Advanced Data Management (CSCI 640/490)

Data Fusion

Dr. David Koop
Football Game Data

- Have each game store the id of the home team and the id of the away team (one-to-one)
- Have each player store the id of the team he plays on (many-to-one)
Concatenation

- Take two data frames with the same columns and add more rows
- `pd.concat([data-frame-1, data-frame-2, ...])`
- Default is to add rows (`axis=0`), but can also add columns (`axis=1`)
- Can also concatenate Series into a data frame.
- `concat` preserves the index so this can be confusing if you have two default indices (0,1,2,3...)—they will appear twice
  - Use `ignore_index=True` to get a 0,1,2...
Merges (aka Joins)

- Want to join the two tables based on the location and date
- Location and date are the **keys** for the join
- Merges are **ordered**: there is a left and a right side

<table>
<thead>
<tr>
<th>Id</th>
<th>Location</th>
<th>Date</th>
<th>Home</th>
<th>Away</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Boston</td>
<td>9/2</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>1</td>
<td>Boston</td>
<td>9/9</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>Cleveland</td>
<td>9/16</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>San Diego</td>
<td>9/23</td>
<td>21</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>wId</th>
<th>City</th>
<th>Date</th>
<th>Temp</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Boston</td>
<td>9/2</td>
<td>72</td>
</tr>
<tr>
<td>1</td>
<td>Boston</td>
<td>9/3</td>
<td>68</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>7</td>
<td>Boston</td>
<td>9/9</td>
<td>75</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>21</td>
<td>Boston</td>
<td>9/23</td>
<td>54</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>36</td>
<td>Cleveland</td>
<td>9/16</td>
<td>81</td>
</tr>
</tbody>
</table>

No data for San Diego
Types of Joins

- Inner: intersection of keys (match on both sides)
- Outer: union of keys (if there is no match on other side, still include with NaN to indicate missing data)
- Left: always have rows from left table (no unmatched right data)
- Right: like left, but with no unmatched left data
Data Merging in Pandas

- `pd.merge(left, right, ...)`
- Default merge: join on matching column names
- Better: specify the column name(s) to join on via `on` kwarg
  - If column names differ, use `left_on` and `right_on`
  - Multiple keys: use a list
- `how` kwarg specifies type of join ("inner", "outer", "left", "right")
- Can add suffixes to column names when they appear in both tables, but are not being joined on
- Can also merge using the index by setting `left_index` or `right_index` to True
Data Integration

```sql
select title, startTime
from Movie, Plays
where Movie.title=Plays.movie AND
    location=“New York” AND
director=“Ava DuVernay”
```

Sources S1 and S3 are relevant, sources S4 and S5 are irrelevant, and source S2 is relevant but possibly redundant.

Movie: Title, director, year, genre
Actors: title, actor
Plays: movie, location, startTime
Reviews: title, rating, description
Data Integration

• Lots of data sources, how do we answer questions where we need to access data from more than one?
• Schema matching
• Problem of heterogeneity
• AI-Complete problem: difficulty is the same as making computers as intelligent as people
• Two techniques:
  - Mediation
  - Data Warehouses
Data Integration Application: Biomedical

OMIM
HUGO
GeneClinics
SwissProt
LocusLink
Entrez
GEO

Phenotype
Gene
Sequenceable Entity
Nucleotide Sequence
Structured Vocabulary
Experiment
Microarray Experiment

[A. Doan et al., 2012]
Data Warehouses: Offline Replication

- Determine physical schema
- Define a database with this schema
- Define procedural mappings in an “ETL tool” to import the data and clean it.
- Periodically copy all of the data from the data sources
  - Note that the sources and the warehouse are basically independent at this point

[Data Warehouse Query Results]

[A. Doan et al., 2012]
Virtual Data Warehouses

Query

Mediated Schema

Independence of:
- source & location
- data model, syntax
- semantic variations
- ...

