

<http://poloclub.gatech.edu/cse6242>

CSE6242/CX4242: **Data** & **Visual** Analytics

Data Integration

Max Mahdi Roozbahani

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.

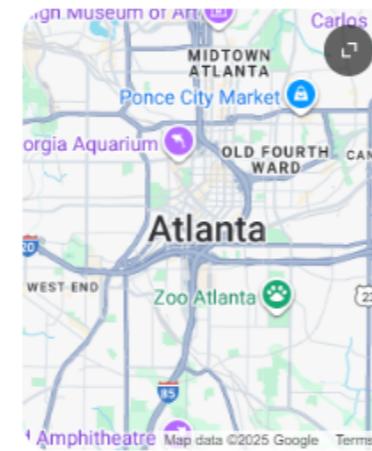
Why is it Important?

Think about the apps, websites, and services that you use every day.

**Businesses derive value
through data integration.**

Atlanta

City in Georgia



Weather

Wed



81°

Thu



82°

Fri



79°

Google Weather

Directions

23m

from Sandy Springs



Wikipedia

<https://en.wikipedia.org/wiki/Atlanta>

Atlanta

The capital and most populous city of the US state of Georgia. It is the county seat of Fulton County and extends into neighboring DeKalb County.



About

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

Population: 520,070 (2024)

Area codes: 404/678/770/470/943

(City of Atlanta): December 29, 1847; 177 years ago

Demonym: Atlantan

Elevation: 1,050 ft (320 m)

Founded (Terminus): 1837; 188 years ago

Feedback

People also ask :

What is the city of Atlanta known for?

What does 404 mean in Atlanta?

Is Atlanta a cheap place to live?

Is \$90,000 a good salary in Atlanta?

What do rappers call Atlanta?

What is the abandoned field in Atlanta?

People also search for



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?

Music and Podcasts

Hey Siri, play Charlie

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, what's the ETA?

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

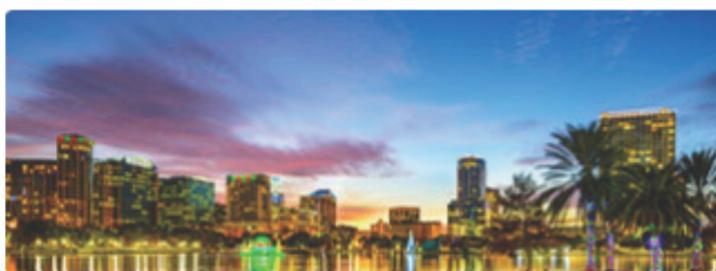


Stay up-to-date

Subscribe now and receive the latest travel news.

[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 Find... Browse Query Help Sign In or Sign Up English

Important! Freebase is read-only and will be shut-down. [More...](#)

3,179,263,202 Facts (and counting)

A community-curated database of well-known people, places, and things

Data Schema Queries Apps Loads Review Tasks Users

Explore Freebase Data

Domain	ID	Topics	Facts
Music	/music	33M	240M
Books	/book	6M	15M
Media	/media_common	6M	17M
People	/people	4M	20M
Film	/film	2M	22M
Location	/location	2M	20M
TV	/tv	2M	19M
Business	/business	1M	4M
Fictional Universes	/fictional_universe	1M	1M
Organization	/organization	996K	4M
Biology	/biology	966K	5M

How can you get started?

Learn how it works
Discover what kind of information Freebase contains, how it's organized, and how Freebase allows you to uniquely identify identities anywhere on the web
[Keep reading »](#)

Use Freebase data
Freebase data is free to use under an [open license](#). You can:

- Query Freebase using our [Search](#), [Topic](#), or [MQL](#) APIs
- [Download](#) our weekly data dumps

Join the Community

- Follow [Freebase on G+](#)

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

Learn more about Freebase at <https://en.wikipedia.org/wiki/Freebase>

Freebase

(a graph of entities)

