
DSE 203

DAY 1: INFORMATION HETEROGENEITY

Information Heterogeneity – a typology

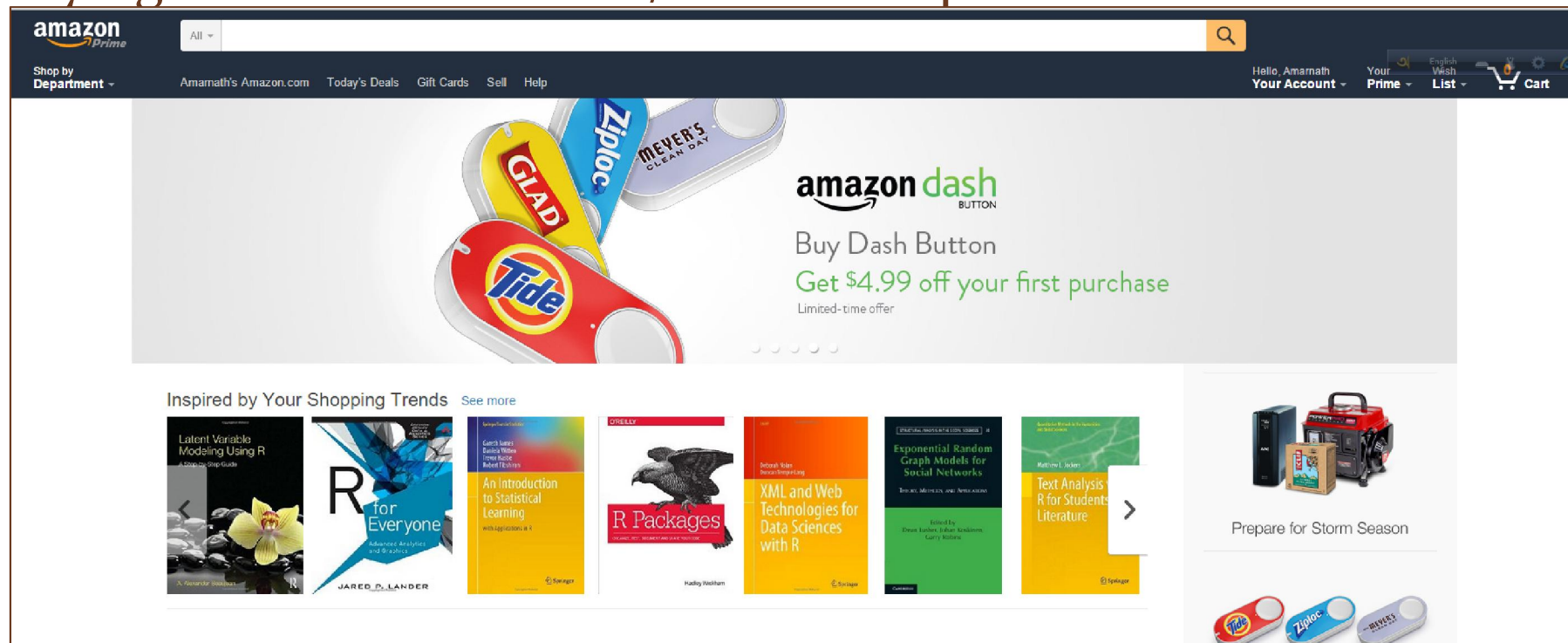
- Heterogeneity
 - System Level
 - Hardware/Software Level
 - Interface Level
 - Structural
 - Data Model Related
 - Schema Related
 - Semantic
 - Identity Conflicts
 - Naming Conflicts
 - Value Conflicts

Heterogeneity in Hardware/System Software

- Hardware differences
 - Mobile phone vs. laptop vs. servers
 - Not every platform can do every query equally effectively
 - “Find a McDonalds near me” while driving a car
 - “Outer Join policy holders with bank customers”
- Different protocols, binary file formats, ...
 - Order information stored in text files: line ending differs between Mac/Window/Linux, character encoding
- Different access control mechanism
 - FTP-access to files: public, ssh authentication, ..

