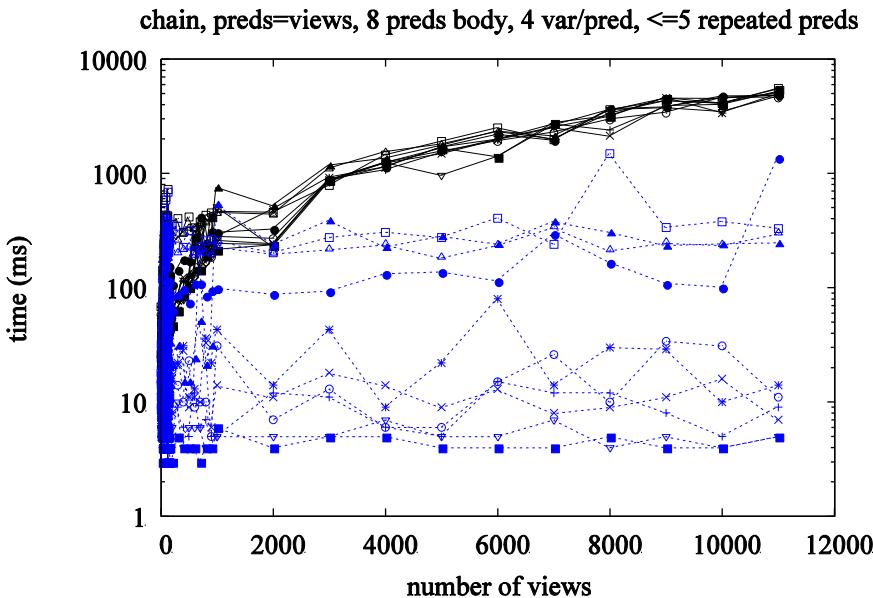
$Q'(\text{subjectID}, \text{age}, \text{'Bipolar'}, \text{eID}) \leftarrow$

$S1_EXPSEGMENT(\text{eID}, \text{projectID}, \text{subjectID}, \text{prtclID}, \text{date} \dots) \wedge$
 $S2_PROTOCOL(\text{prtclID}, \text{protocolDescription}, \dots) \wedge$
 $S3_SUBJECT(\text{subjectID}, \text{age}, \text{gender}, \dots) \wedge \text{age} > 50$

- Q' is translated to evaluation plan or ETL workflow (next slide)

Scalable Data Integration

Query rewriting is NP-hard, but
GQR [Konstantinidis & Ambite, SIGMOD 2011] can
rewrite user query over 11,000 source mappings
(i.e., 100 tables in 110 DBs) in under 1 second



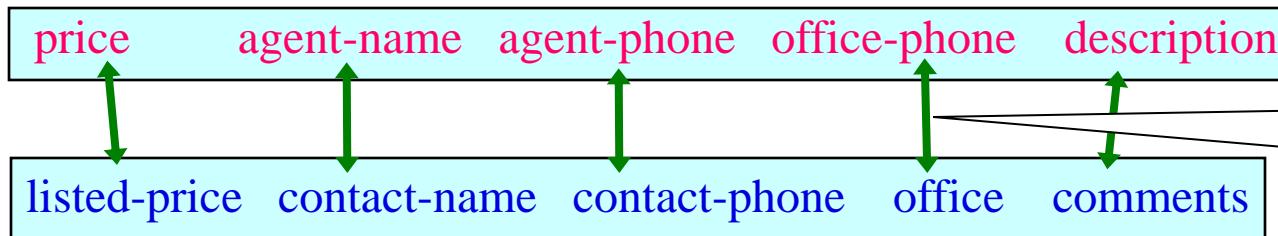
Schema Mapping

- Given two different sources with different schemas, how do we automatically align the information
- Research Topics
 - Automatic schema alignment based on structure and naming
 - Automatic alignment based on the source contents



