Data Integration

Duen Horng (Polo) Chau
Associate Professor, College of Computing
Associate Director, MS Analytics
Machine Learning Area Leader, College of Computing
Georgia Tech

Partly based on materials by
Professors Guy Lebanon, Jeffrey Heer, John Stasko, Christos Faloutsos
What is **Data Integration**?
Combining data from multiple sources to provide the user with a unified view.

**Why is it Important?**
Think about the apps, websites, and services that you use every day.
Businesses derive value through data integration.
**Apple Siri**

**Getting Answers**

- “How is the Nikkei doing?”
- “When is daylight saving time?”
- “What’s the latest in San Francisco?”

See what people are saying on social media about a place or event.

“Was that an earthquake?”
Search hundreds of travel sites at once.

**Stay up-to-date**
Subscribe now and receive the latest travel news.

**Recommended for you**

[Images of recommended destinations]
More Examples?

• **Social media** (data from users, businesses)
  • Facebook: your posts, advertisements, review

• **Search engine**: Google, Bing, Yahoo, etc.

• **Smart assistants**: Siri, Cortana, Alexa

• **Price comparison**: Kayak

• Uber, Lyft: drivers, traffic data, customers

• google maps: users, restaurants, traffic….
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>Lee</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>id</th>
<th>salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>$40k</td>
</tr>
<tr>
<td>222</td>
<td>$60k</td>
</tr>
<tr>
<td>333</td>
<td>$50k</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>id</th>
<th>name</th>
<th>salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>Smith</td>
<td>$40k</td>
</tr>
<tr>
<td>222</td>
<td>Johnson</td>
<td>$60k</td>
</tr>
<tr>
<td>333</td>
<td>Lee</td>
<td>$50k</td>
</tr>
</tbody>
</table>

2. Open Refine
   http://openrefine.org  (Video #3 “Reconcile and Match Data”)
IDs are really important, and can simplify data integration!

But who creates the IDs?
Crowd-sourcing Approaches: Freebase

Freebase intro video: https://youtu.be/TJfrNo3Z-DU

Learn more about Freebase at https://en.wikipedia.org/wiki/Freebase
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 the Freebase knowledge graph?

Hint: Google acquired it in 2010.
The Knowledge Graph

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

Google Knowledge Graph video: https://youtu.be/mmQl6VGvX-c
Freebase replaced by Google Knowledge Graph API

Example:
What does Google know about Taylor Swift?

https://developers.google.com/knowledge-graph/
What does Google know about Taylor Swift?
https://developers.google.com/knowledge-graph/

```json
"@type": "ItemList",
"ItemListElement": [
{
"@type": "EntitySearchResult",
"result": {
"@id": "kg:/m/0d1567",
"name": "Taylor Swift",
"@type": [
"Thing",
"Person"
],
"description": "Singer-songwriter",
"image": {
"contentUrl": "https://t1.gstatic.com/images?q=tbn:ANd9GcQmVDAhjhWnN2OWys2ZMO3P GAhu*
"url": "https://en.wikipedia.org/wiki/Taylor_Swift",
"license": "http://creativecommons.org/licenses/by-sa/2.0"
},
"detailedDescription": {
"articleBody": "Taylor Alison Swift is an American singer-songwriter and actress. R
"url": "http://en.wikipedia.org/wiki/Taylor_Swift",
},
"url": "http://taylorswift.com/
```
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** so difficult?

Let’s understand it through shopping for an iPhone on Apple, Amazon and eBay.
Buy your new iPhone X.

Get free next-business-day delivery on any in-stock iPhone ordered by 5:00 p.m.*

iPhone X
5.8-inch display*

From $49.91/mo. with the iPhone Upgrade Program.¹
Or pay in full from $999.
Find deals and best selling products for Apple iPhone X Cell Phones & Smartphones

Shop by Model

- iPhone X
- iPhone 8 Plus
- iPhone 8
- iPhone 7 Plus
- iPhone 7

Apple iPhone X 64GB - GSM & CDMA Unlocked - USA Model - $990.00
Buy It Now
Free Shipping
914+ Sold

Apple iPhone X 256GB - GSM&CDMA Unlocked-USA - $1,145.00
Buy It Now
Free Shipping
1523+ Sold

Apple iPhone X - 64GB - Space Gray [Factory Unlocked] - Brand
$1,350.00 or Best Offer
Free Shipping
256 Sold

24K Gold Plated Apple iPhone X 256GB - Silver Unlocked Custom
$1,999.00 or Best Offer
Free Shipping
15 Watching

BRAND NEW, Apple iPhone X MOCSLL/A A1865 64GB Silver
D-Dupe

Interactive Data Deduplication and Integration
TVCG 2008

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

https://linqspub.soe.ucsc.edu/basilic/web/Publications/2006/bilgic:vast06/
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)

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

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

Excellent Tutorial on Entity Resolution


by Lise Getoor and Ashwin Machanavajjhala