http://poloclub.gatech.edu/cse6242

CSE6242/CX4242: Data & Visual Analytics

Data Integration

Duen Horng (Polo) Chau

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.

Google

atlanta

All | News | Maps | Images | Videos | More | Tools

About 1,650,000,000 results (0.96 seconds)

Top stories

News about grand jury

INSIDER

Atlanta DA: Trump grand jury report should be secret

7 hours ago

The New York Times

Atlanta D.A. Wants Grand Jury Findings Kept Private in Trump Inquiry

1 hour ago

Axios

Atlanta district attorney pushes to keep Trump 2020 election report secret

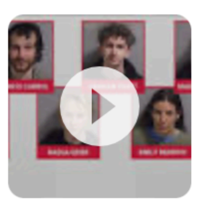
2 hours ago

Protests in Atlanta over police shooting of activist >

11 ALIVE

All but 1 arrested during protest that turned violent in Downtown Atlanta are from…

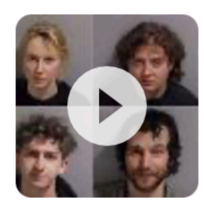
2 days ago



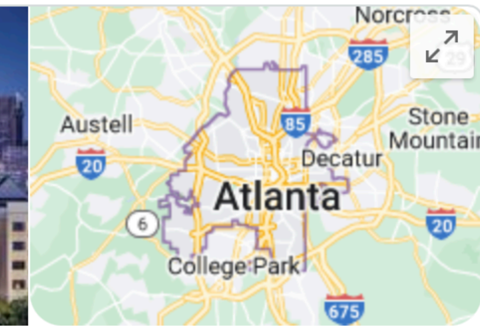
FOX NEWS

Two Atlanta riot suspects granted bond, 4 others denied after domestic…

7 hours ago



Also in the news



Atlanta

City in Georgia

Atlanta is the capital of the U.S. state of Georgia. It played an important part in both the Civil War and the 1960s Civil Rights Movement. Atlanta History Center chronicles the city's past, and the Martin Luther King Jr. National Historic Site is dedicated to the African-American leader's life and times. Downtown, Centennial Olympic Park, built for the 1996 Olympics, encompasses the massive Georgia Aquarium. — Google

Mayor: Andre Dickens Trending

Elevation: 738′

Weather: 52°F (11°C), Wind SE at 8 mph (13 km/h), 37% Humidity

More on weather.com

Local time: Tuesday 4:57 PM

Population: 496,461 (2021)

Siri does all this. And more.

All | Calls and Texts | Knowledge and Answers | Smart Home | Everyday Tasks | Navigation and Maps | Music and Podcasts | TV and Movies | Sports

Knowledge and Answers
Hey Siri, what's tomorrow's forecast for Honolulu?

Navigation and Maps
Hey Siri, show me the map

Navigation and Maps
Hey Siri, biking directions to Golden Gate Park

Calls and Texts
Hey Siri, tell Nisha I'll be home in 15 minutes

Calls and Texts
Hey Siri, return my last call

Music and Podcasts
Hey Siri, play more like this

Knowledge and Answers
Hey Siri, how's the stock market doing?

TV and Movies
Hey Siri, how are the reviews for Tenet?