Schema Mapping

Mediated schema



If "office" occurs in name
=> office-phone

Schema of realestate.com

realestate.com

listed-price	contact-name	contact-phone	office	comments
\$250K	James Smith	(305) 729 0831	(305) 616 1822	Fantastic house
\$320K	Mike Doan	(617) 253 1429	(617) 112 2315	Great location
.....

homes.com

sold-at	contact-agent	extra-info
\$350K	(206) 634 9435	Beautiful yard
\$230K	(617) 335 4243	Close to Seattle

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

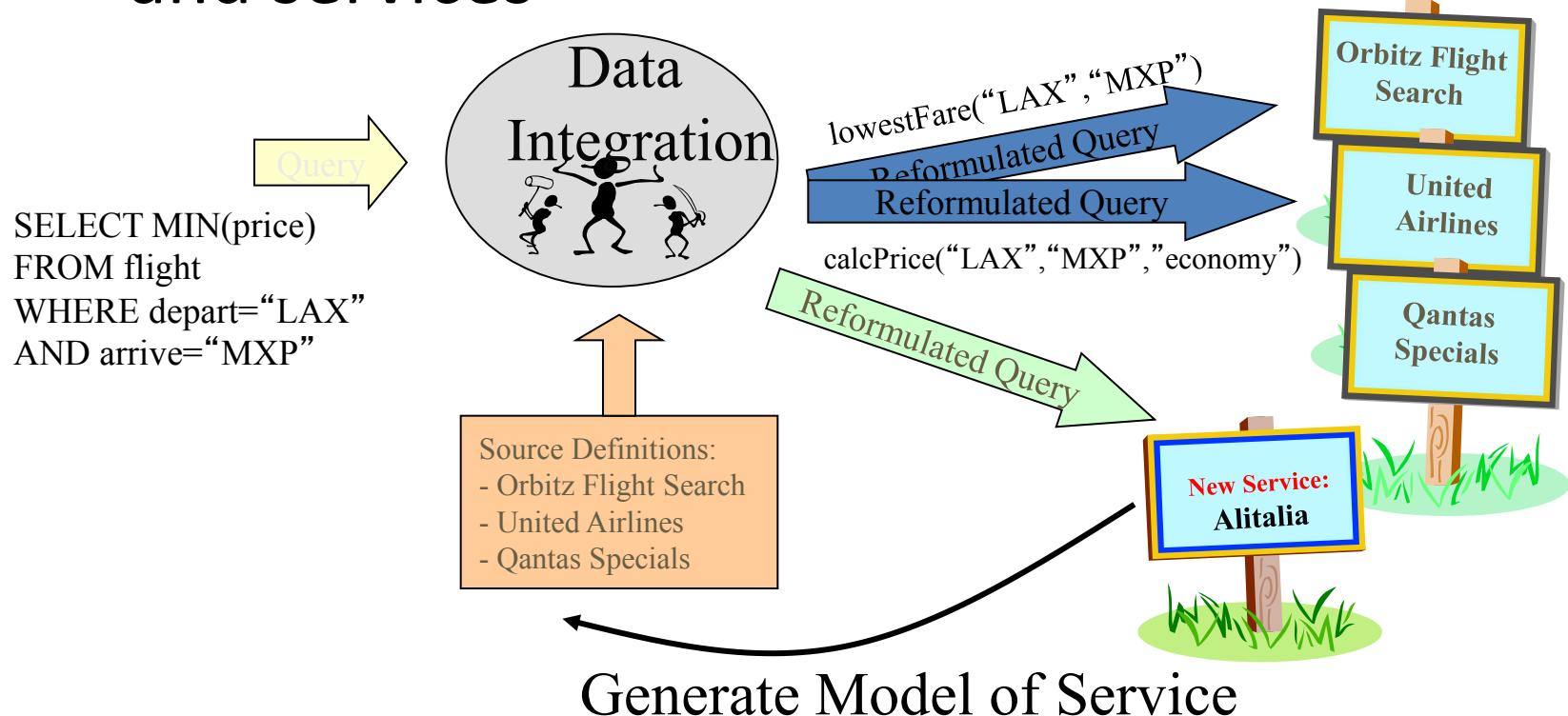
Source Modeling

- Semantic typing
- Source discovery
- Automatic source modeling
- Interactive source modeling

