DS105 : DATA INTEGRATION

Outline

- This week
  - Data Integration

- Next week
  - Start data visualization and ML topics
  - Visualization techniques (like TSNE)
  - Linear and Logistic regression
  - Clustering
  - Random Forest
  - Ensemble Methods

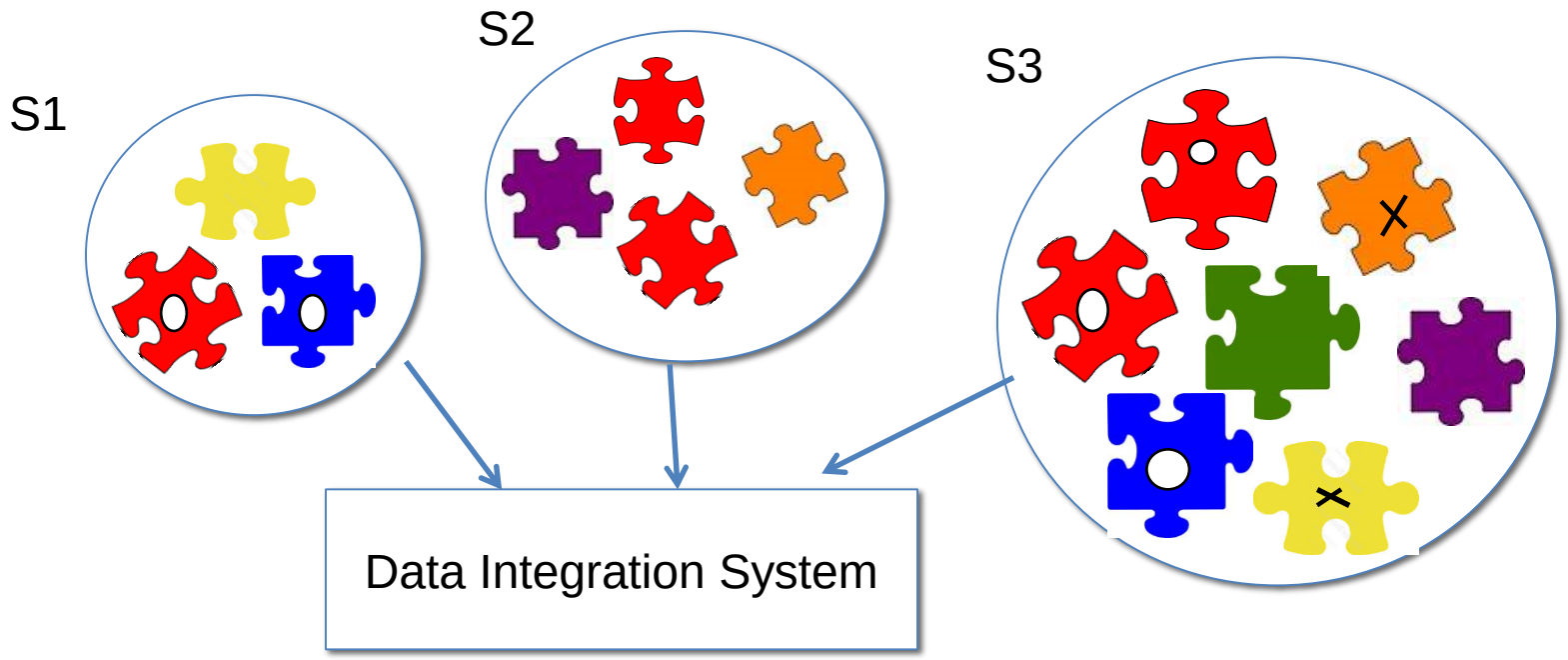
Reading Assignment

- Papers
  - Data Fusion: Resolving Conflicts from Multiple Sources by Luna Dong et. al. [PDF]

- Data Preprocessing / Cleaning
  - Chapter 2 from Data Mining : Concepts and Techniques [PDF]

What is “Data Integration”?

- Data integration is much like a jigsaw puzzle
  - There are many sources that have relevant pieces, but no one source provides a complete picture.
  - A data integration system gathers relevant pieces from various sources to provide a comprehensive view of the data.



What is “Data Integration”? (Cont.)

- Combining information from multiple autonomous and heterogeneous sources with the goal of proving a uniform, complete and consistent interface of answering queries over the integrated sources.

- Uniform access
  - Same query posed once to all sources
- Multiple, autonomous and heterogeneous sources
  - Multiple – usually hundreds to thousands of sources
  - Heterogeneous – data source ‘schema’ are different
    - Schema refers to how a data source represents its data internally
  - Autonomous – data sources don’t report to the data integration system when data or schema changes
- Complete
  - Containing all relevant answers (records) from the various sources
  - For a given record, having all relevant attributes
- Consistent
  - Duplicate records are removed
  - Invalid or conflicting data is resolved

Why is “Data Integration” a hard problem?

- When considering data integration, its all about the Vs
  - Size:
    - Large **volume** of data
      - Each data source may be large, but there are many sources for a single domain.
    - Collected and analyzed at high **velocity**
      - Data is generated at a high rate.
  - Complexity:
    - Huge **variety** of data
      - Sources (even in the same domain, have heterogeneous schemas.
    - Questionable **veracity**
      - Some sources may provide inaccurate data (or data that is not fresh).

Applications of Data Integration

- Many Applications
  - **WWW**
    - Comparison shopping (GoogleFlights, Kayak, etc.)
    - News, or business aggregators
    - Business-to-business, electronic marketplaces
  - **Science informatics**
    - Integrating genomic data, geographic data, archaeological data, astro-physical data, etc.
  - **Enterprise data integration**
    - An average company has 49 different databases and spends 35-50% of its IT dollars on integration efforts
  - **Knowledge Databases**
    - YAGO – a knowledge database that integrates wikipedia and other sources
    - DBpedia