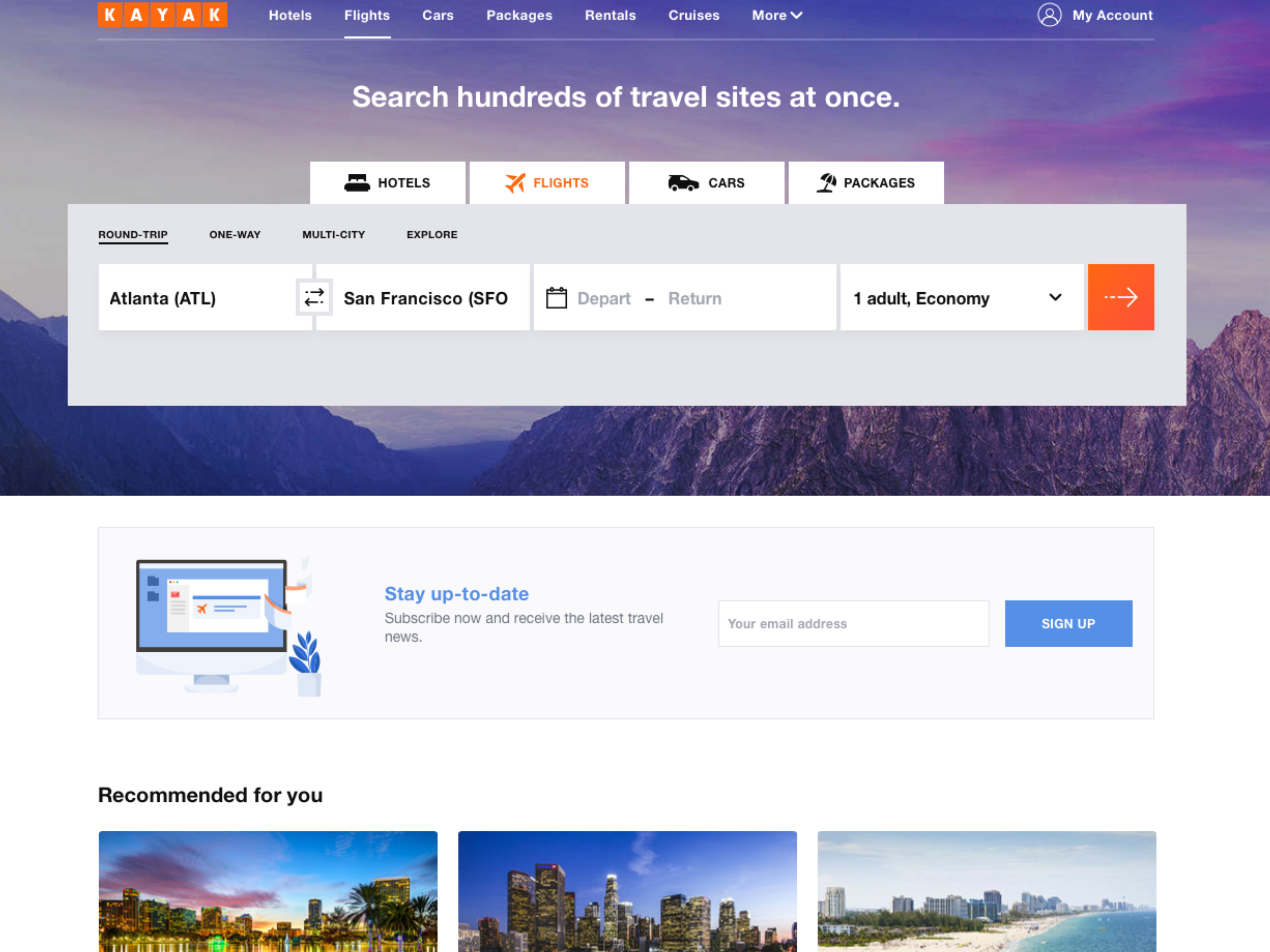
d Podcasts
iri, play Charli

Music and Podcasts
Hey Siri, play something I can dance to

Navigation and Maps
Hey Siri, find coffee shops that take Apple Pay

Navigation and Maps
Hey Siri, open Maps

Navigation and
Hey Siri, w ETA?



KAYAK
Hotels
Flights
Cars
Packages
Rentals
Cruises
More
My Account
Search hundreds of travel sites at once.
HOTELS
FLIGHTS
CARS
PACKAGES
ROUND-TRIP
ONE-WAY
MULTI-CITY
EXPLORE
Atlanta (ATL)
San Francisco (SFO
Depart – Return
1 adult, Economy
Stay up-to-date
Subscribe now and receive the latest travel news.
Your email address
SIGN UP
Recommended for you

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?)