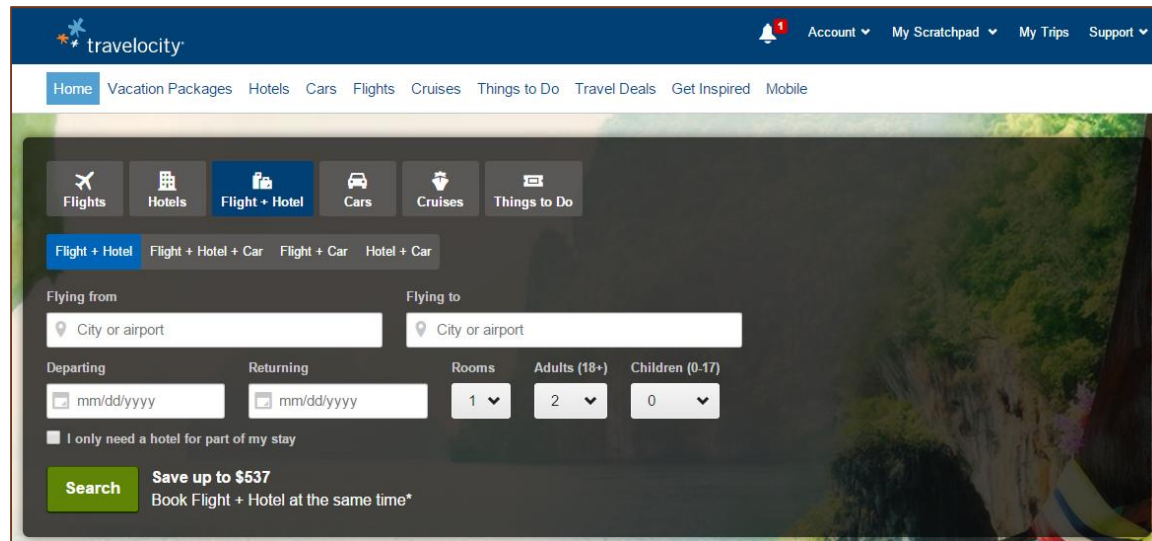
Heterogeneity in Query Interfaces

- Two systems might use the same data model and yet may significantly differ in the way they allow access to their data
- Querying Amazon – what can/can't we express?



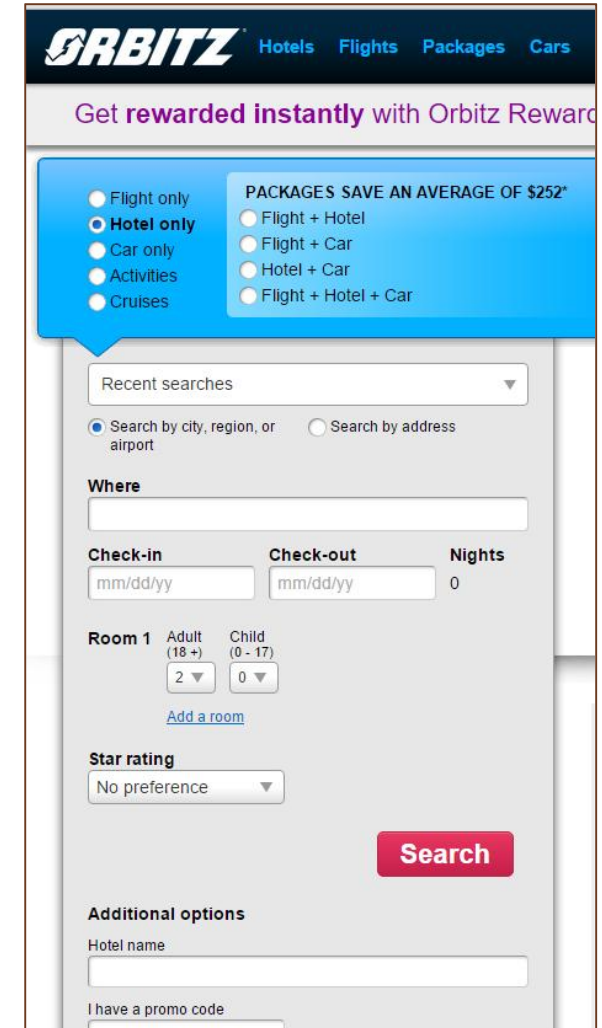
Heterogeneity in Query Interfaces

- Travelocity and Orbitz
 - Is there a difference in their query capabilities?