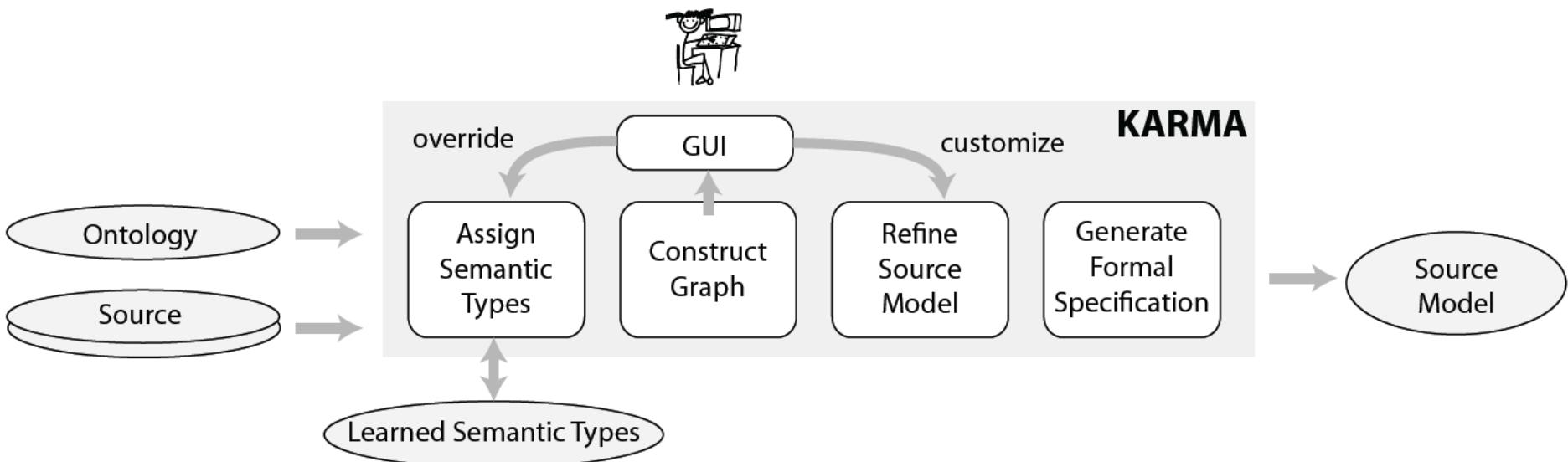


Automatic Source Modeling

- How to learn semantic descriptions of sources and services



Semi-Automatic Source Modeling



Relational database to RDF mapping

Anything to RDF mapping



Karma

A Data Integration Tool

Karma is available as open source (Apache 2 License)

[Download »](#)

Karma Demonstration: Integrating Bioinformatics Sources

The screenshot shows a semantic web interface for integrating bioinformatics sources. At the top, a navigation bar includes a back arrow, forward arrow, and a help icon. Below it is a hierarchical diagram showing relationships between Gene, Pathway, and Drug entities. A modal dialog box titled "Drug_Name" is open, displaying a table of drugs and their properties. The table has columns for Name, Gene_Accession_ID/Gene_Name, and pharmGKBId*. The "Drug_Name" column contains entries like Methotrexate Pathway, Thiopurine Pathway, Statin Pathway (PK), Phenylbutyrate Pathway (PK), etc. The "Gene_Accession_ID/Gene_Name" column lists genes such as ABCB1, ABCC4, ABCB1, CYP1A2, IL10RA, PARK2, AP2A1, CDC42, and ABCB1. The "pharmGKBId*" column lists IDs like PA267, PA397, PA267, PA27093, PA29779, PA32942, PA24852, PA26266, and PA267. To the right of the table is a "Semantic types:" section with a "Primary" checkbox next to "property::name of category:Drug". Other options include "property::pharmGKBId of category:Drug", "property::name of category:Disease", and "property::name of category:Pathway". There are also "Edit (CRF Suggested)" buttons for each. At the bottom of the dialog are "Add synonym semantic type", "Mark as key for the class.", and "Cancel/Submit" buttons.