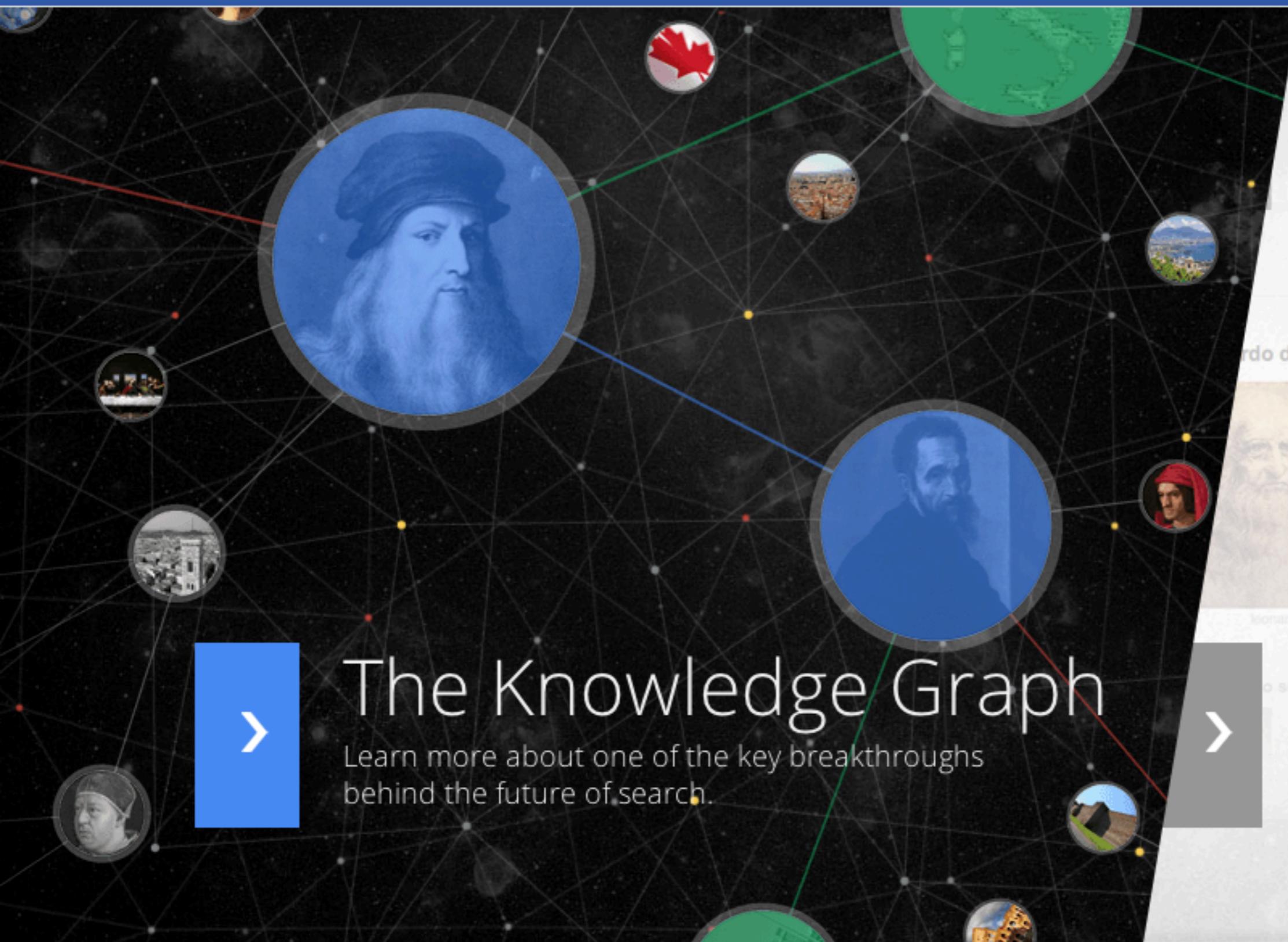
“...a large collaborative knowledge base consisting of metadata composed mainly by its **community members...**”

Wikipedia.

So what?

What can you do with the
Freebase knowledge graph?

Hint: Google acquired it in 2010.



