Data Integration

Duen Horng (Polo) Chau
Assistant Professor
Associate Director, MS Analytics
Georgia Tech

Partly based on materials by
Professors Guy Lebanon, Jeffrey Heer, John Stasko, Christos Faloutsos
What is **Data Integration**?
Why is it Important?
Data Integration

Combining data from different sources to provide the user with a unified view

How to help people effectively leverage multiple data sources?
(People: analysts, researchers, practitioners, etc.)
Examples businesses that derive value via data integration
City of Atlanta, GA: Home
www.atlantaga.gov/
Mayor Reed delivers the first 96-gallon recycling cart to a home in Southwest Atlanta. The citywide distribution of the carts known as "Cartlanta" is a major...

Atlanta - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Atlanta
Atlanta (pron.: /aɪˈtælənə/ or /ætˈlaːntə/ locally /ætˈlæna/)) is the capital of and the most populous city in the U.S. state of Georgia, with an estimated 2011 population of 432,427. Demographics of Atlanta - Atlanta metropolitan area - Colleges and Universities

Atlanta, Georgia - Hotels, Events & Things to Do in Atlanta, GA
www.atlanta.net/
Explore Atlanta, GA events, attractions, restaurants, hotels and packages with this official Atlanta, Georgia guide for travelers and locals, brought to you by the...

50 Fun Things to Do in Atlanta - Atlanta Convention and Visitor's ... www.atlanta.net/50fun/
Check out our guide to the top 50 Fun Things to Do in Atlanta by activity or neighborhood. The Atlanta Convention & Visitors Bureau is your guide to finding fun...

Things to do in Atlanta | www.accessatlanta.com
www.accessatlanta.com/
1 hour ago – Find things to do in Atlanta: Concerts, shows, arts, special events, movies & restaurants. Blogs, celeb news & photos. In Atlanta, it's...

News for atlanta
Winter weather advisory posted for metro Atlanta
Atlanta Journal Constitution - 5 hours ago
Metro Atlanta began the day Thursday under a flood watch, and will end the day under a winter weather advisory for the chance of snow and ...

Five Giant losses: Awful in Atlanta
ESPN (blog) - 1 hour ago
Josh Smith suspended one game

Atlanta

Area: 132.4 sq miles (342.9 km²)
Founded: 1837
Weather: 48°F (9°C), Wind N at 0 mph (0 km/h), 93% Humidity
Local time: Thursday 12:10 PM

Upcoming events
Jan 17
Blue Man Group
Fox Theatre Atlanta
Jan 17
Purity Ring at Variety Playhouse on Jan 17, 2013
Variety Playhouse
Jan 18
Ellie Goulding w/ St. Lucia
The Tabernacle

Points of interest
Know the score.
Ask Siri for baseball, basketball, football, hockey, and soccer scores as well as schedules, rosters, and stats.

Choose a movie.
Ask Siri to get showtimes, look up movie facts, play trailers, show you reviews, and more.

Find a restaurant.
Ask Siri to search by different criteria or a combination. Siri gets you photos, reviews, and reservations.
Craigslist now has map view! What problem has it solved?

https://atlanta.craigslist.org/search/hhh
Compare hundreds of travel sites at once.
Find the best deals faster.

Round-trip  One-way  Multi-city

ATL  To

Depart  Return

1 adult

add nearby airports  add nearby airports  find hotels  find car rentals

More search options

I am a student

New York  Las Vegas  Los Angeles  San Francisco
More Examples?
How to do data integration?
“Low” Effort Approaches

1. Use database’s “Join”! (e.g., SQLite)

When does this approach work?
(Or, when does it NOT work?)

<table>
<thead>
<tr>
<th>id</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>Smith</td>
</tr>
<tr>
<td>222</td>
<td>Johnson</td>
</tr>
<tr>
<td>333</td>
<td>Obama</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>id</th>
<th>state</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>GA</td>
</tr>
<tr>
<td>222</td>
<td>NY</td>
</tr>
<tr>
<td>333</td>
<td>CA</td>
</tr>
</tbody>
</table>

2. Google Refine

http://openrefine.org  (video #3)
So, it’s great to assign an **ID** to everything!