Principal Investigators



Craig
Knoblock



Pedro
Szekely

<http://www.isi.edu/integration/karma/>



String Similarity: Why Strings Don't Match Perfectly?

typos "Joh" vs "John"

OCR errors "J0hn" vs "John"

formatting conventions "03/17" vs "March 17"

abbreviations "J. S. Sargent" vs "John Singer Sargent"

nick names "John" vs "Jock"

word order "Sargent, John S." vs "John S. Sargent"



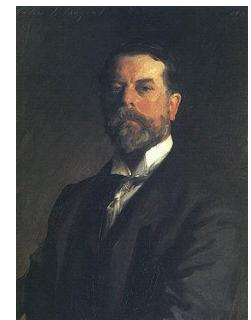
String Similarity Problem Definition

Given X and Y sets of strings

Find pairs (x, y)
such that both x and y
refer to the same real world entity

"John S. Sargent"

"John Singer Sargent"

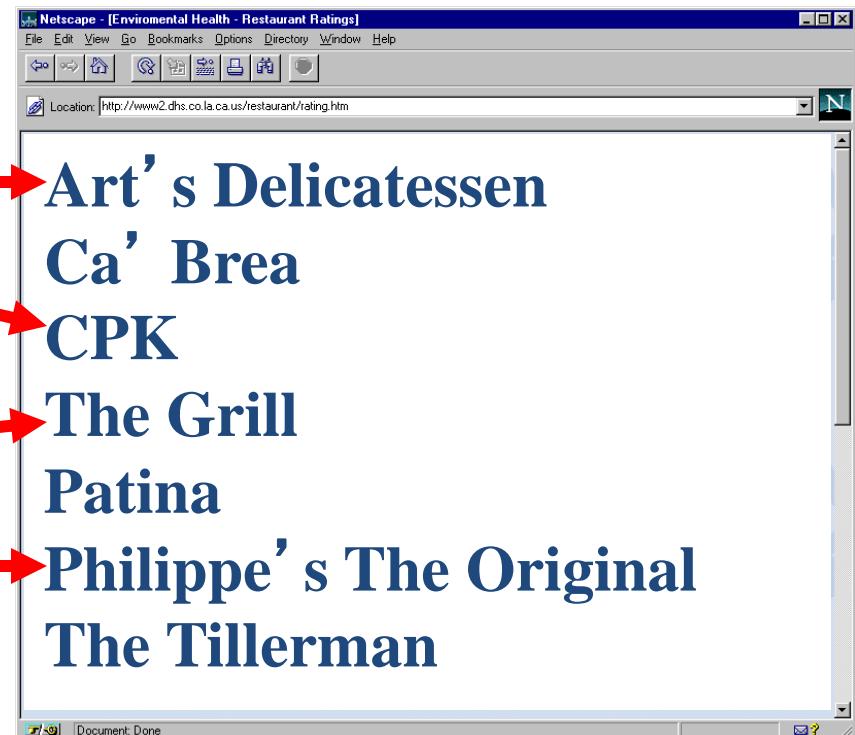


Record Linkage

Zagats Restaurant
Guide Source



Department of Health
Restaurant Source



How can the same objects be identified
when they are stored in inconsistent text formats?



Record Linkage

- Align entities across sources
- Research Topics:
 - Blocking
 - Matching individual attributes
 - Matching records
 - Matching entities

Silk - A Link Discovery Framework for the Web of Data



[Robert Isele](#) (Freie Universität Berlin)

[Anja Jentzsch](#) (Freie Universität Berlin)

[Chris Bizer](#) (Freie Universität Berlin)

[Julius Volz](#) (Google)



Mashup Construction

 pipes Home My Pipes Browse Discuss Documentation [Create a pipe](#) Sign in with your Yahoo! ID or [Join Now](#)

About Pipes

Pipes is a powerful composition tool to aggregate, manipulate, and mashup content from around the web.