Semantic Mappings

S1

S2

S3

[<cd> <title> The best of ... </title> <artist> Carreras </artist> <artist> Pavarotti </artist> <artist> Domingo </artist> <price> 19.95 </price> </cd>]

[A. Doan et al., 2012]
Integrated Schema Example

Movie (title, director, year, genre)
Actors (title, actor)
Plays (movie, location, startTime)
Reviews (title, rating, description)

S1 Movies (name, actors, director, genre)
S2 Cinemas (place, movie, start)
S3 CinemasInNYC (cinema, title, startTime)
S4 CinemasInSF (location, movie, startingTime)
S5 Reviews (title, date, grade, review)

[A. Doan et al., 2012]
Why is Data Integration Hard?

- Systems-level reasons:
  - Managing different platforms
  - SQL across multiple systems is not so simple
  - Distributed query processing

- Logical reasons:
  - Schema (and data) heterogeneity

- ‘Social’ reasons:
  - Locating and capturing relevant data in the enterprise.
  - Convincing people to share (data fiefdoms)
    - Security, privacy and performance implications
Data Fusion

- Definition: Resolving conflicting data and verifying facts.
- Example: "Google, How long is the Mississippi River?"

[X. L. Dong and T. Rekatsinas, 2018]
### Data Fusion

#### Source Observations

<table>
<thead>
<tr>
<th>Source</th>
<th>River</th>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>KG</td>
<td>Mississippi</td>
<td>Length</td>
<td>2,320 mi</td>
</tr>
<tr>
<td>KG</td>
<td>Missouri</td>
<td>Length</td>
<td>2,341 mi</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Mississippi</td>
<td>Length</td>
<td>2,202 mi</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Missouri</td>
<td>Length</td>
<td>2,341 mi</td>
</tr>
<tr>
<td>USGS</td>
<td>Mississippi</td>
<td>Length</td>
<td>2,340 mi</td>
</tr>
<tr>
<td>USGS</td>
<td>Missouri</td>
<td>Length</td>
<td>2,540 mi</td>
</tr>
</tbody>
</table>

#### True Facts

<table>
<thead>
<tr>
<th>River</th>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mississippi</td>
<td>Length</td>
<td>?</td>
</tr>
<tr>
<td>Missouri</td>
<td>Length</td>
<td>?</td>
</tr>
</tbody>
</table>

**Goal:** Find the **latent** true value of facts.

[X. L. Dong and T. Rekatsinas, 2018]
Data Fusion

Source Observations

<table>
<thead>
<tr>
<th>Source</th>
<th>River</th>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>KG</td>
<td>Mississippi River</td>
<td>Length</td>
<td>2,320 mi</td>
</tr>
<tr>
<td>KG</td>
<td>Missouri River</td>
<td>Length</td>
<td>2,341 mi</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Mississippi River</td>
<td>Length</td>
<td>2,202 mi</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Missouri River</td>
<td>Length</td>
<td>2,341 mi</td>
</tr>
<tr>
<td>USGS</td>
<td>Mississippi River</td>
<td>Length</td>
<td>2,340 mi</td>
</tr>
<tr>
<td>USGS</td>
<td>Missouri River</td>
<td>Length</td>
<td>2,540 mi</td>
</tr>
</tbody>
</table>

True Facts

<table>
<thead>
<tr>
<th>River</th>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mississippi River</td>
<td>Length</td>
<td></td>
</tr>
<tr>
<td>Missouri River</td>
<td>Length</td>
<td></td>
</tr>
</tbody>
</table>

Idea: Use redundancy to infer the true value of each fact.

[X. L. Dong and T. Rekatsinas, 2018]
Data Fusion

Source Observations

<table>
<thead>
<tr>
<th>Source</th>
<th>River</th>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>KG</td>
<td>Mississippi River</td>
<td>Length</td>
<td>2,320 mi</td>
</tr>
<tr>
<td>KG</td>
<td>Missouri River</td>
<td>Length</td>
<td>2,341 mi</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Mississippi River</td>
<td>Length</td>
<td>2,202 mi</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Missouri River</td>
<td>Length</td>
<td>2,341 mi</td>
</tr>
<tr>
<td>USGS</td>
<td>Mississippi River</td>
<td>Length</td>
<td>2,340 mi</td>
</tr>
<tr>
<td>USGS</td>
<td>Missouri River</td>
<td>Length</td>
<td>2,540 mi</td>
</tr>
</tbody>
</table>

Majority voting can be limited. What if sources are correlated (e.g., copying)?