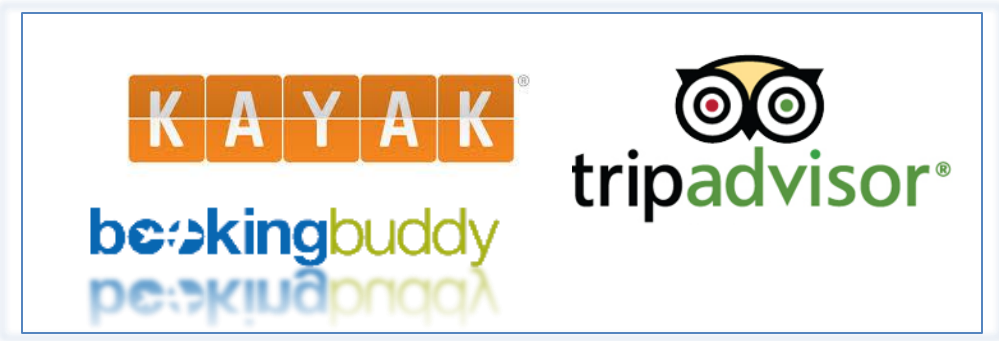
Application Area : The Web

Business Listings

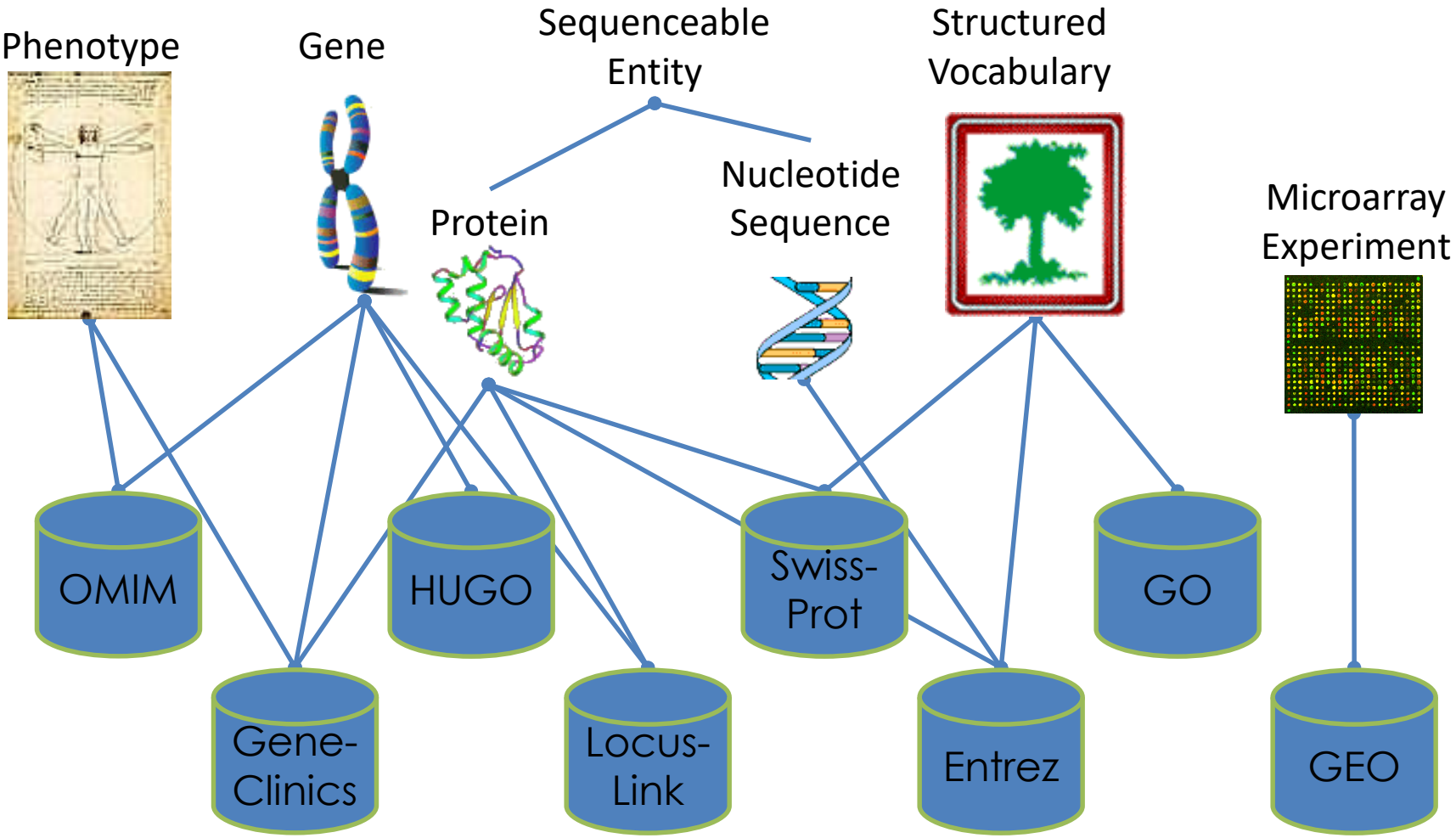
News Aggregators

Legal article/patents Aggregators

Travel (Flight/Hotel) Aggregators



Application Area : Science



Hundreds of biomedical data sources available

Application Area: Knowledge Bases

- Building Web-Scale Knowledge bases

Google Knowledge Graph

ProBase

MSR knowledge base

A Little Knowledge Goes a Long Way.