Like Unix pipes, simple commands can be combined together to create output that meets your needs:

- combine many feeds into one, then sort, filter and translate it.
- geocode your favorite feeds and browse the items on an interactive map.

Featured Pipe: eBay Price Watch



This pipe is designed to use eBay's RSS API to find items within a certain price range.
Photo by Carlos

Yahoo! Query Language YQL Console Code Examples Documentation Blog Forum

What is YQL?

The Yahoo! Query Language is an expressive SQL-like language that lets you query, filter, and join data across Web services. With YQL, *apps run faster with fewer lines of code and a smaller network footprint*.

Yahoo! and other websites across the Internet make much of their structured data available to developers, primarily through Web services. To access and query these services, developers traditionally endure the pain of locating the right URLs and documentation to access and query each Web service.

With YQL, developers can access and shape data across the Internet through one simple language, eliminating the need to learn how to call different APIs.



Ready to get started?
Your use of YQL is subject to the [YQL Terms of Service](#)

[Try the console](#)

[Read the Documentation](#)
[Download the PDF Documentation](#)

DOCUMENTATION AND RELATED LINKS
[- YQL Documentation](#)



Information Extraction (IE)

“1988 Honda Accrd for sale! Only
80k miles, Runs Like New, V6,
2WD... \$2,500 obo. SUPER
DEAL.”



Information Extraction

- Finding structure in unstructured text
- Research Topics
 - Extraction using NLP techniques
 - Extraction with Conditional Random Fields
 - Exploiting reference sets for extraction



Ontology-based data access and integration

- Use ontology language as domain model
 - OWL2 profiles
- Answering queries under description logic constraints
 - Unions of conjunctive queries
 - Datalog

Ontology-based data access and integration: Example Rewriting

Ontology:

$\text{Professor} \sqsubseteq \exists \text{teaches},$
 $\exists \text{teaches}^- \sqsubseteq \text{Student}.$

Clauses:

$\text{teaches}(x, f(x)) \leftarrow \text{Professor}(x),$
 $\text{Student}(x) \leftarrow \text{teaches}(y, x).$

Query: $Q(x) \leftarrow \text{teaches}(x, y) \wedge \text{Student}(y).$

Rewriting:

$Q(x) \leftarrow \text{teaches}(x, y) \wedge \text{Student}(y).$
 $Q(x) \leftarrow \text{teaches}(x, y) \wedge \text{teaches}(z, y),$
 $Q(x) \leftarrow \text{Professor}(x).$

Main idea: compile the ontology inferences into the query

Wrapper Generation and Learning

- Turning online sources into structured information
- Research Topics
 - Wrapper Learning
 - Automatic Wrapper Generation
 - Wrapper Maintenance



Wrapper Generation and Learning

Yellow Pages

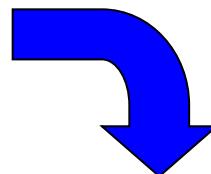
Home → Yellow Pages → Results

RESULTS Restaurants (1 - 1 of 1)

Casablanca Restaurant
220 Lincoln Boulevard, Venice, CA 90291
(310) 392-5751

Appears in the Category:
[Restaurants](#)

[Jump to Top](#)