id	name
111	Smith
222	Johnson
333	Lee

id	salary
111	$40k
222	$60k
333	$50k

id	name	salary
111	Smith	$40k
222	Johnson	$60k
333	Lee	$50k

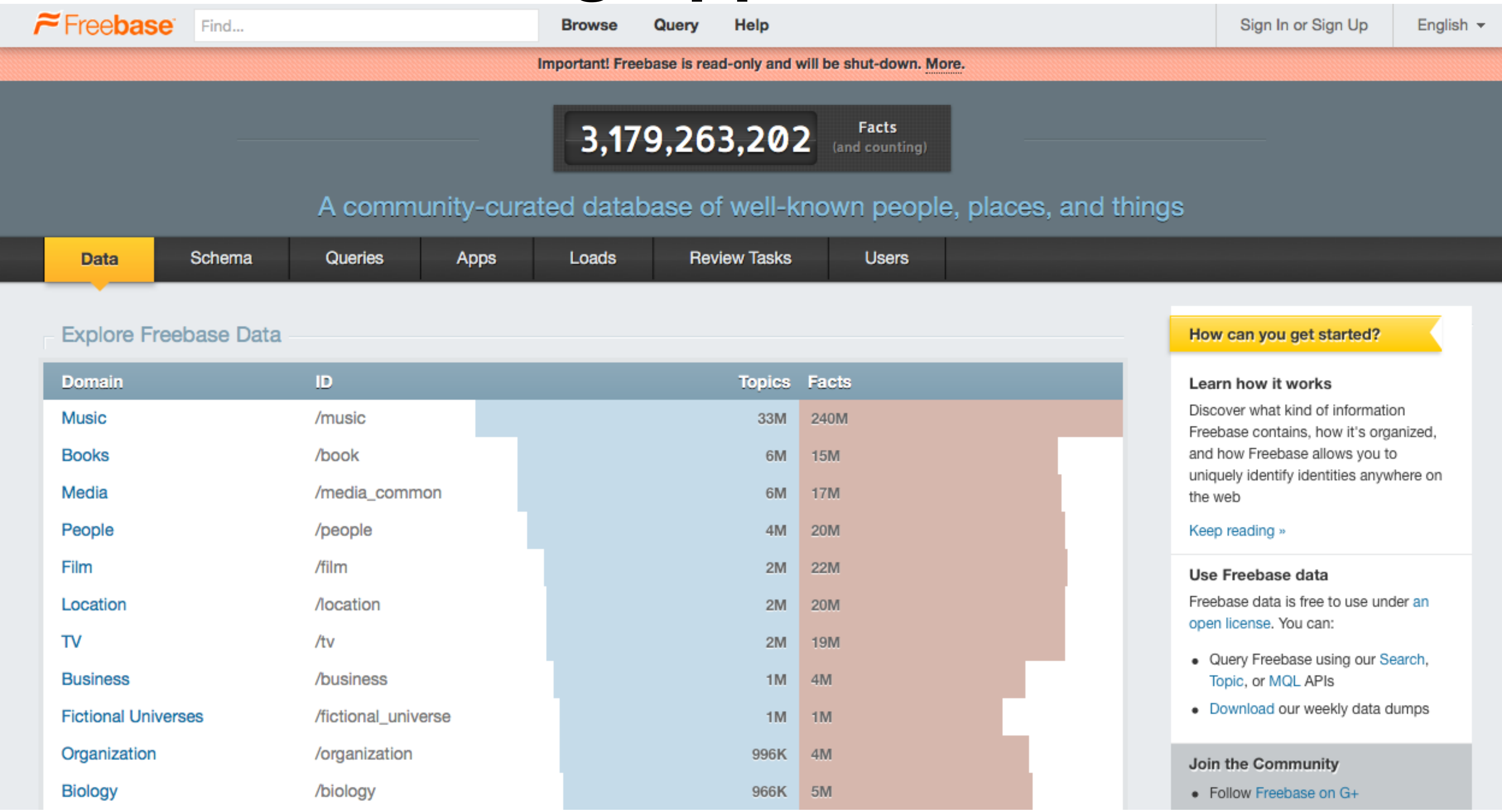
2. **Open Refine**

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)

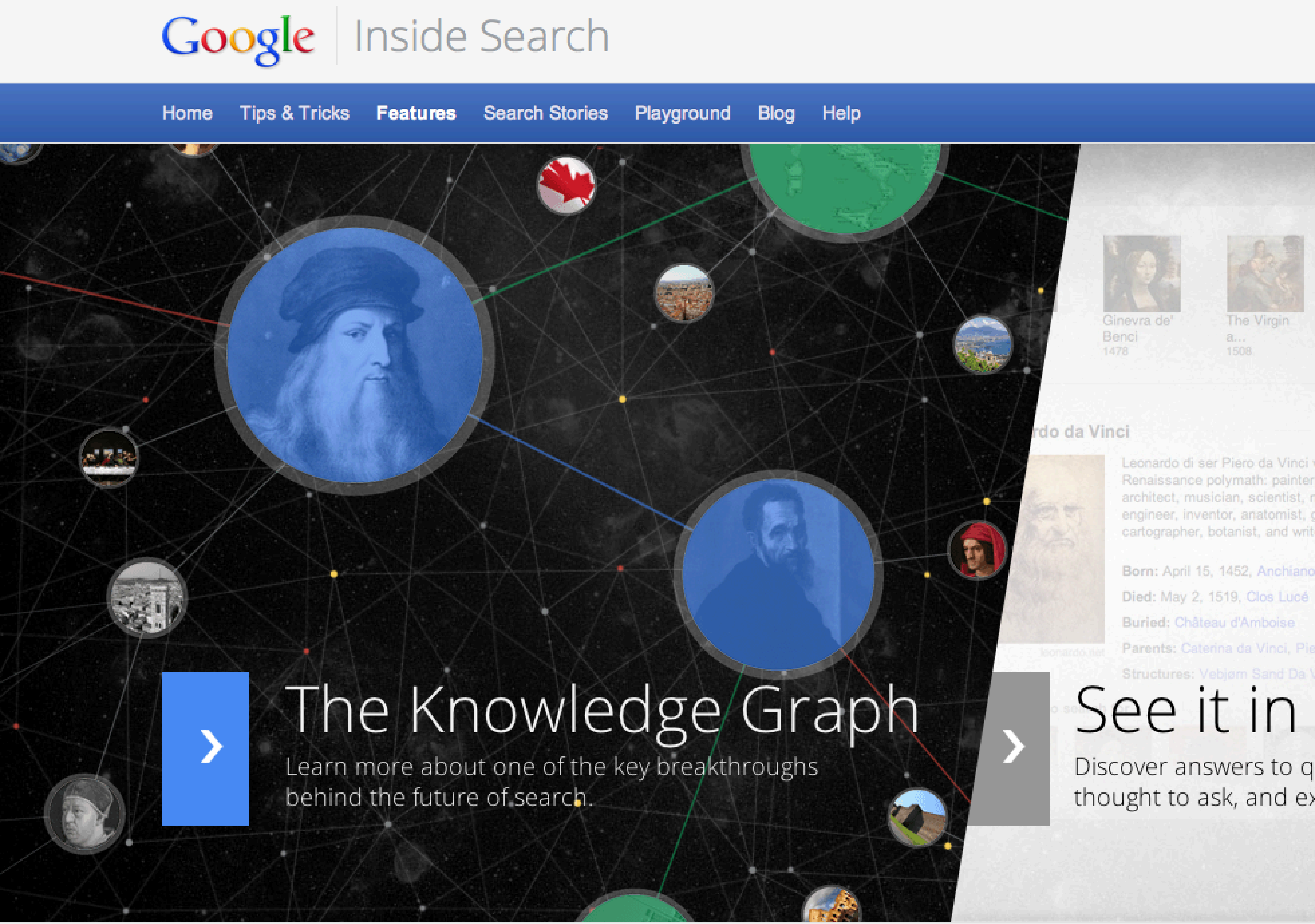
"…a large collaborative knowledge base consisting of metadata composed mainly by its **community members**…"