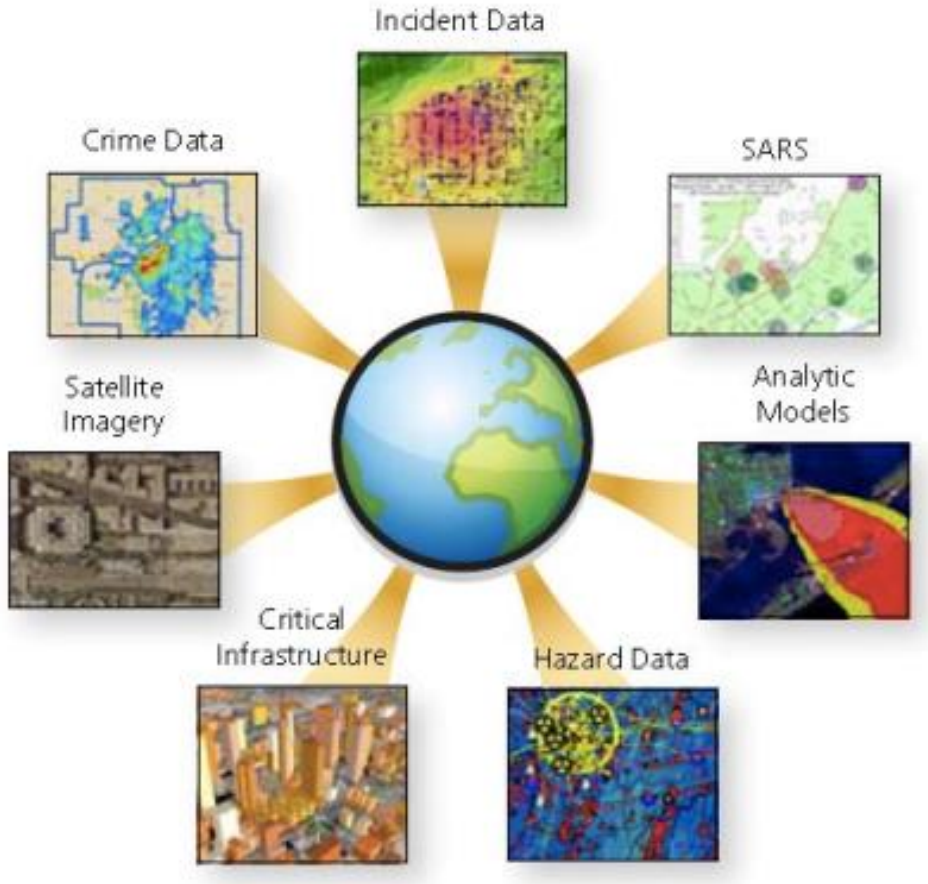
NELL

yago

select knowledge

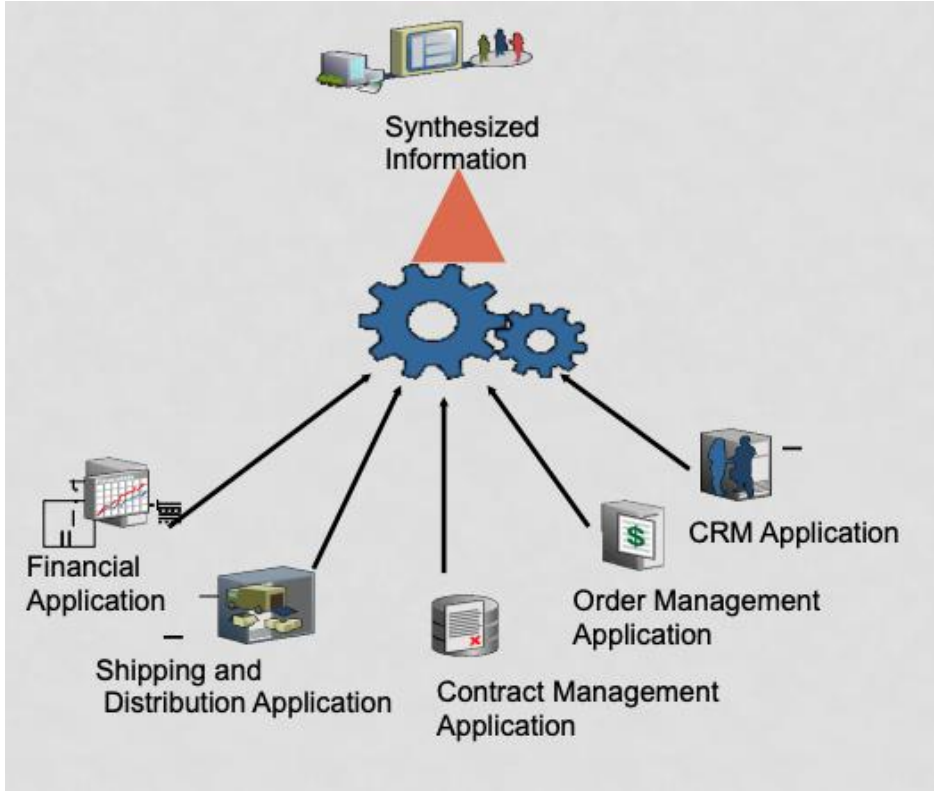
Application Area: GIS

- Geo-spatial data fusion



Data Integration is standard

- Data integration is also valid within a single organization.
  - Integrating data from different departments or sectors
  - Usually referred to as ETA (Extract – Transform – Load)



HETEROGENEITY PROBLEMS

- The main problem is the heterogeneity among the data sources

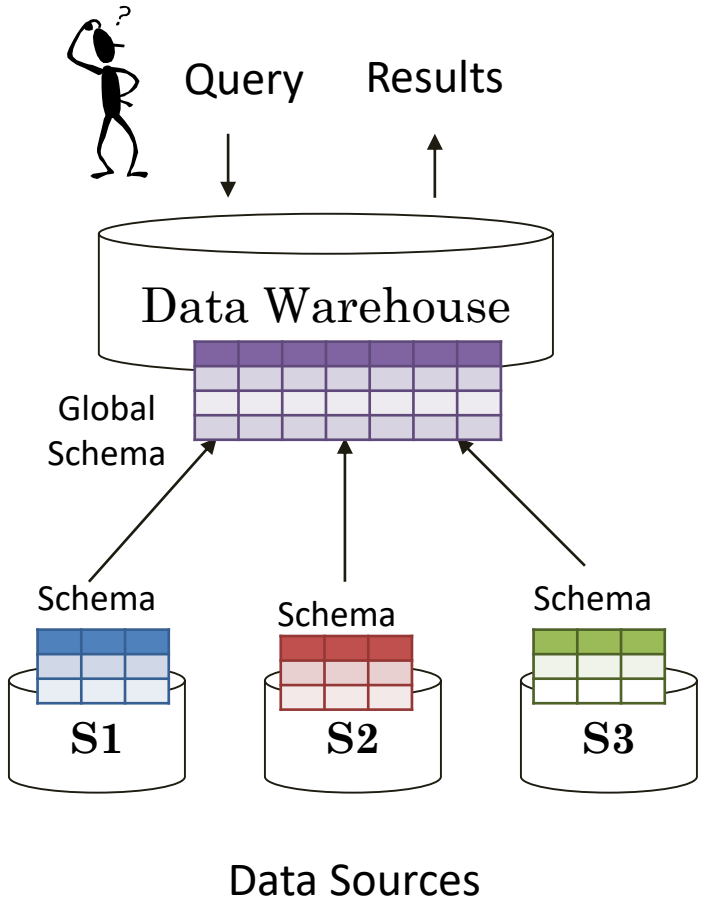
- Source Type Heterogeneity
  - Systems storing the data can be different

Approaches to Data Integration

- Data Warehouses
  - Usually utilized for static or slowly changing data
- Virtual Data Integration
  - Also called ‘Mediated Schema Integration’
  - Usually used for web data sources which publish new content at a high rate

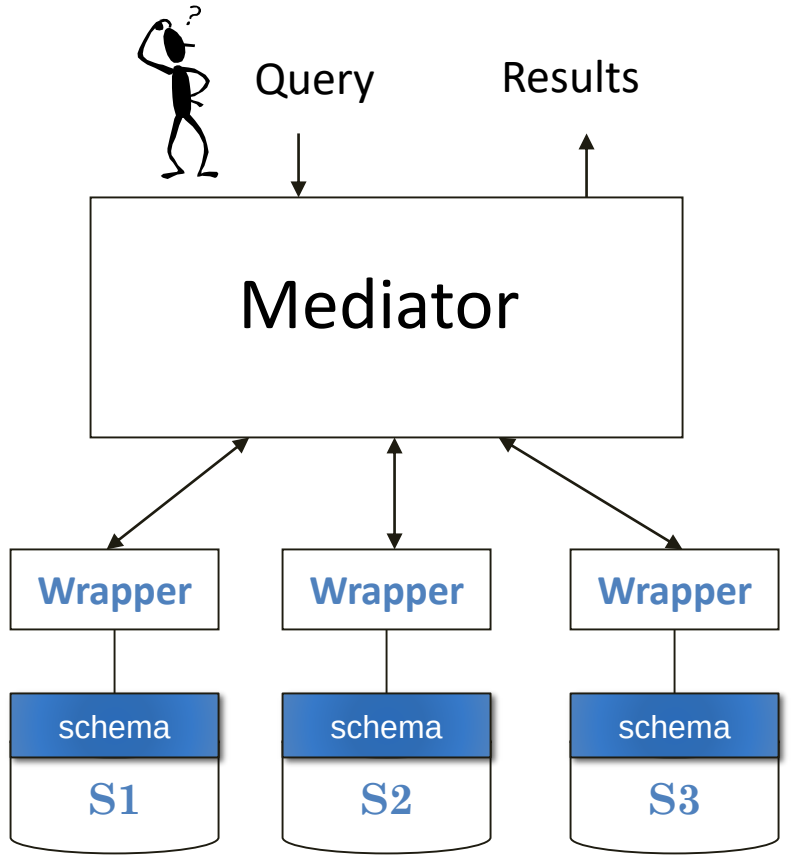
Traditional Data Integration Solution

- Data Warehouse
  - The warehouse system extracts, transforms and loads data from heterogeneous sources into a single data source so data becomes easily queryable.
- Problems
  - Data has to be cleaned due to different formats
  - Requires a central storage system for all data
  - Data needs to be updated periodically to maintain ‘fresh’ data.
    - Data sources are autonomous, hence, content can change without notice.
    - Expensive because of the large quantities of data and data cleaning costs.