The screenshot shows the Travelocity homepage with a navigation bar at the top containing links for Home, Vacation Packages, Hotels, Cars, Flights, Cruises, Things to Do, Travel Deals, Get Inspired, and Mobile. Below the navigation bar is a search area with tabs for Flights, Hotels, Flight + Hotel (selected), Cars, Cruises, and Things to Do. Under the Flight + Hotel tab, there are sub-tabs for Flight + Hotel, Flight + Hotel + Car, Flight + Car, and Hotel + Car. The main search form includes fields for Flying from (City or airport), Flying to (City or airport), Departing (mm/dd/yyyy), Returning (mm/dd/yyyy), Rooms (1), Adults (18+) (2), and Children (0-17) (0). There is a checkbox for "I only need a hotel for part of my stay" and a green Search button. A promotional message says "Save up to \$537 Book Flight + Hotel at the same time*".

- Querying with web forms – what are the benefits and limitations?
 - Relations with binding patterns



The screenshot shows the Orbitz homepage with a navigation bar at the top containing links for Hotels, Flights, Packages, and Cars. Below the navigation bar is a search area with a blue banner that says "Get rewarded instantly with Orbitz Rewards". The banner includes a list of search options: Flight only, Hotel only (selected), Car only, Activities, Cruises, and a section for Packages that says "PACKAGES SAVE AN AVERAGE OF \$252*" with options for Flight + Hotel, Flight + Car, Hotel + Car, and Flight + Hotel + Car. Below the banner is a search form with a Recent searches dropdown, a radio button for Search by city, region, or airport (selected), and a radio button for Search by address. The form includes fields for Where, Check-in (mm/dd/yy), Check-out (mm/dd/yy), and Nights (0). There is a section for Room 1 with Adult (18+) (2) and Child (0-17) (0) dropdowns, and a link to Add a room. There is a Star rating dropdown set to No preference. A red Search button is at the bottom right. There is also a section for Additional options with a Hotel name field and a field for I have a promo code.

Heterogeneity in Query Interfaces

Advanced Search

Find pages with...

all these words:

this exact word or phrase:

any of these words:

none of these words:

numbers ranging from:

Then narrow your results by...

language:

region:

last update:

site or domain:

terms appearing:

SafeSearch:

file type:

usage rights:

Search Results

chinese economy volatility US OR USA "stock market"

[Web](#) [News](#) [Videos](#) [Images](#) [Shopping](#) [More](#) [Search tools](#)


Search English pages Past month Sorted by relevance All results Clear

Extreme volatility rocks China stocks - Sep. 1, 2015
[money.cnn.com/2015/09/.../china-stocks-asia-world-markets...](#) CNNMoney
Sep 1, 2015 - Turbulence in Asia comes after a very rough Tuesday for U.S. stocks.... Analysts are concerned that **China's economy** is slowing faster than anticipated.

Can the global gloom sink the U.S. economy? - Aug. 29, 2015
[money.cnn.com/.../stocks-market-lookahead-us-economy/](#) CNNMoney
Aug 29, 2015 - That immediately raised questions about whether **China's economy** is really ... Amid the **volatility**, New York Fed President William Dudley said Wednesday that a ... In short, the **stock market** turmoil does impact the Fed's decision on a rate hike.

