Data Cleaning & Integration

Duen Horng (Polo) Chau
Georgia Tech

Partly based on materials by Professors Guy Lebanon, Jeffrey Heer, John Stasko, Christos Faloutsos
Data Cleaning

Why data can be dirty?
How dirty is real data?

Examples

• ...

How dirty is real data?

Examples

- duplicates
- empty rows
- abbreviations (different kinds)
- difference in scales / inconsistency in description/ sometimes include units
- typos
- missing values
- trailing spaces
- incomplete cells
- synonyms of the same thing
- skewed distribution (outliers)
- bad formatting / not in relational format (in a format not expected)
More to read

Big Data's Dirty Problem [Fortune]
http://fortune.com/2014/06/30/big-data-dirty-problem/

For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights [New York Times]
Data Janitor
Data Cleaners

Watch videos

• Open Refine  (previously Google Refine)
• Data Wrangler  (research at Stanford)

Write down

• Examples of **data dirtiness**
• Tool’s **features** demo-ed (or that you like)

Will collectively summarize similarities and differences afterwards

Open Refine: [http://openrefine.org](http://openrefine.org)
Data Wrangler: [http://vis.stanford.edu/wrangler/](http://vis.stanford.edu/wrangler/)
Welcome!

OpenRefine (formerly Google Refine) is a powerful tool for working with messy data: cleaning it; transforming it from one format into another; extending it with web services; and linking it to databases like Freebase.

Please note that since October 2nd, 2012, Google is not actively supporting this project, which has now been rebranded to OpenRefine. Project development, documentation and promotion is now fully supported by volunteers. Find out more about the history of OpenRefine and how you can help the community.

Using OpenRefine - The Book

Using OpenRefine, by Ruben Verborgh and Max De Wilde, offers a great introduction to OpenRefine. Organized by recipes with hands on examples, the book covers the following topics:

1. Import data in various formats
2. Explore datasets in a matter of seconds
Wrangler is an interactive tool for data cleaning and transformation. Spend less time formatting and more time analyzing your data.

UPDATE: The Wrangler research project is complete, and the software is no longer actively supported. The team behind Wrangler has moved on to work on a commercial venture, Trifacta.

Why wrangle?

- Too much time is spent manipulating data just to get analysis and visualization tools to read it. Wrangler is designed to accelerate this process: spend less time fighting with your data and more time learning from it.
- Wrangler allows interactive transformation of messy, real-world data into the data tables analysis tools expect. Export data for use in Excel, R, Tableau, Protovis, ...
- Want to learn more about Wrangler's design? Take a look at our research paper.
- Wrangler is still a work-in-progress. Please share your feedback and feature requests!

TRY IT NOW
How are the tools similar or different?

- [G] cluster similar entities (e.g., T&M), synonyms
- [G, W] history
- [G] trailing spaces
- [W] text extraction
- [W] export script, code (work on other systems? interoperability)
- [W, G] one-click (usability)
- [G] distribution of data (apply log scale)
- [W] pivot data (unfold)
- [W] suggestions (even more usable)
- [W + G] preview changes

G = Google Refine
W = Data wrangler
The videos only show some of the tools’ features. Try them out.

Google Refine: http://code.google.com/p/google-refine/
Data Wrangler: http://vis.stanford.edu/wrangler/
Data Integration
Course Overview

- Collection
- Cleaning
- Integration
- Analysis
- Visualization
- Presentation
- Dissemination
What is **Data Integration**?
Why is it Important?
Data Integration

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

As data’s **volume**, **velocity** and **variety** increase, and **veracity** decreases, data integration presents new (and more) **opportunities and challenges**

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.: /ætˈlæntə/, stressed /ætˈlæntə/, locally /ætˈlæna/) is the capital of and the most populous city in the U.S. state of Georgia, with an estimated 2011 ... 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.

[Search options]

[Find Flights]
More Examples?

- Google Now
- Yelp?
- Amazon — different kinds of product (dpreview.com)
- Dealnews, slickdeals, fatwallet?
- tinder (facebook and instagram)
- facebook (news stories, etc.)
- walmart (different merchants)
- search engines
- ? any websites with advertising (e.g., new york times)
More Examples?

• [FREE] Mint: account app, integrates multiple account (credit card, bank, etc.), can parse receipts
• Google News
• Crime mapping
• Feedly
• app that check gas prices, coupons
• zillow-trulia/redfin
• imdb (movie database)
• coin: combine multiple credits
• ebay
More Examples?