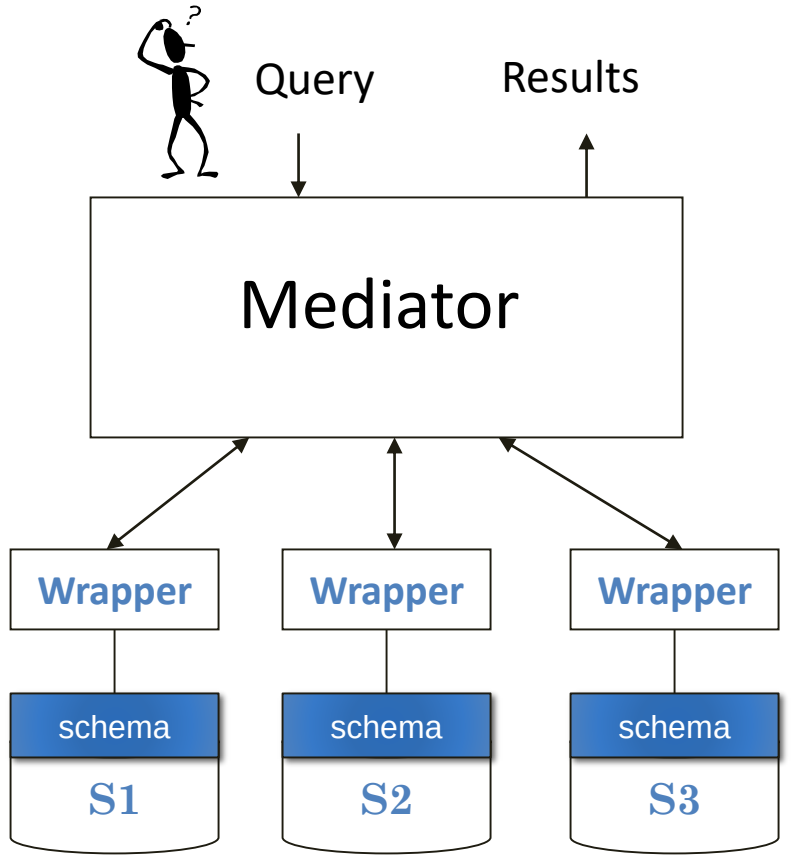
Data Integration using Mediated Schema

- Mediated Schema Approach
  - Mediator is a virtual view over the data (it does not store any data)
    - Data is stored only at the sources
  - Mediator has a virtual schema that combines all schemas from the sources
  - The mapping takes place at query time
    - This is unlike warehousing where mapping takes place at upload time



Data Integration using Mediated Schema

- Query is mapped to multiple other queries
- Each query (or set of queries) are sent to the sources
- Sources evaluate the queries and return the results
- Results are merged (combined) together and passed to the end-user

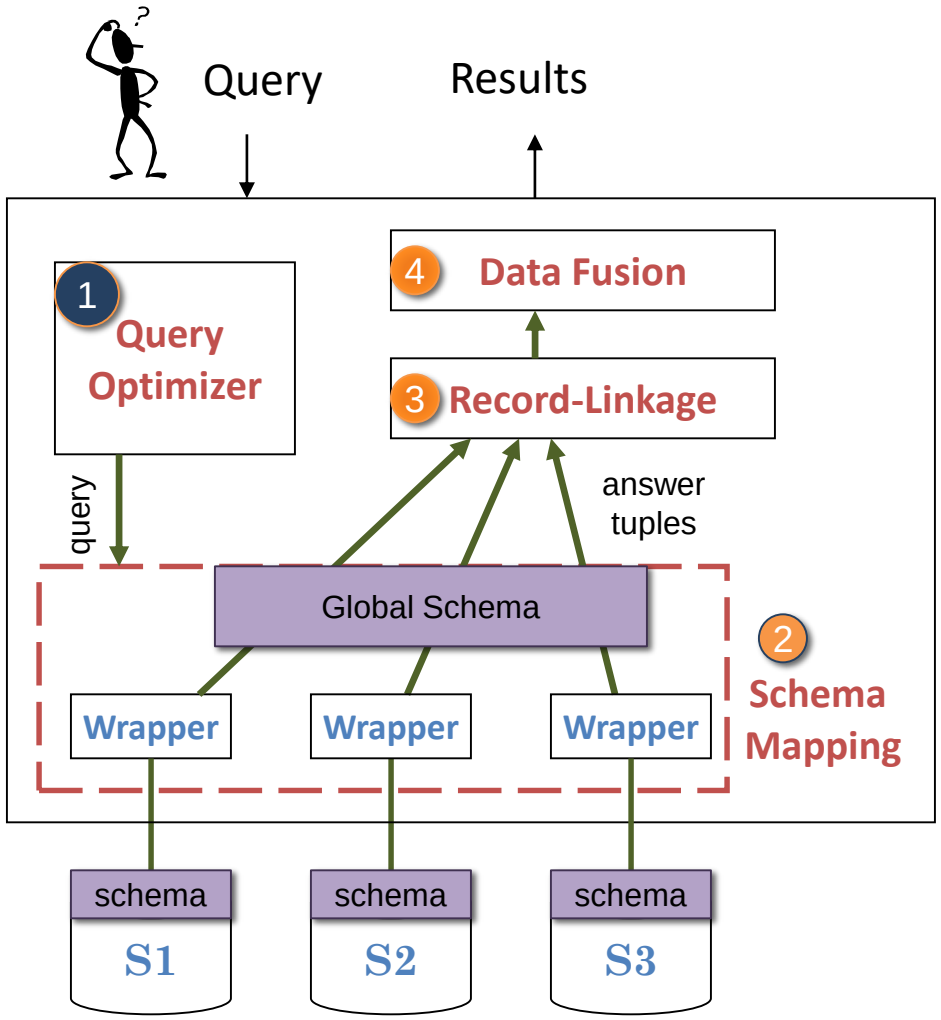


Problems in Data Integration

- Data integration (whether building a data warehouse or mediated schema) usually involves the following phases.
  - Source Selection
  - Schema Mapping
  - Records Linkage
  - Data Fusion