Wikipedia.

So what?

What can you do with the Freebase knowledge graph?

Hint: Google acquired it in 2010.



Google Knowledge Graph video: https://youtu.be/mmQl6VGvX-c

Freebase replaced by **Google Knowledge Graph API**

Example:
What does Google know about Taylor Swift?

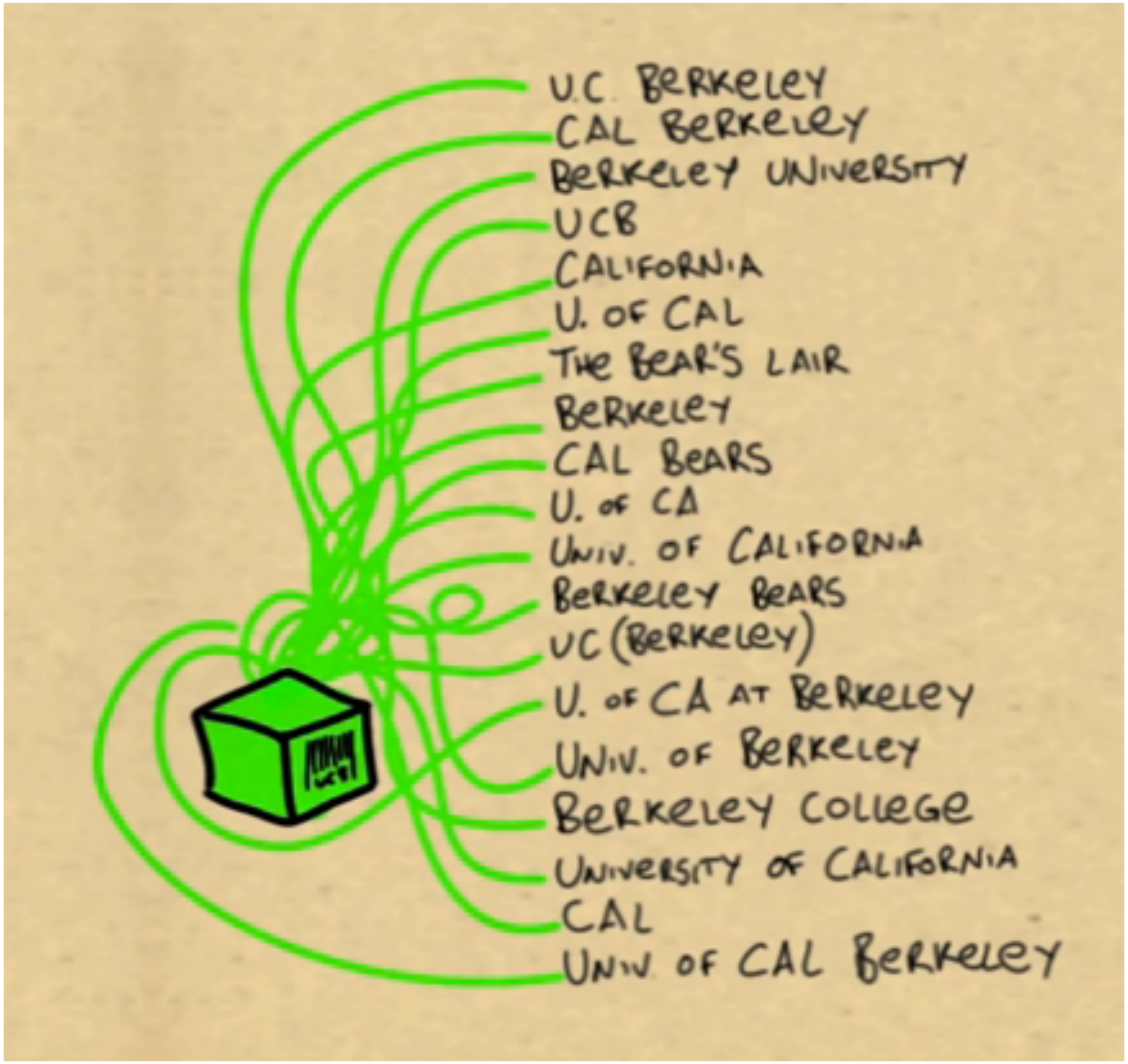
https://developers.google.com/knowledge-graph/