Idea: Model source quality for accurate results.

True Facts

<table>
<thead>
<tr>
<th>River</th>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mississippi River</td>
<td>Length</td>
<td>?</td>
</tr>
<tr>
<td>Missouri River</td>
<td>Length</td>
<td>2,341</td>
</tr>
</tbody>
</table>

MV's assumptions
1. Sources report values independently
2. Sources are better than chance.

[X. L. Dong and T. Rekatsinas, 2018]
Reading Quiz
Assignment 3

• Salary Data
• Use Pandas (not loops)
• Part 2: CSCI 640 students need to do (b), CSCI 490 students can choose
• Part 5: use melt/pivot or a similar high-level operation
Today

• Due to faculty candidate visit, office hours changed (12:30-1:30pm)
• Faculty candidate talks both today and tomorrow
  - 3:30 to 4:30 FW 201
Record Linkage Motivation

- Often data from different sources need to be integrated and linked
  - To allow data analyses that are impossible on individual databases
  - To improve data quality
  - To enrich data with additional information

- **Lack of unique entity identifiers** means that linking is often based on personal information

- When databases are linked across organisations, maintaining privacy and confidentiality is vital

- The linking of databases is challenged by **data quality**, **database size**, and **privacy concerns**
Motivating Example

- Preventing the outbreak of epidemics requires monitoring of occurrences of unusual patterns of symptoms, ideally in real time.
- Data from many different sources will need to be collected (including travel and immigration records; doctors, emergency and hospital admissions; drug purchases; social network and location data; and possibly even animal health data).

[P. Christen, 2019], image: [Pharexia, Wikipedia]
Record Linkage

P. Christen
Record Linkage Process

1. **Database A**
   - Data pre-processing
   - Indexing / Searching
   - Comparison

2. **Database B**
   - Data pre-processing
   - Indexing / Searching
   - Comparison

3. **Classification**
   - Matches
   - Non-matches
   - Potential Matches
   - Evaluation

4. **Clerical Review**
   - Potential Matches
Record Linkage Techniques

- Deterministic matching
  - Rule-based matching (complex to build and maintain)

- Probabilistic record linkage [Fellegi and Sunter, 1969]
  - Use available attributes for linking (often personal information, like names, addresses, dates of birth, etc.)
  - Calculate match weights for attributes

- “Computer science” approaches
  - Based on machine learning, data mining, database, or information retrieval techniques
  - Supervised classification: Requires training data (true matches)
  - Unsupervised: Clustering, collective, and graph based

[P. Christen, 2019]
Record Linkage/Entity Resolution Recipe

- Problem: Link references to the same entity
- Short Answers:
  - Random Forest with attribute similarity features
  - Deep Learning to handle text and noise
  - End-to-end solutions still being worked on

[X. L. Dong and T. Rekatsinas, 2018]
Data Integration and Data Fusion

• Data Integration: focus on integrating data from different sources
• When sources are orthogonal, no problems
• What happens when two sources provide the same type of information and they conflict?
• Data Fusion: create a single object while resolving conflicting values
Data Fusion—
Resolving Data Conflicts in Integration

X. L. Dong and F. Naumann
Data Fusion Summary

• Conflict resolution strategies
• "Truth-discovery" techniques
  - Accuracy
  - Freshness
  - Dependence
• Fusion Issues
  - Accuracy
  - Efficiency
  - Usability
  - How fusion fits with the rest of data integration?
Data Conflicts

Integrated data

Schering CRM

Bayer CRM

[Rx]

[A]

[Caduceus]

[Plus]

[L. Dong and F. Naumann, 2009]
Information Integration

[L. Dong and F. Naumann, 2009]

Source A

Source B

Schema Mapping

Data Transformation

Duplicate Detection

Data Fusion

[Scheth & Larson, 1990]
Information Integration

Source A

Source B

Schema Mapping
Data Transformation
Duplicate Detection
Data Fusion

Preserve lineage

[L. Dong and F. Naumann, 2009]
Data Fusion

• Problem: Given a duplicate, create a single object representation while resolving conflicting data values.