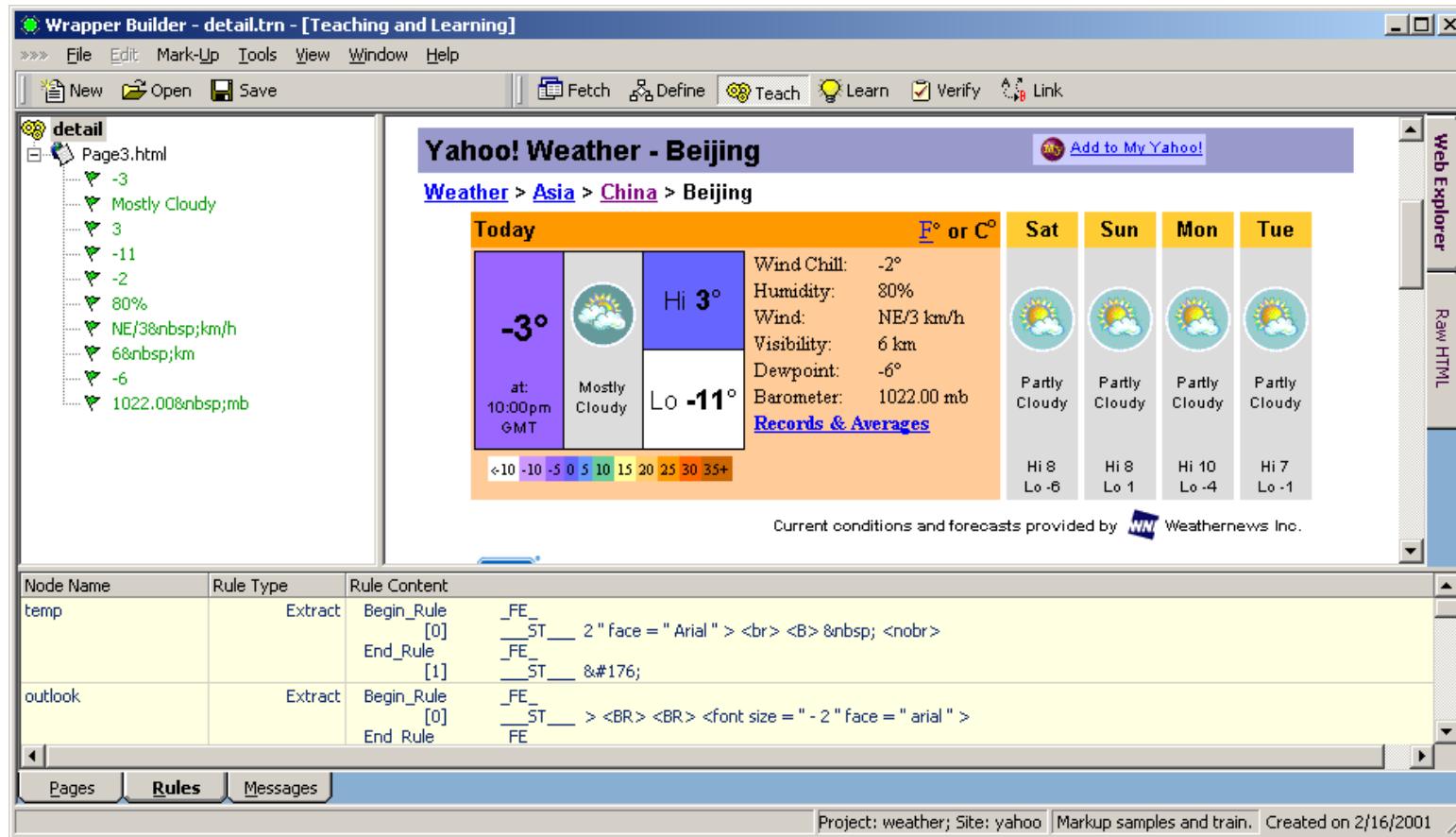


NAME	Casablanca Restaurant
STREET	220 Lincoln Boulevard
CITY	Venice
PHONE	(310) 392-5751



Wrapper generation and learning

- Wrapper automatically learned based on a few examples



→ Startup: Fetch Technologies (www.fetch.com)

Intellectual Property



Patents



Trademarks

A screenshot of the Java Code Expert software. The window has two main panes: "Input" on the left and "Output" on the right. The input pane contains Java code, and the output pane contains C/C++ code. The Java code includes imports for "java.util.List", "java.util.ArrayList", and "java.util.StringBuilder". The C/C++ code includes imports for "string.h", "vector", and "string". The code itself defines a class "Java2C" with methods for converting Java's List and ArrayList into C/C++ equivalents.