What does Google know about Taylor Swift?

https://developers.google.com/knowledge-graph/


```
  "@type": "ItemList",
  "itemListElement": [
    {
      "@type": "EntitySearchResult",
      "result": {
        "@id": "kg:/m/0dl567",
        "name": "Taylor Swift",
        "@type": [
          "Thing",
          "Person"
        ],
        "description": "Singer-songwriter",
        "image": {
          "contentUrl": "https://t1.gstatic.com/images?q=tbn:ANd9GcQmVDAhjhWnN2OWys2ZMO3PGAhu
          "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",
          "license": "https://en.wikipedia.org/wiki/Wikipedia:Text_of_Creative_Commons_Attrib
        },
        "url": "http://taylorswift.com/"
```

What if we don't have the luxury of having IDs ?



(Screenshot from FreeBase video)

A common problem in academia:

Polo Chau
Duen Horng Chau
Duen Chau
D. Chau

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 and eBay

Store Mac iPad iPhone Watch Vision AirPods TV & Home Entertainment Accessories Support

Carrier Deals at Apple
See all deals

Pay as low as $2.75/mo after trade-in.[Δ]

Pay as low as $0 after trade-in.[°]

Pay as low as $0 after trade-in.[ΔΔ]

New

Buy iPhone 15

From $799 or $33.29/mo. for 24 mo.*

Get $30–$620 for your trade-in.[◊◊]

Get 3% Daily Cash back with Apple Card.

Model. Which is best for you?

iPhone 15
6.1-inch display[1]
From $799 or $33.29/mo. for 24 mo.*

iPhone 15 Plus
6.7-inch display[1]
From $899 or $37.45/mo. for 24 mo.*

Hi! Sign in or register Daily Deals Brand Outlet Gift Cards Help & Contact Sell Watchlist My eBay

Shop by category

iphone 15 | Cell Phones & Acc... | Search | Advanced

Related: iphone 14 iphone 14 pro iphone 14 pro max iphone 13 iphone 15 unlocked iphone 15 pro iphone 15 pro max iphone 12 iphone 14 plus iphone 11 .. Include description

Category

All

Cell Phones & Accessories

- Cell Phone Accessories
- Cell Phone & Smartphone Parts
- Cell Phones & Smartphones
- Phone Cards & SIM Cards
- More +

Computers/Tablets & Networking

Consumer Electronics

Home & Garden

Business & Industrial

Art

Show More +

Compatible Model

- For Apple iPhone 13 (4,005,262)
- For Apple iPhone 14 Pro Max (3,951,206)
- For Apple iPhone 13 Pro Max (3,934,393)
- For Apple iPhone 14 (3,883,230)
- For Apple iPhone 13 Pro (3,874,039)
- For Apple iPhone 12 Pro Max (3,868,639)
- For Apple iPhone 14 Pro (3,802,709)
- For Apple iPhone 12 (3,684,734)

see all

Brand

- Unbranded (4,840,340)
- Apple (313,402)
- Samsung (82,575)