• Difficulties:
  - Null values: Subsumption and complementation
  - Contradictions in data values
  - Uncertainty & truth: Discover the true value and model uncertainty in this process
  - Metadata: Preferences, recency, correctness
  - Lineage: Keep original values and their origin
  - Implementation in DBMS: SQL, extended SQL, UDFs, etc.
Conflict Resolution Strategies

Conflict ignorance
- PASS IT ON

Conflict avoidance
- instance based
  - TAKE THE INFORMATION NO GOSSIPING

Conflict resolution
- instance based
  - deciding
    - CRY WITH THE WOLVES ROLL THE DICE
  - mediating
    - MEET IN THE MIDDLE

- metadata based
  - deciding
    - TRUST YOUR FRIENDS
  - mediating
    - NOTHING IS OLDER THAN THE NEWS FROM YESTERDAY

[Scholten and Naumann, 2009]
Integrating Conflicting Data: The Role of Source Dependence

X. L. Dong, L. Berti-Equille, and D. Srivastava
Example Problem
### Example Problem

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stonebraker</td>
<td>MIT</td>
<td>Berkeley</td>
<td>MIT</td>
</tr>
<tr>
<td>Dewitt</td>
<td>MSR</td>
<td>MSR</td>
<td>UWisc</td>
</tr>
<tr>
<td>Bernstein</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
</tr>
<tr>
<td>Carey</td>
<td>UCI</td>
<td>AT&amp;T</td>
<td>BEA</td>
</tr>
<tr>
<td>Halevy</td>
<td>Google</td>
<td>Google</td>
<td>UW</td>
</tr>
</tbody>
</table>
Naive Voting Works

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stonebraker</td>
<td>MIT</td>
<td>Berkeley</td>
<td>MIT</td>
</tr>
<tr>
<td>Dewitt</td>
<td>MSR</td>
<td>MSR</td>
<td>UWisc</td>
</tr>
<tr>
<td>Bernstein</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
</tr>
<tr>
<td>Carey</td>
<td>UCI</td>
<td>AT&amp;T</td>
<td>BEA</td>
</tr>
<tr>
<td>Halevy</td>
<td>Google</td>
<td>Google</td>
<td>UW</td>
</tr>
</tbody>
</table>

[X L Dong et al., 2009]
Naive Voting Only Works if Data Sources are Independent

[X L Dong et al., 2009]
Naive Voting Only Works if Data Sources are Independent

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stonebraker</td>
<td>MIT</td>
<td>Berkeley</td>
<td>MIT</td>
<td>MIT</td>
<td>MS</td>
</tr>
<tr>
<td>Dewitt</td>
<td>MSR</td>
<td>MSR</td>
<td>UWisc</td>
<td>UWisc</td>
<td>UWisc</td>
</tr>
<tr>
<td>Bernstein</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
</tr>
<tr>
<td>Carey</td>
<td>UCI</td>
<td>AT&amp;T</td>
<td>BEA</td>
<td>BEA</td>
<td>BEA</td>
</tr>
<tr>
<td>Halevy</td>
<td>Google</td>
<td>Google</td>
<td>UW</td>
<td>UW</td>
<td>UW</td>
</tr>
</tbody>
</table>

[X L Dong et al., 2009]
S4 and S5 copy from S3

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stonebraker</td>
<td>MIT</td>
<td>Berkeley</td>
<td>MIT</td>
<td>MIT</td>
<td>MS</td>
</tr>
<tr>
<td>Dewitt</td>
<td>MSR</td>
<td>MSR</td>
<td>UWisc</td>
<td>UWisc</td>
<td>UWisc</td>
</tr>
<tr>
<td>Bernstein</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
</tr>
<tr>
<td>Carey</td>
<td>UCI</td>
<td>AT&amp;T</td>
<td>BEA</td>
<td>BEA</td>
<td>BEA</td>
</tr>
<tr>
<td>Halevy</td>
<td>Google</td>
<td>Google</td>
<td>UW</td>
<td>UW</td>
<td>UW</td>
</tr>
</tbody>
</table>