Copyrights



Trade Secrets



Intellectual Property

- What types of things can you patent
- When do you copyright something
- When would you file a trademark
- What is a trade secret
- Are we allowed to extract data from online sources
- When you do, who owns the data
- ..and so on...



Course Mechanics

Where to find Professor Ambite...

Research Assistant Professor, Computer Science
Project Leader, Information Sciences Institute

- Office hours:
 - Main campus: Outside ZHS 163 or THH 114
(Immediately after class)
 - Information Sciences Institute, Marina del Rey
310-448-8472 (by appointment)
- Email: ambite@isi.edu

Where to find Professor Knoblock...

- Research Professor
Computer Science and Spatial Sciences
- Office location:
 - Spatial Sciences AHF B55a
- Office hours:
 - Wed 10-11am in AHF B55a (Spatial Sciences)
 - By appointment at ISI: 310-448-8786
- Email: knoblock@isi.edu



Teaching Assistant

- TA: Bo Wu
 - Office Hours: Tuesday, 10am-12pm
- Location: plaza between RTH café and EEB
- Email: bowu@isi.edu



Course Materials and Web Pages

- Dropbox folder:
 - <https://www.dropbox.com/sh/5si94aql8inubb1/AABXH91b2iQHaPhidUo1Z7nXa?dl=0>
 - All readings, slides, homeworks, etc will be posted on the shared Dropbox folder
- Google discussion board
 - All questions should be posted
 - Expect response from TA in 24 hours, otherwise let Profs know
 - If you know the answer to a posted question, please try to provide helpful suggestions
 - But please don't post answers to homeworks!
- Blackboard – blackboard.usc.edu
 - Submit homeworks on Blackboard



Prerequisites & Recommended

- Prerequisites
 - CS561 -- Introduction to AI
- Recommended
 - CS585 – Database Systems
 - Some programming experience



Grading (CSCI 548)

- Homework: 50%
 - Must be turned in on time, but
 - You can turn in a homework assignment up to 1 week late with a 20% grade penalty
- Quizzes: 25%
 - One per week, at the beginning of the class
 - Questions based on the lectures, readings, and homeworks from previous week
 - No make ups
 - We will drop the lowest grade
- Final Exam: 25%
 - 3:30pm class: Friday, May 8, 2-4 p.m.
 - You must take the final for your class – no exceptions!



More on Grading

- This is a hard class, but you will learn a lot!
 - Principles and theory
 - Technical readings and lectures (quizzes, final exam)
 - Putting principles into practice
 - Homeworks
 - We do give B's, C's, and even D's and F's
- Grade distribution
 - Roughly half A's and B's
(consider a C a failing grade)

94 - 100 = A

90 - 93 = A-

87 - 89 = B+

84 - 86 = B

83 - 83 = B-

77 - 79 = C+

74 - 76 = C

70 - 73 = C-

67 - 69 = D+

64 - 66 = D

60 - 63 = D-

Below 60 is an F



Readings

- Posted on Dropbox each week
 - You can read it online or print them
- Read all required readings before class
 - You will get more out of the lectures
- Optional readings go in more depth
 - Read them to learn more



Slides

- Available online before the lecture
- Not intended as a replacement for the lecture
- You can bring them to class
- Final version of the slides available after class



Working Together

- Each person must do their own homework
 - We will check for overlap in homeworks
 - If we find any plagiarism, all parties lose credit so
 - Don't share your answers
 - Don't leave printouts in the trash with your answers
 - Don't give out your password
 - Don't copy others (they may have the wrong answer anyway!)
- You can ask the professors or TAs for help



Cheating

- Not tolerated!
- All infractions will be reported
- Examples:
 - Turning in someone else's homework
 - Doing the homework in collaboration with someone else and then turning in your own copy
 - Copying from someone else during a quiz or exam