Ask an Economist: What's Going on With the Stock Market ...
[www.usnews.com/.../ask-an-economist-whats...](#) U.S. News & World Report
Sep 8, 2015 - The **U.S. stock market** has been all over the place in the last few weeks, with all three ... Given where the **U.S. economy** is right now, how alarming are these fluctuations? ... What does **Chinese volatility** mean to the every day U.S. consumer?

In the news

 **China's Xi seeks to reassure on cybercrime, stock market**
Global Risk Insights - 1 day ago
Xi Jinping's first official visit to the **U.S** has seen an important speech which ... The world was shocked by the sudden **volatility** in the **Chinese stock market** and Beijing's responses to said crisis.

The Fed, NOT China, is at fault for market volatility
Asia Times - 2 days ago

- Querying search engines
 - Keywords + partly structured search
- What can we express with search engines that we cannot with standard databases?

Heterogeneity in Query Interfaces

All services are subject to change at any time.

The services include the ability to:

Retrieve a federation summary, e.g., http://nif-services.neuinfo.org/servicesv1/v1/summary?q=*

Retrieve data records from a NIF federation source for a search,
e.g., <http://nif-services.neuinfo.org/servicesv1/v1/federation/data/nif-0000-00007-1?q=purkinje>

Retrieve registry data records from NIF, e.g., http://nif-services.neuinfo.org/servicesv1/v1/federation/data/nlx_144509-1?q=miame

Retrieve a complete search summary, e.g., <http://nif-services.neuinfo.org/servicesv1/v1/federation/search?q=cortex>

Retrieve NIF auto-complete suggestions, e.g., <http://nif-services.neuinfo.org/servicesv1/v1/vocabulary?prefix=hippo>

Use the NIF annotator for arbitrary text, e.g., <http://nif-services.neuinfo.org/servicesv1/v1/annotate?content=The%20cerebellum%20is%20a%20wonderful%20thing>

http://www.neuinfo.org/developers/nif_web_services.shtml

- Web services – function calls made over the web
- What schema are we querying against?
 - Every web service call has a response schema
 - Can be thought of as a **view** over an unknown relation

Heterogeneity in Query Interfaces

- Interface Heterogeneity – Expressiveness
 - Keyword-search vs. query language
- Query Language
 - Predicates: equality (=), inequality (<, !=)
 - Logical connectives: conjunctive (AND)
 - disjunctive (OR), negation
 - Complex operations: aggregation, quantification
 - Limitations: user's knowledge of the schema, predicates, fixed queries with parameters, ...

Heterogeneity in Query Interfaces

- Impacts on the integration system
- Unknown domain semantics
- Bounded Parameters
 - Not all attributes are available for query
 - The effective domain of an attribute can be restricted
 - Not all operations are available for an attribute
- Consequence
 - A reasonable query may have null result
 - An integration system may have to formulate multiple queries
 - “Find all children’s books for \$10 or less”
 - An integration system has to evaluate part of the query
 - “What’s the name and price of Avi Silberschatz’s latest book on Operating Systems?”


Structural Heterogeneity

- Data model
 - Different structure
 - Different semantics and expressiveness
- Schema
 - Integrity constraints, keys
 - Schema elements
 - Differently modeled attribute and relations
- Structure
 - Flat relational vs. nested relational
 - Deep vs. shallow trees in XML

Relational Conflicts

- Naming and Schema conflicts
 - Person(Id, name, gender, birthday)
 - Person(Id, firstname, lastname, male, female)
 - Manager(Id, name, gender, age)

what kind of
conflict is this?



- Conflicting integrity constraints
 - Person(Id, name, gender, birthday) – gender is not null, today().year-birthday.year > 18
 - Person(Id, name, gender='female', birthday)

Relational Conflicts

- Attribute conflicts
 - Person(SSN:integer, name:varchar(128), gender:string, birthday:datetime)
 - Person(SSN:varchar(9), name:varchar(64), gender:{'M', 'F', 'unwilling to reveal', default='unknown'}, birthdate:integer) – counting in days from 1/1/1900.
 - Person(SSN:string, name:string, gender:{0, 1,-99}, DOB:string(format:'dd-mm-yy'))

Structural Mismatch – relations, attributes, values

LabResults(orderID, date, PatientID, Test1, Test2, Test3,....)