[X L Dong et al., 2009]
S4 and S5 copy from S3

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stonebraker</td>
<td>MIT</td>
<td>Berkeley</td>
<td>MIT</td>
<td>MIT</td>
<td>NS</td>
</tr>
<tr>
<td>Dewitt</td>
<td>MSR</td>
<td>MSR</td>
<td>UWisc</td>
<td>UWisc</td>
<td>UWisc</td>
</tr>
<tr>
<td>Bernstein</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
</tr>
<tr>
<td>Carey</td>
<td>UCI</td>
<td>AT&amp;T</td>
<td>BEA</td>
<td>BEA</td>
<td>BEA</td>
</tr>
<tr>
<td>Halevy</td>
<td>Google</td>
<td>Google</td>
<td>UW</td>
<td>UW</td>
<td>UW</td>
</tr>
</tbody>
</table>

[X L Dong et al., 2009]
## Challenges in Dependence Discovery

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stonebraker</td>
<td>MIT</td>
<td>Berkeley</td>
<td>MIT</td>
<td>MIT</td>
<td>MS</td>
</tr>
<tr>
<td>Dewitt</td>
<td>MSR</td>
<td>MSR</td>
<td>UWisc</td>
<td>UWisc</td>
<td>UWisc</td>
</tr>
<tr>
<td>Bernstein</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
</tr>
<tr>
<td>Carey</td>
<td>UCI</td>
<td>AT&amp;T</td>
<td>BEA</td>
<td>BEA</td>
<td>BEA</td>
</tr>
<tr>
<td>Halevy</td>
<td>Google</td>
<td>Google</td>
<td>UW</td>
<td>UW</td>
<td>UW</td>
</tr>
</tbody>
</table>

[X L Dong et al., 2009]
Challenges in Dependence Discovery

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stonebraker</td>
<td>MIT</td>
<td>Berkeley</td>
<td>MIT</td>
<td>MIT</td>
<td>MS</td>
</tr>
<tr>
<td>Dewitt</td>
<td>MSR</td>
<td>MSR</td>
<td>UWisc</td>
<td>UWisc</td>
<td>UWisc</td>
</tr>
<tr>
<td>Bernstein</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
</tr>
<tr>
<td>Carey</td>
<td>UCI</td>
<td>AT&amp;T</td>
<td>BEA</td>
<td>BEA</td>
<td>BEA</td>
</tr>
<tr>
<td>Halevy</td>
<td>Google</td>
<td>Google</td>
<td>UW</td>
<td>UW</td>
<td>UW</td>
</tr>
</tbody>
</table>

[X L Dong et al., 2009]
## Challenges in Dependence Discovery

2. With only a snapshot it is hard to decide which source is a copier.

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stonebraker</td>
<td>MIT</td>
<td>Berkeley</td>
<td>MIT</td>
<td>MIT</td>
<td>MS</td>
</tr>
<tr>
<td>Dewitt</td>
<td>MSR</td>
<td>MSR</td>
<td>UWisc</td>
<td>UWisc</td>
<td>UWisc</td>
</tr>
<tr>
<td>Bernstein</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
</tr>
<tr>
<td>Carey</td>
<td>UCI</td>
<td>AT&amp;T</td>
<td>BEA</td>
<td>BEA</td>
<td>BEA</td>
</tr>
<tr>
<td>Halevy</td>
<td>Google</td>
<td>Google</td>
<td>UW</td>
<td>UW</td>
<td>UW</td>
</tr>
</tbody>
</table>