170,000+ results for **iphone 15** Save this search Shipping to: 30349

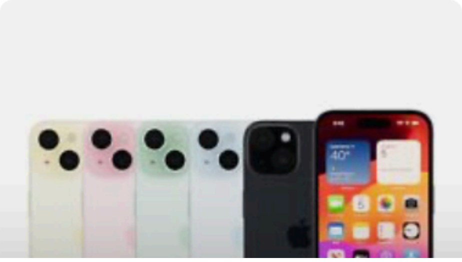
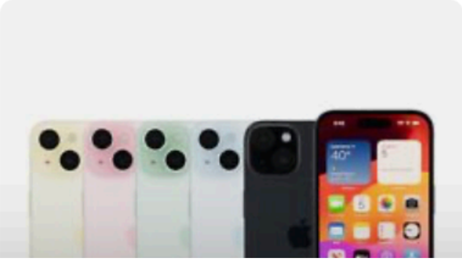
All | Auction | Buy It Now | Condition | Shipping | Local | Sort: Best Match

iPhone 15 Charger OEM Original Genuine Apple iPhone 15 Charger...
Brand New
$4.87 to **$10.87**
Was: ~~$5.60~~ 13% off
Buy It Now
Free shipping
Free returns
Buy 1, get 1 20% off with coupon
Sponsored

Apple - iPhone 15 PRO MAX - 256gb - Unlocked - Factory Warranty - All...
Brand New
★★★★★ (7)
$1,194.99
Buy It Now
Free shipping
Free returns
116+ sold

Apple iPhone 15 128GB Factory Unlocked AT&T T-Mobile Verizon...
Very Good - Refurbished
$664.95
Buy It Now
Free 3 day shipping
Free returns
$5 off 2+ with coupon

Apple iPhone 15 5G - 128GB - All Colors - Factory Unlocked -E SiM -...
Excellent - Refurbished
$749.99 to **$759.99**
Buy It Now
Free 4 day shipping
Free returns
eBay Refurbished