Source Selection Phase

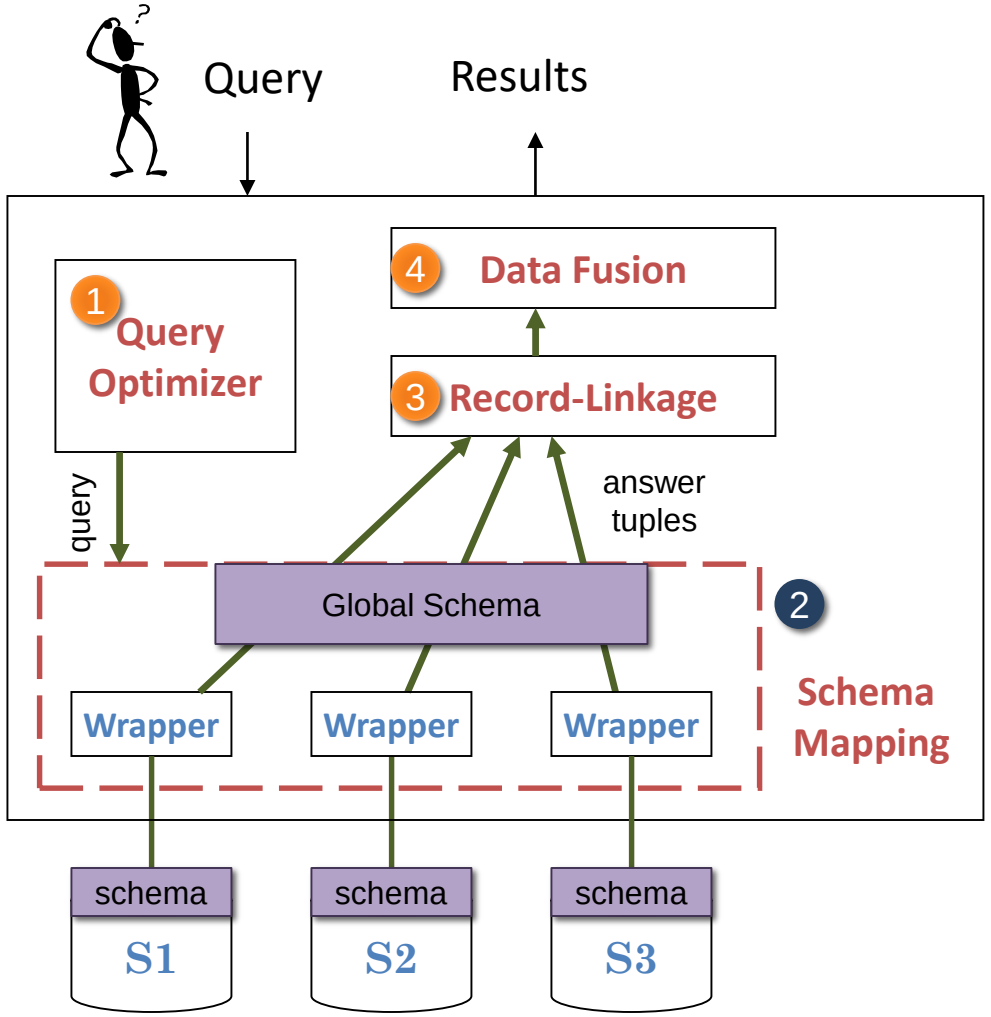
- Source Selection (Query Optimizer)
  - Determine which sources are most relevant to the given query.
  - Determine the order of how sources should be accessed.
  - For this purpose, the query optimizer needs to access statistics about the coverage of the sources, as well as the overlap between sources.



Query Optimization

- Select relevant sources for the given query.
- Select order in which to query sources.
- Consider quality of data provided sources (knowledge acquired from previous queries)
  - Coverage (# of answers)
  - Completeness (# of attributes)
  - Accuracy
  - Freshness

Schema Mapping Phase



- Schema Mapping
  - Since each source is independent and autonomous, it has its own schema.
  - **Objective:** Generate a global virtual schema, and create a mapping between the global schema and the individual source schemas.

Heterogeneity Problems

- **Schema Heterogeneity**
  - The structure of the tables storing the data can be different (even if storing the same data)
- **Data Type Heterogeneity**
  - Storing the same data (and values) but with different data types
  - E.g., Storing the phone number as String or as Number
  - E.g., Storing the name as fixed length or variable length
- **Value Heterogeneity**
  - Same logical values stored in different ways
  - E.g., ‘Prof’, ‘Prof.’, ‘Professor’
  - E.g., ‘Right’, ‘R’, ‘1’ …… ‘Left’, ‘L’, ‘-1’

Heterogeneity Problems (Cont.)

- Semantic Heterogeneity
  - Same values in different sources can mean different things
  - E.g., Column ‘Title’ in one database means ‘Job Title’ while in another database it means ‘Person Title’

Schema Matching Example

Need to resolve differences in schema and data representation.