[X L Dong et al., 2009]
### Challenges in Dependence Discovery

1. Sharing common data does not in itself imply copying.

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stonebraker</td>
<td>MIT</td>
<td>Berkeley</td>
<td>MIT</td>
<td>MIT</td>
<td>MS</td>
</tr>
<tr>
<td>Dewitt</td>
<td>MSR</td>
<td>MSR</td>
<td>UWisc</td>
<td>UWisc</td>
<td>UWisc</td>
</tr>
<tr>
<td>Bernstein</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
</tr>
<tr>
<td>Carey</td>
<td>UCI</td>
<td>AT&amp;T</td>
<td>BEA</td>
<td>BEA</td>
<td>BEA</td>
</tr>
<tr>
<td>Halevy</td>
<td>Google</td>
<td>Google</td>
<td>UW</td>
<td>UW</td>
<td>UW</td>
</tr>
</tbody>
</table>

2. With only a snapshot it is hard to decide which source is a copier.

3. A copier can also provide or verify some data by itself, so it is inappropriate to ignore all of its data.

[X L Dong et al., 2009]
Source Dependence

- Source dependence: two sources S and T deriving the same part of data directly or transitively from a common source (can be one of S or T).
  - Independent source
  - Copier
    - copying part (or all) of data from other sources
    - may verify or revise some of the copied values
    - may add additional values
- Assumptions
  - Independent values
  - Independent copying
  - No loop copying

[X L Dong et al., 2009]
Core Case

• Conditions
  - Same source accuracy
  - Uniform false-value distribution
  - Categorical value

• Proposition: With independent “good” sources, Naïve voting selects values with highest probability to be true.

[X. L. Dong et al., 2009]
Ideas

- If two sources share a lot of false values, they are more likely to be dependent.
- S1 is more likely to copy from S2, if the accuracy of the common data is highly different from the accuracy of S1.
Combining Accuracy and Dependence

Truth Discovery

Source-accuracy Computation

Dependence Detection

[X L Dong et al., 2009]
Combining Accuracy and Dependence

Step 1

Source-accuracy Computation

Step 2

Truth Discovery

Step 3

Dependence Detection

[X L Dong et al., 2009]
## The Motivating Example

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stonebraker</td>
<td>MIT</td>
<td>Berkeley</td>
<td>MIT</td>
<td>MIT</td>
<td>MS</td>
</tr>
<tr>
<td>Dewitt</td>
<td>MSR</td>
<td>MSR</td>
<td>UWisc</td>
<td>UWisc</td>
<td>UWisc</td>
</tr>
<tr>
<td>Bernstein</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
</tr>
<tr>
<td>Carey</td>
<td>UCI</td>
<td>AT&amp;T</td>
<td>BEA</td>
<td>BEA</td>
<td>BEA</td>
</tr>
<tr>
<td>Halevy</td>
<td>Google</td>
<td>Google</td>
<td>UW</td>
<td>UW</td>
<td>UW</td>
</tr>
</tbody>
</table>

![Diagrams showing the Motivating Example](image_url)

[X L Dong et al., 2009]
## The Motivating Example

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Round 1</td>
<td>.52</td>
<td>.42</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
</tr>
<tr>
<td>Round 2</td>
<td>.63</td>
<td>.46</td>
<td>.55</td>
<td>.55</td>
<td>.55</td>
</tr>
<tr>
<td>Round 3</td>
<td>.71</td>
<td>.52</td>
<td>.53</td>
<td>.53</td>
<td>.37</td>
</tr>
<tr>
<td>Round 4</td>
<td>.79</td>
<td>.57</td>
<td>.48</td>
<td>.48</td>
<td>.31</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Round 11</td>
<td>.97</td>
<td>.61</td>
<td>.40</td>
<td>.40</td>
<td>.21</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value Confidence</th>
<th>Carey</th>
<th>Halevy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UCI</td>
<td>AT&amp;T</td>
</tr>
<tr>
<td>Round 1</td>
<td>1.61</td>
<td>1.61</td>
</tr>
<tr>
<td>Round 2</td>
<td>1.68</td>
<td>1.3</td>
</tr>
<tr>
<td>Round 3</td>
<td>2.12</td>
<td>1.47</td>
</tr>
<tr>
<td>Round 4</td>
<td>2.51</td>
<td>1.68</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Round 11</td>
<td>4.73</td>
<td>2.08</td>
</tr>
</tbody>
</table>

[X L Dong et al., 2009]