We Follow USC Policies



UNIVERSITY OF
SOUTHERN CALIFORNIA



EXAMINATION BLUE BOOK

NAME *Joe College*
SUBJECT *Trojan Integrity Quiz*
INSTRUCTOR *Judicial Affairs and Community Standards*
EXAM SEAT NO. _____
SECTION _____
DATE _____
GRADE _____

TROJAN INTEGRITY

A Guide to Understanding and Avoiding Academic Dishonesty

Introduction

One key value at USC, as in all academic communities, is academic integrity: honesty in all academic endeavors. Those who fail to uphold these standards not only suffer severe grade consequences, but also cheat themselves and others out of learning, degrade the value of their degree, and diminish the prestige of a USC education.

What is Academic Dishonesty?

What constitutes academic dishonesty at the University of Southern California is spelled out in the student handbook, *SCampus*. It includes, but is not limited to: plagiarism, cheating on exams, unauthorized collaboration and falsifying academic records. Abbreviated definitions follow:

Plagiarism: Using someone else's work in any academic assignment without appropriate acknowledgment (such as paraphrasing another's ideas or copying text, phrases or ideas from a book, journal, electronic source or another person's paper, without acknowledgment).

Cheating on Exams: Unauthorized use of external assistance during an examination (such as using crib notes, talking with fellow students, or looking at another person's exam).

Unauthorized Collaboration: Preparing academic assignments with another person without faculty authorization (such as discussing or sharing work on homework or projects).

Falsifying Academic Records: Alteration or misrepresentation of official or unofficial records including academic transcripts, applications for admission, exam papers, registration materials, medical excuses or lab attendance forms.

What are the Consequences?

In addition to a grade penalty ranging from a "zero" on an assignment to an "F" in the course, the student may also face the following sanctions: dismissal from an academic unit, revocation of admission, suspension from the university, revocation of degree and expulsion from the university.

What is the Procedure?

If a student is accused of academic dishonesty, the student has an opportunity to meet with the faculty member to discuss the basis for the allegation. The faculty member may assess an academic penalty for the course and must report the action to the Office for Student Judicial Affairs and Community Standards, and he or she may recommend additional sanctions.

If the student denies the allegation, he or she has an opportunity for a review of the matter. Such a review may be conducted by an administrator or a panel. Refer to *SCampus* for the official statement of policies and procedures.

The decision from the review may be appealed to an appellate body. The decisions rendered by the appeals panel are final.

What are Your Responsibilities?

- Don't do it! Remember that a poor grade on an assignment or exam is better than failing the course and facing suspension or expulsion.
- Report cheating to the faculty.
- Protect your work from others, and do not take unfair advantage of other students' work.
- Prepare yourself. Learning to budget time to ensure optimal preparation for an exam or assignment is an absolutely essential tool to success at any university.
- Know exactly what constitutes academic dishonesty. Read *SCampus*, talk to your professors and TAs.
- Make sure you understand the specific standards for an assignment or class. If you don't know, ask your professor or TA.
- Don't sit next to friends during an exam. It may put you or them in a compromising position.
- Get help. Extensive campus resources including the Center for Academic Support, Student Judicial Affairs and Community Standards and The Writing Center are available, but you have to take the first step.
- Discourage your friends and classmates from committing acts of academic dishonesty by providing them with support, information and a good example: you!

Print it (from Blackboard), read it, sign it and return it to us

Waiting List

- Even if you are in the waiting list, come to the classes and do the quizzes/homeworks
- Likely there will be additional openings in the next couple of weeks

When the Course is Over

- Directed research (MS or Phd Students)
- M.S. Thesis
- Summer interns (MS or Phd)
- Research Assistantships (1-2 Phd Students)
 - We can also recommend you for positions in other groups
- Graders/Course Producers (MS)
- Recommendation letters (anyone that gets at least an A-)
- Positions at related companies
 - Companies are often looking for recommendations of students



Questions?