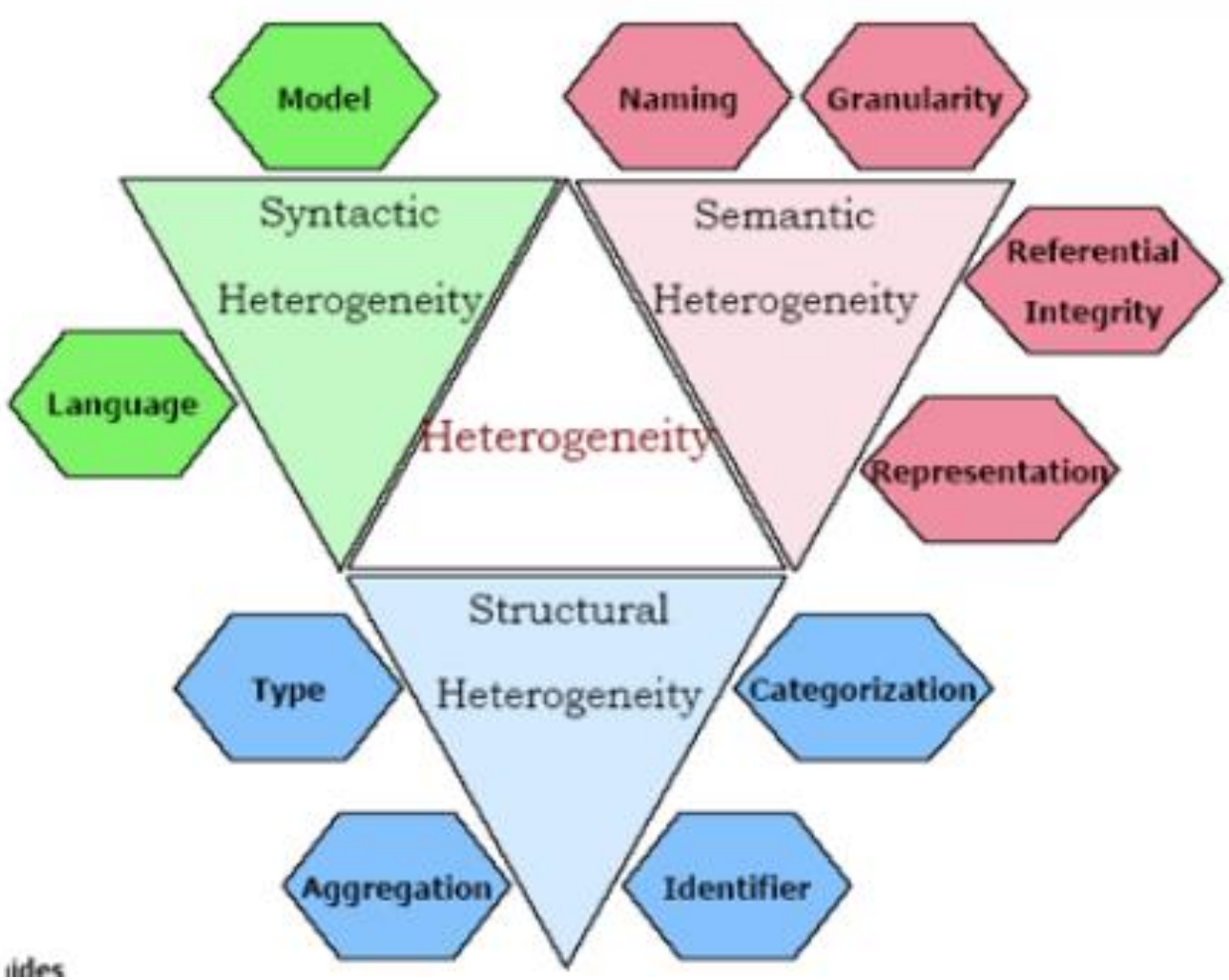
S1

Business Name	Phone	Specialty	Address	Website
Thai Boat	(805) 594-1638	Asian	3212 Broad St San Luis Obispo, CA	thaiboatslo.com
Jaffa Cafe	(805) 543-2400	Mediterranean	1308 Monterey St San Luis Obispo, CA	jaffacafe.us
Taste	(805) 541-5860	American, Burgers	2900 Broad St San Luis Obispo, CA	taste2900.com

S2

Listing Name	Phone	Cuisine	Street Address	City	State
Thai Boat	805-594-1638	Thai	3212 Broad St #100	San Luis Obispo	CA
Jaffa Cafe	805-543-2449	Lebanese	1308 Monterey St	San Luis Obispo	CA
Taste	805-541-5860	American	2900 Broad St	San Luis Obispo	CA
Oasis	805-543-1155	Greek	675 Higuera St	San Luis Obispo	CA

Challenges in Schema Matching



Schema Mapping Approaches (Cont.)

- Current Approaches
  - Manual or Semi-Supervised Schema Matching

- Schema Matching Approaches
  - Linguistic approach
    - Find semantically similar schema elements
    - Ex. S1.make = S2.brand
    - How is similarity defined?
      - Equality of names
      - Equality of names after stemming, deals with prefixes/suffixes
      - Equality of synonyms
      - Similarity of names based on common substring (edit-distance)
      - User provided name matches

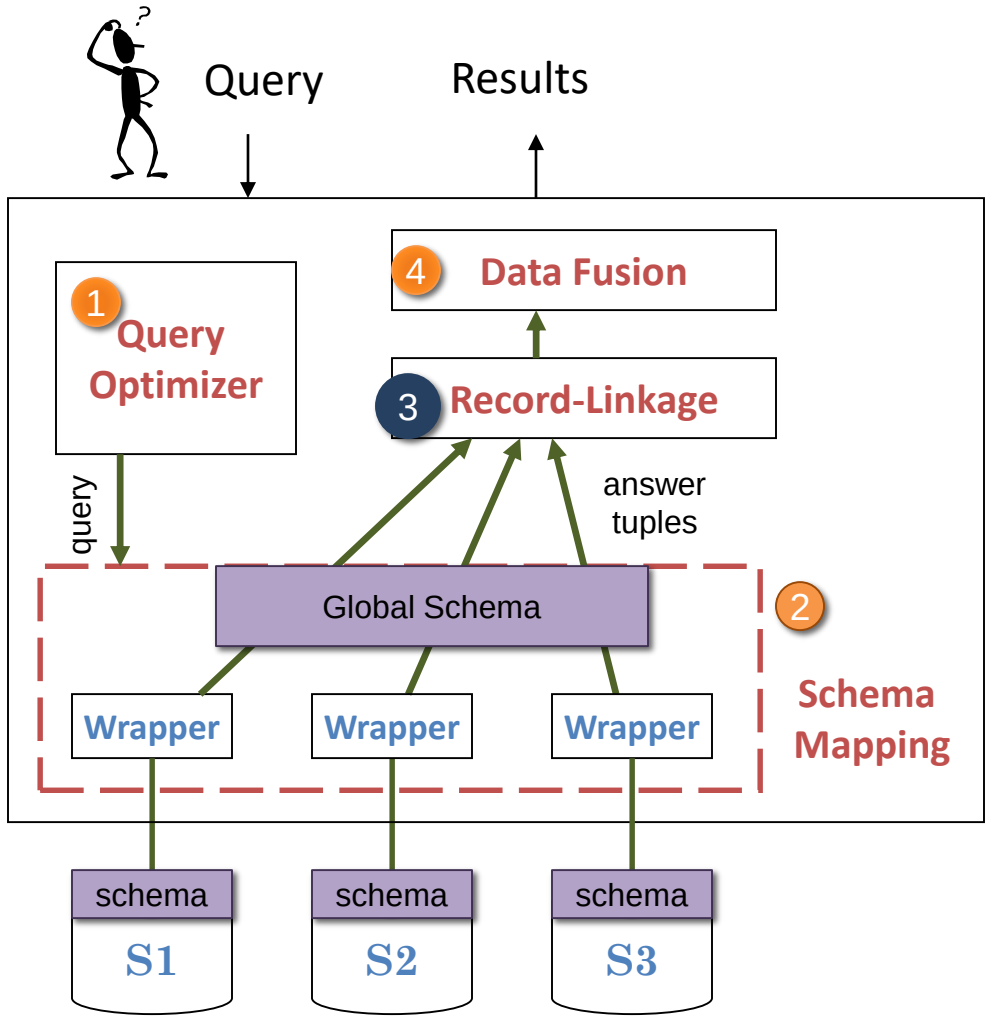
Schema Mapping Approaches (Cont.)

- Schema Matching Approaches
  - Constraint-based approach
    - Utilize contains to determine similar schema elements
      - Data types
      - Uniqueness-constraints
      - Value-range constraints
      - Relationship-types

S1 elements	S2 elements
Employee	Personnel
EmpNo – int, primary key	Pno - int, unique
EmpName – varchar (50)	Pname – string
DeptNo – int, references Department	Dept - string
Salary - dec (15,2)	Born - date
Birthdate – date	
Department	
DeptNo – int, primary key	
DeptName – varchar (40)	

* A Survey of Approaches to Automatic Schema Matching by E. Rahm , P. Bernstein

Record Linkage Phase



- Record Linkage (duplicate detection or entity resolution)
  - **Input:** List of records from sources.
  - **Output:** For each real-world object, it assigns a object-ID to the record.
  - **Objective:** Finds records from different sources that may represent the same real-world object.