LabResults(orderID, date, PatientID, TestID, TestValue)

Test1Results(orderID, date, PatientID, TestValue), Test2Results(orderID, date, PatientID, TestValue), ...

LabResults(orderID, date, PatientID, Tests(TestID, TestValue))

- User's query
 - Find the value for Test#3, Test#4 for PatientID='12345' for date='09/03/2015'
 - Select Tests from LabResults
where PatientID='12345' and date='09/03/2015' and
Tests.TestID = '3' or Tests.TestID = '4'
- Which model is the target schema using?

Handling Schema Heterogeneity

- The integration system needs to provide
 - Unified access to multiple schemas
 - An integrated schema over existing schema
- Schema level actions
 - Schema mapping, model management operators, schema languages
- Data level actions
 - Data transformation (ETL), data exchange, warehousing
- These are topics we will cover over this term

Semantic Conflicts

- The identity problem
 - Also called record linking, deduplication, entity resolution

Hospital A's record:

PID	SSN	FName	MI	LName	DOB	Address	Allergies
15883	555-43-2991	Florence	E	Schwartz	06/21/67	6345 Tony Drive, San Diego, CA 92127	Peanut, Cat fur, Pine pollen

Hospital B's record:

PID	SSN	FName	MI	LName	DOB	Address	Allergies
231834	653-86-9950	Flora	E	Schwartz-Jones	06/21/67	12290 Carmel Pointe, San Diego, CA 92130	Peanut, Cat fur, Pine pollen
Do these records belong to the same person?							

Value Conflicts – a kind of Semantic Conflict

- Non-ontological Conflicts
 - Objects representing the same entities have conflicting values for semantically equivalent attributes
 - Is the name of the company Google or Alphabet? IBM or International Business Machines?
 - In many cases other attributes like Date should be used to resolve value conflicts
 - Measurement related values may conflict due to choice of (often unmentioned) units
- There must be a way to identify that these objects are represent the same entity first!
- Resolving such conflicts require Data Fusion
 - Pick value from conflicting values
 - Numerical methods: e.g., average
 - Preferred value
 - Coarsification of granularity – e.g., latitude longitude

Ontological Conflicts – a kind of Semantic Conflict

- Same attribute, different domains
 - Human vs. homo sapiens
- Same attribute, different levels of granularity
 - Subclass-based granularity
 - Rodents vs. mice
 - Partonomy-based granularity
 - Distal phalanges vs. hand
- Different attributes, different levels granularity
 - Location
 - Named location vs. latitude-longitude

Ontological Conflicts – a kind of Semantic Conflict

- Implicit vs. Explicit constraints

PID	Image Type	Image SubType	View	Region Imaged	Visible Organs	Findings
15883	T1	Spin Echo	Sagittal	Lumber Spine	thoracic spine, spinal cord and sacrum	No visible abnormality

PID	Image Type	Image SubType	View	Region Imaged	Visible Organs	Findings
15883	MRI	Spin Echo	Sagittal	Lumber Spine	thoracic spine, spinal cord and sacrum	No visible abnormality

- Find all MRI images where the region viewed is Lumber Spine

Ontological Conflicts – a kind of Semantic Conflict

Ontological concepts

- Relationships between concepts
 - $\mathbf{A = B}$ - Equivalence
 - $\mathbf{A \subseteq B}$ - Inclusion
 - $\mathbf{A \cap B}$ - Overlap
 - $\mathbf{A \neq B}$ - Disjunction

Handling Ontological Conflicts

- The integration system needs to provide
 - A way to specify which schema elements conform to which ontologies
 - A way to specify ontological relationships as part of the query
 - A way to perform query evaluation using ontological relationships
- Schema level actions
 - Schema mapping to ontologies
- Data level actions
 - Data mappings to ontologies