But how?
Crowd-sourcing Approaches: Freebase

Freebase intro: https://www.youtube.com/watch?v=TJfrNo3Z-DU
Freebase to move over to Wikidata in July (2015): http://goo.gl/3ZDTg7
http://wiki.freebase.com/wiki/What_is_Freebase%3F
Freebase
(a graph of entities)

“...a large collaborative knowledge base consisting of metadata composed mainly by its community members...”

Wikipedia.
So what?
What can you do with Freebase?

Hint: Google acquired it in 2010
The Knowledge Graph

Learn more about one of the key breakthroughs behind the future of search.

https://www.youtube.com/watch?v=mmQI6VGvX-c
Introducing Graph Search

People who like **Cycling** and are from my hometown

Find people who share your interests

Want to start a book club or find a gym buddy? Connect with friends who like the same activities—and meet new people, too.

https://www.youtube.com/watch?v=W3k1USQbq80  
https://www.facebook.com/about/graphsearch
Feldspar
Finding Information by Association
Polo Chau, Brad Myers, Andrew Faulring

CHI 2008

YouTube: http://www.youtube.com/watch?v=Q0TIV8F_o_E&feature=youtu.be&list=ULQ0TIV8F_o_E
What if we don’t have the luxury of having IDs?

A common problem in academia:

Polo Chau
Duen Horng Chau
Duen Chau
D. Chau

(Screenshot from FreeBase video)
Then you need to do…

**Entity Resolution**

(A hard problem in data integration)
Why is entity resolution important?

Case Study
Let’s shop for an iPhone 6 on Apple, Amazon and eBay
Buy iPhone 6

Model

- iPhone 6
  4.7-inch display
  From $199

- iPhone 6 Plus
  5.5-inch display
  From $299

Finish

- Silver
- Gold
- Space Gray

Storage

- 16GB
  From $199
Apple iPhone 6 a1549 16GB (AT&T) - Gold Silver or Gray
Carrier Locked, Includes Charger, Free Shipping
$419.99
List price: $649.00
Buy It Now
Free shipping
82+ sold
35% off

Apple iPhone 6 a1549 16GB - (Unlocked) Gold Gray or Silver
Refurbished w/ 30 Day Guarantee - Charger - Free Ship!
$549.00
List price: $649.00
Buy It Now
Free shipping
678+ sold
15% off

Apple iPhone 6 Plus - 64GB (Factory Unlocked) Smartphone - Gold Silver Gray
Original Open Box & Accessories Included - Top Seller
$764.99
D-Dupe

Interactive Data Deduplication and Integration
TVCG 2008

University of Maryland
Bilgic, Licamele, Getoor, Kang, Shneiderman

http://www.cs.umd.edu/projects/linqs/ddupe/ (skip to 0:55)
Numerous similarity functions

Excellent read: http://infolab.stanford.edu/~ullman/mmds/ch3a.pdf

- Euclidean distance
  Euclidean norm / L2 norm

- TaxiCab/Manhattan distance

- Jaccard Similarity (e.g., used with w-shingles)
  e.g., overlap of nodes’ #neighbors

  \[ \text{Jaccard similarity of sets } S \text{ and } T \text{ is } \frac{|S \cap T|}{|S \cup T|} \]

- String edit distance
  e.g., “Polo Chau” vs “Polo Chan”

- Canberra distance
  (compare ranked items)

Figure 3.1: Two sets with Jaccard similarity 3/8
Core components: **Similarity functions**

Determine how two entities are similar.

D-Dupe’s approach:

**Attribute similarity** + **relational similarity**

\[
sim(e_i, e_j) = (1 - \alpha) \times \text{sim}_A(e_i, e_j) + \alpha \times \text{sim}_R(e_i, e_j),
\]

\[0 \leq \alpha \leq 1,
\]

**Similarity score** for a pair of entities
Attribute similarity (a weighted sum)

\[ sim_A(e_i, e_j) = \sum_{k=1}^{n} w_k \times sim_{\text{fun}_k}(e_i \cdot a_k, e_j \cdot a_k), \]

\[-1 \leq w_k \leq 1 \quad \text{and} \quad \sum_{k=1}^{n} |w_k| = 1,\]