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.
Atlanta, GA: Home

Atlanta - Wikipedia

Atlanta is the capital of and the most populous city in the U.S. state of Georgia, with an estimated 2021 population of 472,522. Atlanta is the cultural and ...
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.

“Do I need an umbrella today?”

“Was that an earthquake?”
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>
<th>id</th>
<th>salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>Smith</td>
<td>111</td>
<td>$40k</td>
</tr>
<tr>
<td>222</td>
<td>Johnson</td>
<td>222</td>
<td>$60k</td>
</tr>
<tr>
<td>333</td>
<td>Lee</td>
<td>333</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/

```
"@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:ANd9GcQmVDAhjhwWnN2OWys2ZMO3PGAhuPZCQ15zWcS66wQ1",
      "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*

Select

From $49.91/mo. with the iPhone Upgrade Program.¹
Or pay in full from $999.
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/
Polo

Palo

Alice
Bob
Carol
Dave
Core components: **Similarity functions**

Determine how two entities are similar.

D-Dupe’s approach:

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

\[
sim(e_i, e_j) = (1 - \alpha) \times \text{sim}_A(e_i, e_j) + \alpha \times \text{sim}_R(e_i, e_j), \quad 0 \leq \alpha \leq 1,
\]

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

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

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

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

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

Figure 3.1: Two sets with Jaccard similarity 3/8
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.

### 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