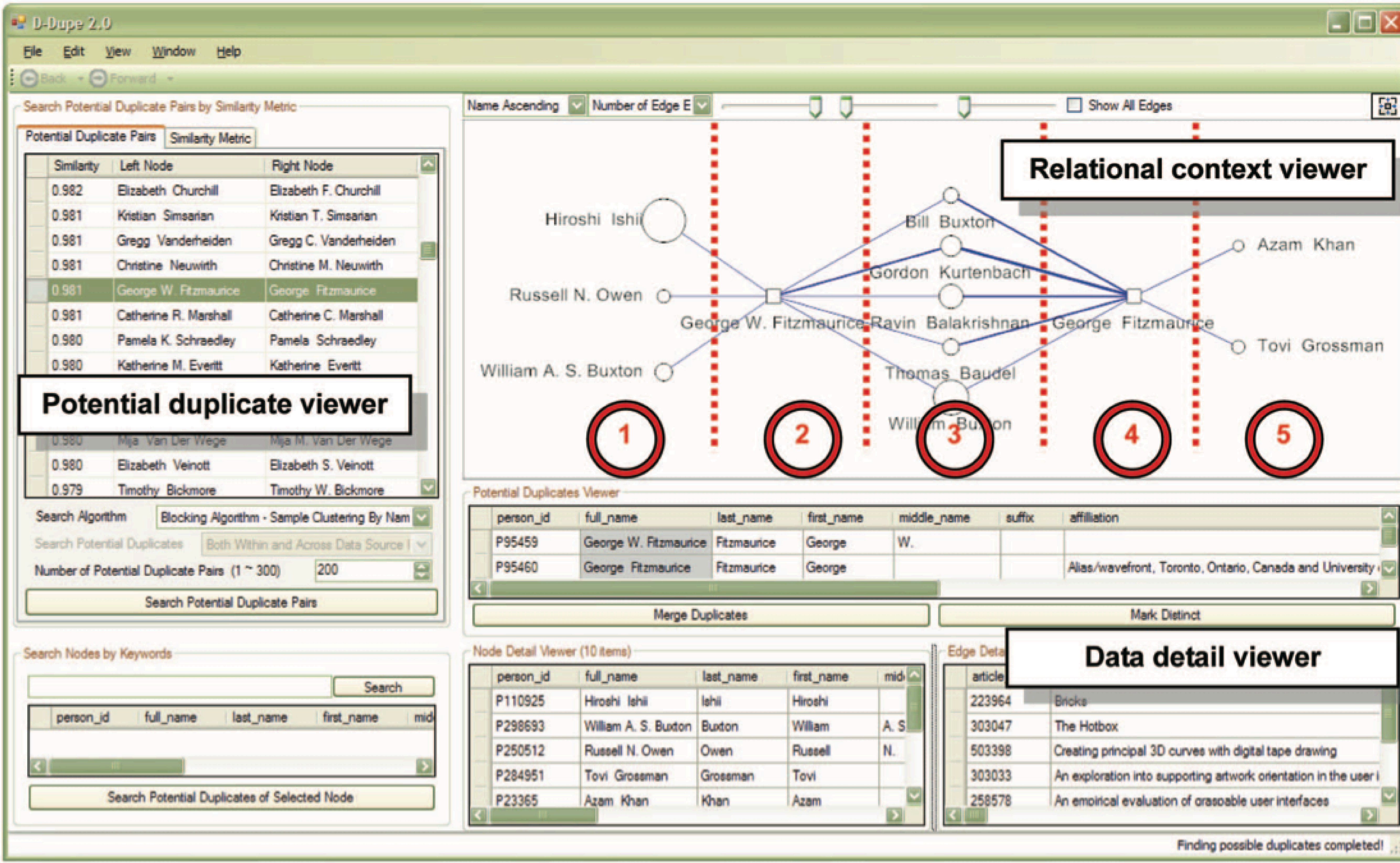



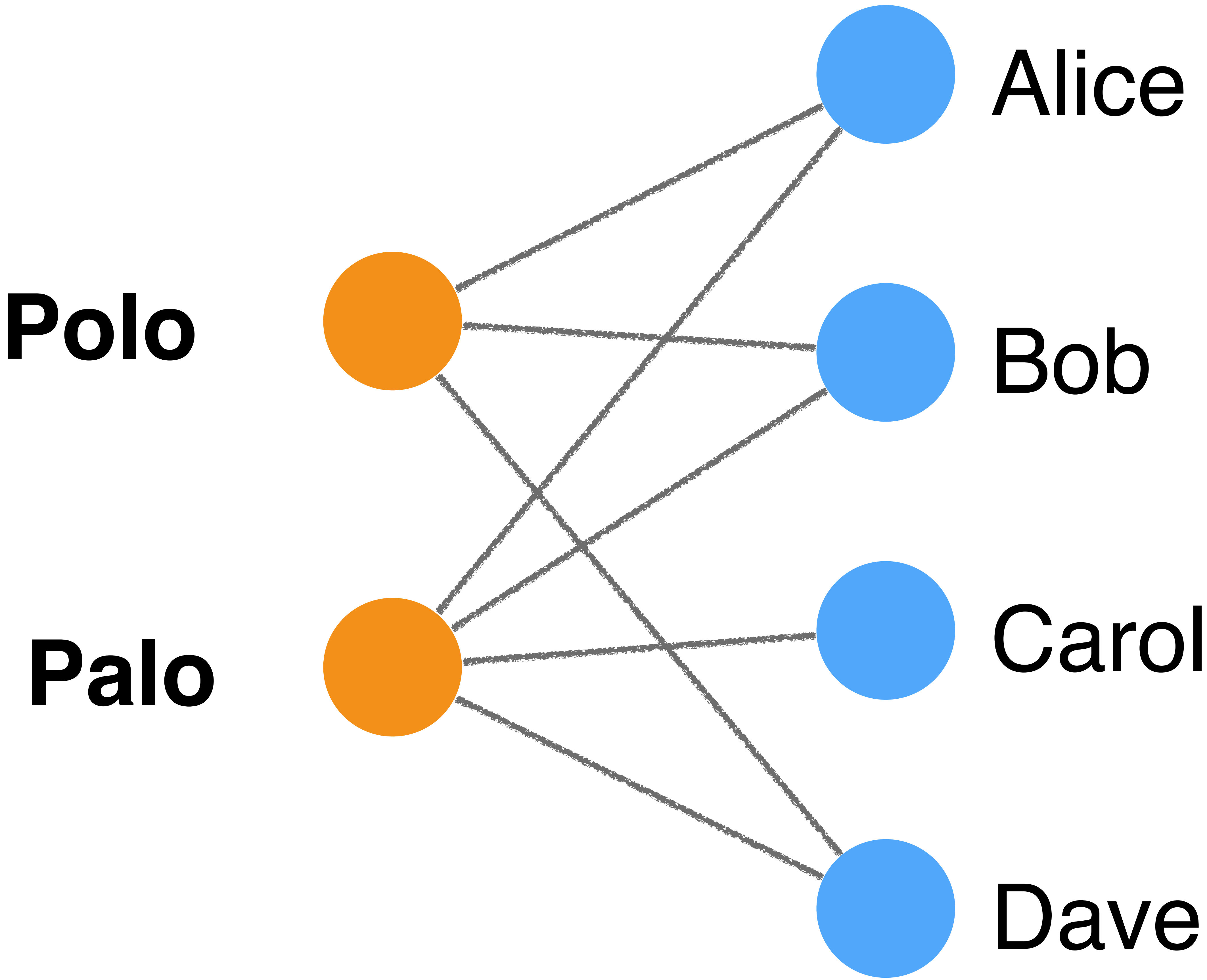
D-Dupe

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/


D-Dupe 2.0
File Edit View Window Help
Search Potential Duplicate Pairs by Similarity Metric
Potential Duplicate Pairs
Similarity Metric
Potential duplicate viewer
Relational context viewer
Data detail viewer
Hiroshi Ishii
Russell N. Owen
William A. S. Buxton
George W. Fitzmaurice
Bill Buxton
Gordon Kurtenbach
Ravin Balakrishnan
Thomas Baudel
William Buxton
George Fitzmaurice
Azam Khan
Tovi Grossman
1
2
3
4
5
Show All Edges
Potential Duplicates Viewer
Merge Duplicates
Mark Distinct
Search Nodes by Keywords
Search
Search Potential Duplicates of Selected Node
Search Potential Duplicate Pairs
Node Detail Viewer (10 items)
Finding possible duplicates completed!