Record Linkage Example

- Link records that correspond to the same 'object'

S1

Business Name	Phone	Specialty	Address	Website
Thai Boat	(805) 594-1638	Asian	3212 Broad St San Luis Obispo, CA	thaiboatslo.com
Jaffa	(805) 543-2400	Mediterranean	1308 Monterey St San Luis Obispo, CA	jaffacafe.us
Taste	(805) 541-5860	American Cafe	2900 Broad St San Luis Obispo, CA	taste2900.com

S2

Listing Name	Phone	Cuisine	Street Address	City	State
Thai Boat	805-594-1638	Thai	3212 Broad St #100	San Luis Obispo	CA
Jaffa Cafe	805-543-2449	Lebanese	1308 Monterey St	San Luis Obispo	CA
Taste	805-541-5860	American	2900 Broad St	San Luis Obispo	CA
Oasis	805-543-1155	Greek	675 Higuera St	San Luis Obispo	CA

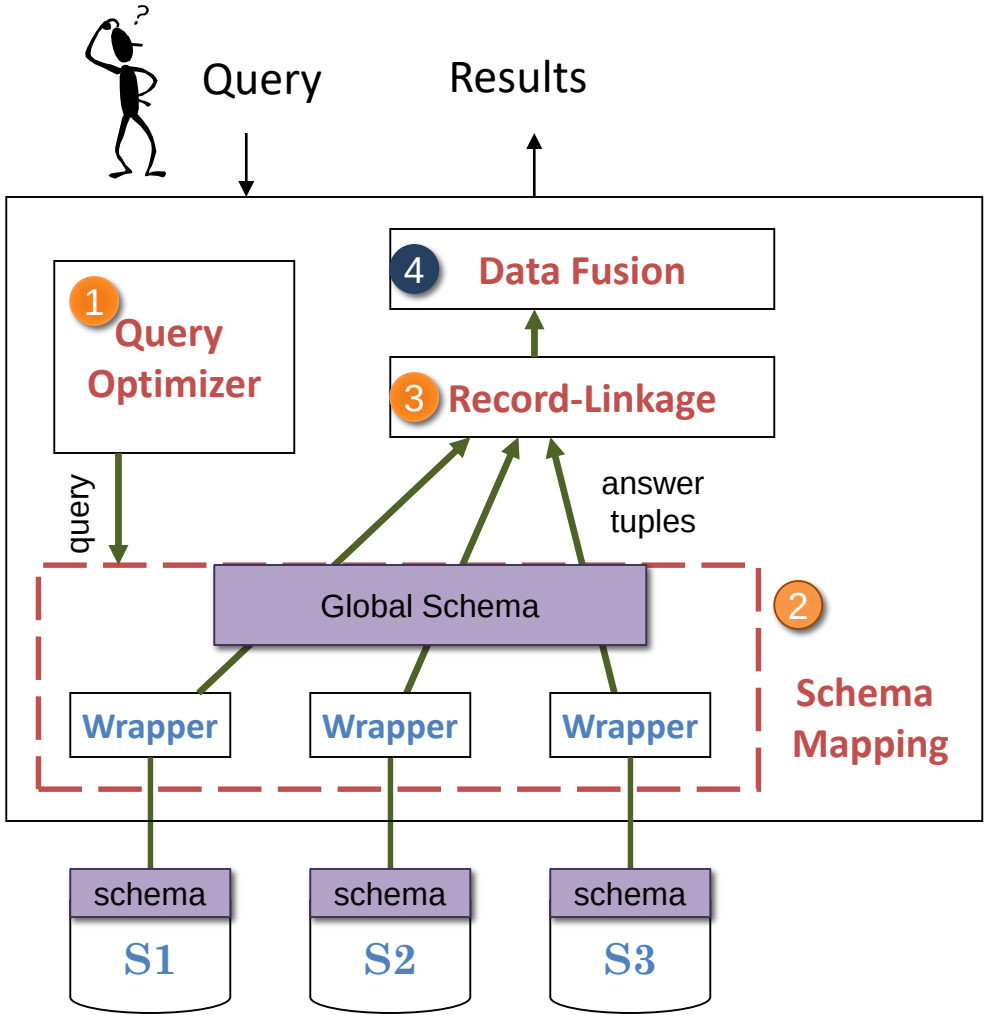
Reasons for mismatching

- Misspelling
  - “Smith”, “Smeth”, “Snith”
- Variant names, synonyms, and abbreviations
  - “St.”, “St”, “Street”…..“Prof”, “Professor”….”car”, “vehicle”
- Different systems
  - “Chin Le”, “Le, Chin”… “10/02/2000”, “10-02-2000”, “02-10-2000”
- Different domains
  - “YES/NO”, “1/0”, “T/F”

MECHANISMS FOR ENTITY RESOLUTION

- Edit Distance
  - Compare string fields using edit distance function
  - Can assign different weights to different fields
- Normalization & Ontology
  - Using a dictionary, replace all abbreviations with a standard forms
  - Ontology helps in synonyms
- Clustering and Partitioning
  - Run a clustering-based algorithm over the returned records
  - Tuples belonging to the same cluster can be further tested for matching

Data Fusion Phase



- Data Fusion
  - **Input:** Receives list of records (with object-ID) from record-linkage phase.
  - **Output:** Concise, complete, and consistent record list (answer list).
  - **Objective:** Fuse records from different sources that relate to the same real-world object and resolve inconsistencies and conflicts.
  - Also known as: data merging, consolidation, entity resolution, finding representatives/survivors, instance-level conflict resolution

Data Fusion Example

- Fuse data and resolve conflicts and inconsistencies
- Basic approach uses voting, or can utilize weighted vote (like PageRank)

For this record, S1 stated the phone # was (805)-543-2400 while S2 stated it was 805-543-2449.

Which source to trust?

Fused Data

Name	Phone	Cuisine	Address	Website
Thai Boat	805-594-1638	Thai	3212 Broad St #100 San Luis Obispo CA	thaiboatslo.com
Jaffa Cafe	805-543-2449	Mediterranean	1308 Monterey St San Luis Obispo CA	jaffacafe.us
Taste	805-541-5860	American	2900 Broad St San Luis Obispo CA	taste2900.com
Oasis	805-543-1155	Greek	675 Higuera St San Luis Obispo CA	

Data Fusion

- Naïve solution is to simply use the information that is asserted by the largest number of data sources (i.e., naive voting).
  - Its inadequate since biased (and even malicious) sources abound, and plagiarism (i.e., copying without proper attribution) between sources may be widespread.

- Ideally, when applying voting, we would like to give a higher vote to more trustworthy sources and ignore copied information.
  - First, we often do not know a priori the trustworthiness of a source.
  - Second, in many applications we do not know how each source obtains its data, so we have to discover copiers from a snapshot of data.

- Paper “Data Fusion: Resolving Conflicts from Multiple Sources” by Dong et. Al. examines the problem of data fusion in more detail. [PDF]

Strategies for Dealing with Missing Data

- After performing data fusion, we must also deal with other inconsistences in the data, such as:
  - Missing values
  - Null values

Missing Values

- There are several strategies for dealing with missing data.

