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?
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.: /ætˈlaːntə/, stressed /ætˈlaːntə/, locally /ætˈlaːna/) is the capital of and the most populous city in the U.S. state of Georgia, with an estimated 2011 population of 432,427. Wikipedia
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 ...

Atlanta Journals 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

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.
More Examples

- Amazon (products): music/movie info
- Facebook: pages from wikipedia (google)
- mint.com: bank/credit card accounts info
- flipboard/google news/apple news: articles
- ...

9
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>
<th>id</th>
<th>state</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>Smith</td>
<td>111</td>
<td>GA</td>
</tr>
<tr>
<td>222</td>
<td>Johnson</td>
<td>222</td>
<td>NY</td>
</tr>
<tr>
<td>333</td>
<td>Obama</td>
<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

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
Feldspar
Feldspar

A system that helps people find things on their computers when typical search or browsing tools don’t work

An example scenario…
“Find the webpage mentioned in the email from the person I met at an event“
“Find the **webpage** mentioned in the **email** from the **person** I met at an **event**“

If I can’t remember the specifics, such as any text in the webpage, email, etc.

→ Can’t search
"Find the **webpage** mentioned in the **email** from the **person** I met at an **event**"

If I can’t remember the specifics, such as any text in the **webpage**, **email**, etc.

→ Can’t search

If I haven’t bookmarked the **webpage**

→ Can’t browse
“Find the **webpage** mentioned in the **email** from the **person** I met at an **event**”
“Find the **webpage** mentioned in the **email** from the **person** I met at an **event**”

But I can describe the **webpage** with a chain of **associations**.
“Find the **webpage** mentioned in the **email** from the **person** I met at an **event**“

But I can describe the **webpage** with a chain of **associations**.

**webpage** – **email** – **person** – **event**
“Find the **webpage** mentioned in the **email** from the **person** I met at an **event**“

But I can describe the **webpage** with a chain of **associations**.

**webpage** – **email** – **person** – **event**

The psychology literature has shown that people often remember things exactly like this.
Natural question:
Can I find things by associations?
Natural question:
Can I find things by associations?

Can I find the webpage by specifying its associated information (email, person, and event)?
Natural question:
Can I find things by associations?

Can I find the **webpage** by specifying its associated information (**email**, **person**, and **event**)?

We created Feldspar, which supports this associative retrieval of information.
Feldspar stands for....
**Feldspar** stands for....

Finding Elements by Leveraging Diverse Sources of Pertinent Associative Recollection
DEMO

YouTube: http://www.youtube.com/watch?v=Q0TIV8F_o_E&feature=youtu.be&list=ULQ0TIV8F_o_E
Implementation: Overview

Create a graph database to store the associations among items on the computer.

Develop an algorithm that processes the query and returns results.
Creating an Association Database (a graph)

Install Google Desktop and let it index all the items on the computer.
Creating an Association Database (a graph)

Focus on 7 types

Install Google Desktop and let it index all the items on the computer

- *filetype:calendar*
- *filetype:email*
- *filetype:doc, etc.*
- *filetype:web*
Creating an Association Database (a graph)

Focus on 7 types

Install **Google Desktop** and let it index all the items on the computer

Identify **associations** and build our database, which is a **directed graph**

```
filetype:calendar
filetype:email
filetype:doc, etc.
filetype:web
```
Creating an Association Database (a graph)

Identify associations and build our database, which is a directed graph.

Focus on 7 types:
- People
- Emails
- Folders
- Files
- Webpages
- Dates
- Events

Install Google Desktop and let it index all the items on the computer.