Polo
Palo
Alice
Bob
Carol
Dave

Core components: **Similarity functions**

Determine how two entities are similar.

D-Dupe's approach:

Attribute similarity + **relational similarity**

$$sim(e_i, e_j) = (1 - \alpha) \times sim_A(e_i, e_j) + \alpha \times sim_R(e_i, e_j),$$
$$0 \leq \alpha \leq 1,$$

Similarity score for a pair of entities

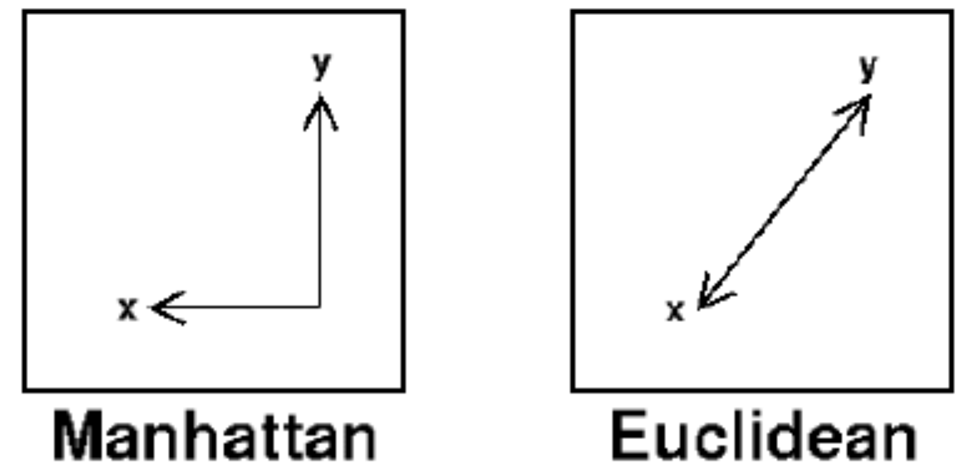
Attribute similarity (a weighted sum)

$$sim_A(e_i, e_j) = \sum_{k=1}^{n} w_k \times sim\_fun_k(e_i \cdot a_k, e_j \cdot a_k),$$

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

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

Jaccard similarity of sets S and T is $|S \cap T|/|S \cup T|$

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

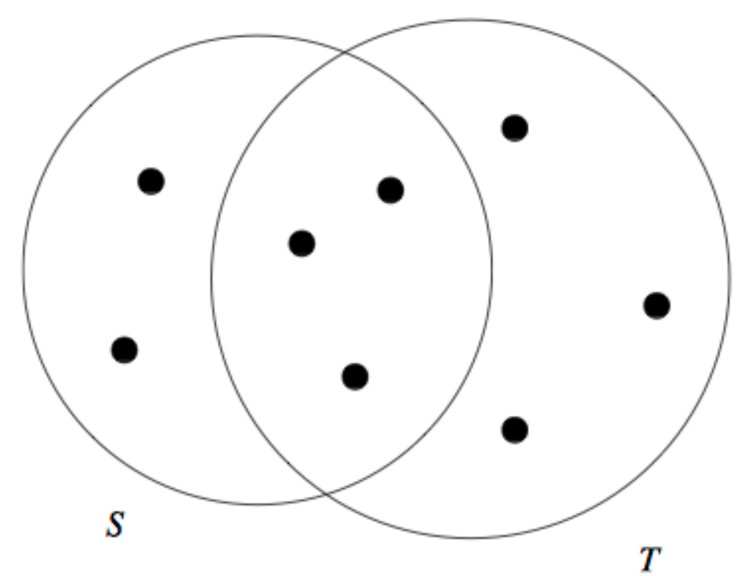


Figure 3.1: Two sets with Jaccard similarity 3/8

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.

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

https://reference.wolfram.com/language/guide/DistanceAndSimilarityMeasures.html

Excellent Tutorial on Entity Resolution

http://www.umiacs.umd.edu/~getoor/Tutorials/ER_KDD2013.pdf

by Lise Getoor and Ashwin Machanavajjhala