- Palantir gotham
- Yelp: restaurant reviews, business reviews
- Facebook friend request: look at your friends’s friends and recommend those friends as your friends
- Trulia / zillow (real estate sites)
- Graph search (facebook)
- Waze
- Yahoo pipe
- Google search engine
- Google transit
- Google now / Apple Siri
How to do data integration?
“Low” Effort Approaches

Use database’s “Join”! (e.g., SQLite)
When would this approach work?
(Or, when it won’t 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>

Google Refine
http://code.google.com/p/google-refine/  (video #3)
Crowd-sourcing Approaches: Freebase

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
Freebase to move over to Wikidata in July (2015): http://goo.gl/3ZDTg7
Given a graph of entities, like Freebase, what other cool things can you do?
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.
Feldspar

Finding Information by Association.

CHI 2008

Polo Chau, Brad Myers, Andrew Faulring

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

We need ways to identify the many ways that one thing may be called. How?
Entity Resolution
(A hard problem in data integration)

Polo Chau
P. Chau
Duen Horng Chau
Duen Chau
D. Chau
Why is entity resolution so 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
Narrow your choices

Cell Phones & Accessories
Cell Phone Cases (11,768,664)
Accessories (459,701)
Cases, Holsters & Clips (12,139,658)
Screen Protectors (26,592)
Wallet Cell Phone Cases (202,237)
Flip Cell Phone Cases (129,872)
Replacement Parts (4,641)
Data Cables (13,724)

Eligible for Free Shipping
Free Shipping by Amazon

Cell Phone Internal Memory

4 GB
8 GB
16 GB
32 GB
64 GB

Brand
Apple
Spigen
Tech Armor
Maxboost
iBlasson
Ringke
DandyCase
JETech
Luvitt
Bear Motion
Samsung
Beluga
Nupro
MYX

Apple iPhone 6, Silver, 16 GB (Unlocked)
by Apple

$698.00 ✓Prime
In stock on August 30, 2015

More Buying Choices
$529.48 new (47 offers)
$505.00 used (34 offers)

FREE Shipping on orders over $35

Apple iPhone 6 16GB (4.7-inch) 4G LTE Factory Unlocked GSMDual-Core Smartphone - Gold
by Apple

$685.99 $1,069.99
Only 5 left in stock - order soon.

More Buying Choices
$499.00 new (60 offers)
$489.00 used (36 offers)

FREE Shipping on orders over $35

iPhone 6, Clear Case, Crystal Clear Case, Hybrid Bumper Case With, Ultra Clear Back Panel Cover, Clear Case, Silicone...
by MXX

$6.40 $24.99 ✓Prime
Get it by Friday, Aug 28

More Buying Choices
$4.25 new (7 offers)

FREE Shipping on orders over $35

Related Searches: iphone 5, iphone 5s, iphone 6 plus.
eBay and PayPal are now separate companies. We’ve updated the eBay and PayPal User Agreements and Privacy Notices. Learn more

Hi! Sign in or register | Daily Deals | Gift Cards | Sell | Help & Contact

My eBay

Search

Advanced

Categories

Cell Phones & Accessories

Cell Phones & Smartphones

Cell Phone Cases, Covers & Skins

More

See all categories

Storage Capacity

- 128GB (666)
- 64GB (2,021)
- 32GB (14)
- 16GB (333)
- Not Specified (518)

Model

- iPhone 6 (5,514)
- iPhone 6 Plus (2,091)
- iPhone 5s (18)

Network

- AT&T (1,501)
- Sprint (655)
- T-Mobile (666)
- Verizon (243)

Color

- Black (221)
- Gold (2,001)
- Gray (2,008)
- Pink (18)
- Silver (1,501)
- White (295)

Screen Size

- 4" or less (1,000)
- 4.7" (1,001)
- 5.5" (1,000)

Sort: Best Match

View:

Popular on eBay

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

Apple iPhone 5C - 16gb - Factory GSM...
$219.88
Buy It Now
Free shipping
Popular

Apple iPhone 5 - 16GB - (Factory Unlocked)...
$219.88
Buy It Now
Free shipping
Popular
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)
Core components: **Similarity functions**

Determine how two entities are similar.

D-Dupe’s approach:

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

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

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

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

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

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

- Enable new services
- Improve existing services
- New ways to interactive with data