- **Reducing dataset**
  - The simplest solution for the missing values imputation problem is the reduction of the data set and elimination of all samples with missing values.

- **Replace Missing Value with Mean**
  - This method replaces each missing value with mean of the attribute (Kantardzic, M. 2003). The mean is calculated based on all known values of the attribute. This method is usable only for numeric attributes and is usually combined with replacing missing values with most common attribute value for symbolic attributes.

- **Replace Missing Value with Median for the Given Class**
  - Since the mean is affected by the presence of outliers it seems natural to use the median instead just to assure robustness. In this case the missing data for a given attribute is replaced by the median of all known values of that attribute in the class where the instance with the missing value belongs (Acuña, E. & Rodriguez, C. 2004). This method is usable only for numeric attributes and requires existence of classes or possibility to create classes as previous method

Missing Values (Cont.)

- **Replace Missing Value with Most Common Attribute Value**
  - This method simply uses most common attribute value for missing value imputation (Grzymala-Busse J. W., Hu M. 2001). The most common value of all values of the attribute is used. This method is usable only for symbolic attributes and is usually combined with replacing missing values with missing values imputation using mean for numeric attributes.

- **Closest Fit**
  - The closest fit algorithm (Grzymala-Busse, J. W., Goodwin, L. K., Grzymala-Busse, W. J. & Zheng X. 2005) for missing attribute values is based on replacing a missing attribute value with an existing value of the same attribute from another case that resembles as much as possible the case with missing attribute values.
  - This approach can be generalized to using K nearest neighbor, which will look for more than one (“k”) similar cases with known values of the attribute (Hand, D., Mannila, H. & Smyth, P. 2001).

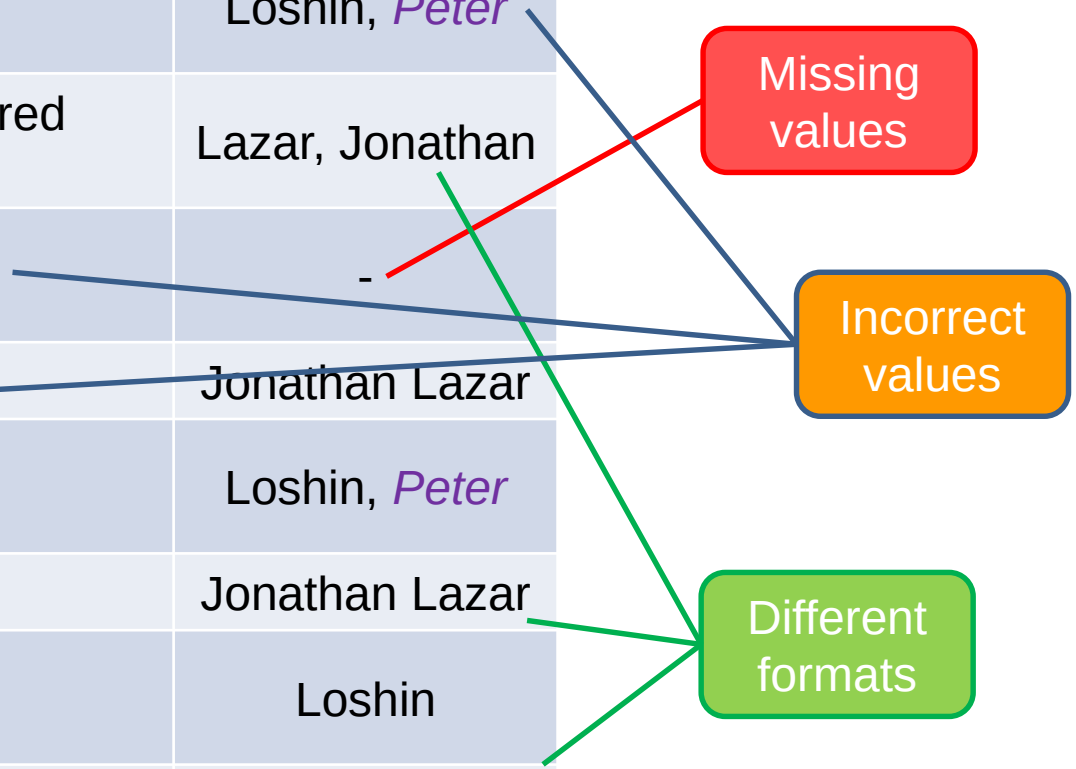
Data Fusion literature

- Data Fusion: Resolving Conflicts from Multiple Sources by Luna Dong et. al. [PDF]
  - Examined the data fusion problem over numerous data sources where sources accuracy is considered.
  - Usually, conflicts in data fusion is solved through voting.
  - Sources may copy from others, and since these sources are dependent on other sources, their vote should not hold as much weight.
  - The paper addressed the following questions:
    - How can we discover source dependence ?
    - How should we fuse data given source dependence ?

Problem Definition—Input

- Objects: a real-world entity, described by a set of attributes
  - Each associated w. a true value
- Sources: each providing data for a subset of objects

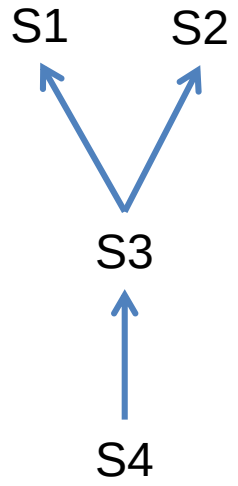
Src	ISBN	Name	Author
S1	1	IPV6: Theory, Protocol, and Practice	Loshin, *Peter*
	2	Web Usability: A User-Centered Design Approach	Lazar, Jonathan
S2	1	*IPV4: Theory, Protocol, and Practice*	-
	2	*Web Usability: A User*	Jonathan Lazar
S3	1	IPV6: Theory, Protocol, and Practice	Loshin, *Peter*
	2	*Web Usability: A User*	Jonathan Lazar
S4	1	IPV6: Theory, Protocol, and Practice	Loshin
	2	*Web Usability: A User*	Lazar



Problem Definition—Output

Src	ISBN	Name	Author
S1	1	IPV6: Theory, Protocol, and Practice	Loshin, *Peter*
	2	Web Usability: A User-Centered Design Approach	Lazar, Jonathan
S2	1	*IPV4: Theory, Protocol, and Practice*	-
	2	*Web Usability: A User*	Jonathan Lazar
S3	1	IPV6: Theory, Protocol, and Practice	Loshin, *Peter*
	2	*Web Usability: A User*	Jonathan Lazar
S4	1	IPV6: Theory, Protocol, and Practice	Loshin
	2	*Web Usability: A User*	Lazar

- For each S1, S2, decide pr of S1 copying directly from S2
  - A copier copies all or a subset of data
  - A copier can add values and verify/modify copied values—independent contribution
  - A copier can re-format copied values—still considered as copied



Summary

- Data Integration is important , the question is can we automate this process??

Ongoing research

- How to match their schemas automatically " schema matching techniques
- How to find matching records " record linkage techniques
- How to find errors, synonyms, etc. and correct them " data cleansing techniques