Example queries:
- filetype:calendar
- filetype:email
- filetype:doc, etc.
- filetype:web
Creating an Association Database (a graph)

Focus on 7 types

Install Google Desktop and let it index all the items on the computer

Identify associations and build our database, which is a directed graph

Creating an Association Database (a graph)

Focus on 7 types

Identify associations and build our database, which is a directed graph

What to Do When Search Fails: Finding Information by Association
Polo Chau, Brad Myers, Andrew Faulring
Algorithm that processes the query
Algorithm that processes the query

Here are the **webpages** related to **emails** related to **persons** related to **events**: 2 results

<table>
<thead>
<tr>
<th>Title</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><a href="http://www.shure.com/index.htm">http://www.shure.com/index.htm</a></td>
</tr>
<tr>
<td></td>
<td><a href="http://www.pittsburghsymphony.org">http://www.pittsburghsymphony.org</a></td>
</tr>
</tbody>
</table>
Algorithm that processes the query

webpages - emails - persons - events
Algorithm that processes the query

webpages - emails - persons - events
Algorithm that processes the query

webpages - emails - persons - events

Sample association graph database
Algorithm that processes the query

webpages - emails - persons - events

webpages - emails - persons

Sample association graph database

Results Generator
Algorithm that processes the query

- webpages - emails - persons - events
- webpages - emails - persons
- webpages - emails

Sample association graph database

Results Generator
Algorithm that processes the query

1. webpages - emails - persons - events
2. webpages - emails - persons
3. webpages - emails
4. webpages

Sample association graph database

Results Generator
Algorithm that processes the query

webpages - emails - persons - events

For 7 data types \( \rightarrow \) needs \( 7 \times 7 = 49 \) results generators

One Results Generator for each pair of associations
Usability Study: Questions to Answer

Do people understand how to use Feldspar?

Does Feldspar perform well on complicated tasks (which are difficult if not using Feldspar)?

Does Feldspar also works well for simple everyday tasks?
Usability Study: Study Design
Usability Study: Study Design

Within-subject design

Two groups of software

**Feldspar**
**Control**: conventional desktop applications, including Outlook and its built-in browsing and querying mechanisms, Google Desktop, and the Windows Explorer

Two similar task sets
  Task set A and B
  Each set has 7 tasks
  Tasks ask people to find things on a computer
Usability Study: Study Design & Participants
Usability Study: Study Design & Participants

Four conditions, counter-balanced for software order and task set order:

(Feldspar + Task Set A) then (Control + Task Set B)
(Feldspar + Task Set B) then (Control + Task Set A)
(Control + Task Set A) then (Feldspar + Task Set B)
(Control + Task Set B) then (Feldspar + Task Set A)
Usability Study: Study Design & Participants

Four conditions, counter-balanced for *software order* and *task set order*

(Feldspar + Task Set A) then (Control + Task Set B)
(Feldspar + Task Set B) then (Control + Task Set A)
(Control + Task Set A) then (Feldspar + Task Set B)
(Control + Task Set B) then (Feldspar + Task Set A)

8 participants, two in each conditions
User Study: Task Set A

1. Open the last email received on July 27, 2007
2. Open all the email attachments of type .txt
3. Find out who had email conversations with the person who sent out the file file.doc
4. Find out who attended the event in which Cara was present.
5. Find all the events that were attended by anyone who has sent you a file.
6. Open the file folders that contain email attachments from Spence.
7. Open the webpage mentioned in the email from the person you met in an event in May.
User Study: Task Set A

1. Open the last email received on **July 27, 2007**
2. Open all the email attachments of type .txt
3. Find out who had email conversations with the person who sent out the file **file.doc**
4. Find out who attended the event in which **Cara** was present.
5. Find all the events that were attended by anyone who has sent you a file.
6. Open the file folders that contain email attachments from **Spence**.
7. Open the webpage mentioned in the email from the person you met in an event in **May**.
User Study: Task Set A

1. Open the last email received on July 27, 2007
2. Open all the email attachments of type .txt
3. Find out who had email conversations with the person who sent out the file file.doc
4. Find out who attended the event in which Cara was present.
5. Find all the events that were attended by anyone who has sent you a file.
6. Open the file folders that contain email attachments from Spence.
7. Open the webpage mentioned in the email from the person you met in an event in May.
Quantitative Results

Average completion time for each task (in seconds)

![Average Task Completion Times (seconds)](chart)

Number of fails for each task

![Number of Participants Failing Tasks](chart)

- Statistically significant tasks marked by * (p<0.05)
- Shorter bars are better
Qualitative Results

Which Software Seemed ...

- Easier to learn
- Easier to use
- Faster
- More liked
- More enjoyable
- More accurate

# of participants

- Feldspar
- Control
- The Same
Conclusions

Feldspar

Provides a usable interface for finding desktop information by interactively and incrementally specifying multiple levels of associations;
Builds graph database for storing associations, identified from Google Desktop;
Uses real-time algorithms to provide answers to queries.

Feldspar could be a useful addition to conventional search and browsing tools.
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
Functional Protection. Battery Case for iPhone 6

Showing results in Cell Phones & Accessories. Show instead results in All Departments.

Related Searches: iphone 5, iphone 5s, iphone 6 plus.

Apple iPhone 6, Silver, 16 GB (Unlocked)
by Apple
$698.00 ✓Prime
In stock on August 30, 2015

Apple iPhone 6 16GB (4.7-inch)
4G LTE Factory Unlocked GSM
Dual-Core Smartphone - Gold
by Apple
$685.99 $1,069.99
Only 5 left in stock - order soon.

iPhone 6, Clear Case, Crystal Clear Case, Crystal Clear Case, Crystal Clear Case, Silicone...
by NXX
$6.40 $24.90 ✓Prime
Get it by Friday, Aug 28

FREE Shipping on orders over $35

FREE Shipping

FREE Shipping

FREE Shipping

FREE Shipping

FREE Shipping

FREE Shipping
eBay and PayPal are now separate companies. We've updated the eBay and PayPal User Agreements and Privacy Notices. Learn more

Search for iPhone 6 on eBay

Related: iPhone 5, iPhone 4, iPhone 5c, Samsung Galaxy S4, iPhone 1, iPhone 5 case, iPhone 5 unlocked, Samsung Galaxy S3, iPhone 3

Sort: Best Match | View:

Categories
- Cell Phones & Accessories
  - Cell Phones & Smartphones
    - Cell Phone Cases, Covers & Skins

See all categories

Storage Capacity
- 128GB
- 64GB
- 32GB
- 16GB
- Not Specified

Model
- iPhone 6
- iPhone 6 Plus
- iPhone 5s

Network
- AT&T
- Sprint
- T-Mobile
- Verizon

Color
- Black
- Gold
- Gray
- Pink
- Silver
- White

Screen Size

Popular on eBay

Apple iPhone 6 a1549 16GB (AT&T) - Gold Silver or Gray
Carrier Locked, Includes Charger, Free Shipping
$419.99
List price: $640.00
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 Shipping
$549.00
List price: $646.00
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

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

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

- **Canberra distance**
  (compare ranked items)
Distance and Similarity Measures

Different measures of distance or similarity are convenient for different types of analysis. The Wolfram Language provides built-in functions for many standard distance measures, as well as the capability to give a symbolic definition for an arbitrary measure.

**Reference**

**Numerical Data**
- EuclideanDistance
- SquaredEuclideanDistance
- NormalizedSquaredEuclideanDistance
- ManhattanDistance
- ChessboardDistance
- BrayCurtisDistance
- CanberraDistance
- CosineDistance
- CorrelationDistance
- BinaryDistance
- TimeWarpingDistance

**Boolean Data**
- HammingDistance
- JaccardDissimilarity
- MatchingDissimilarity
- DiceDissimilarity
- RogersTanimotoDissimilarity
- RussellRaoDissimilarity
- SokalSneathDissimilarity
- YuleDissimilarity

**String Data**
- EditDistance
- DamerauLevenshteinDistance
- HammingDistance
- SmithWatermanSimilarity
- NeedlemanWunschSimilarity

**Images & Colors**
- ImageDistance
- ColorDistance

**Geospatial & Temporal Data**
- GeoDistance
- DateDifference

Core components: \textbf{Similarity functions}

Determine how two entities are similar.

D-Dupe’s approach:

\textbf{Attribute similarity} + \textbf{relational similarity}

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

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

\textbf{Similarity score} for a pair of entities
Attribute similarity (a weighted sum)

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

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