Freebase replaced by Google Knowledge Graph API



Example:
**What does Google know
about Taylor Swift?**

[https://developers.google.com/
knowledge-graph/](https://developers.google.com/knowledge-graph/)

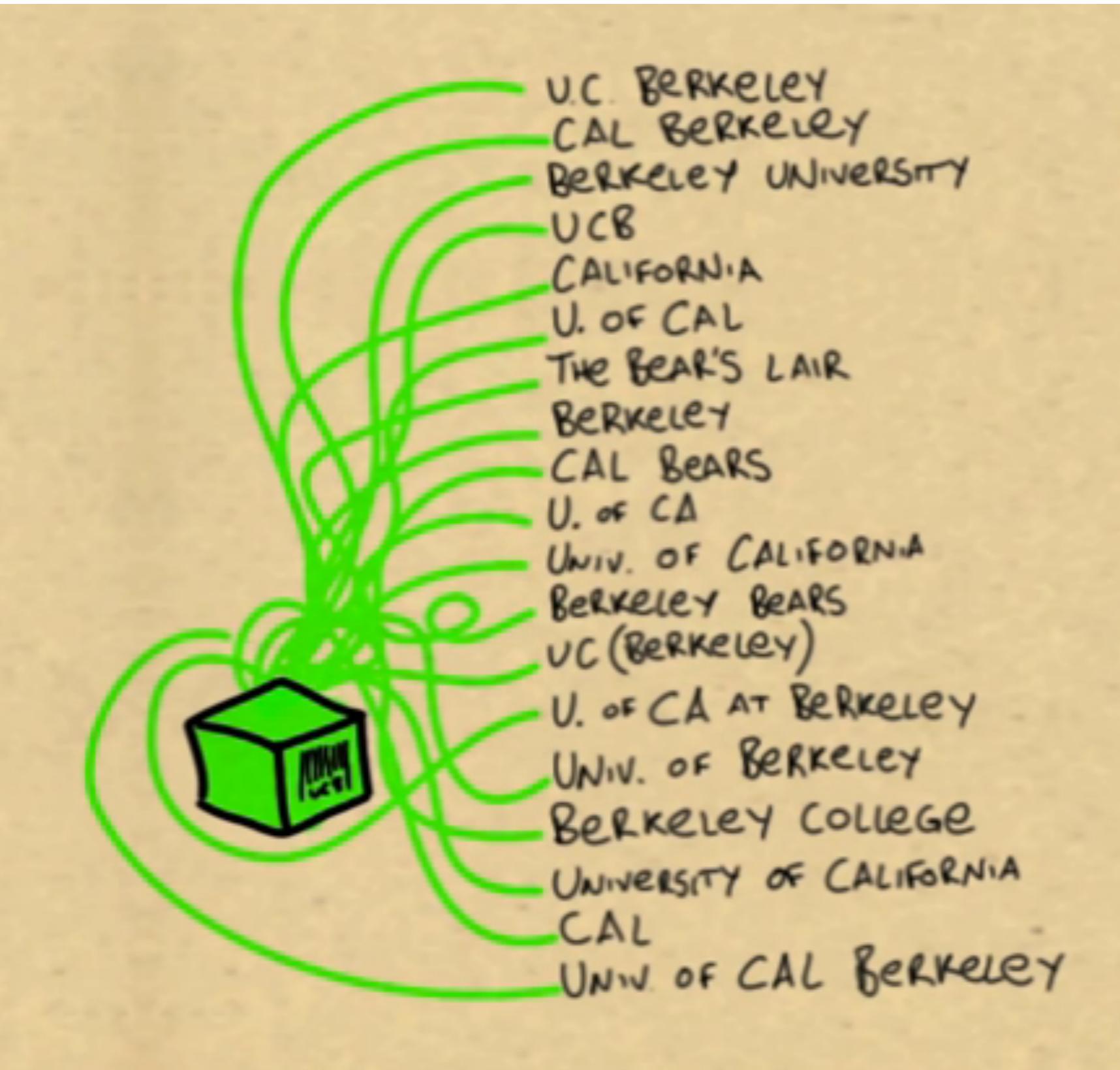


What does Google know about Taylor Swift?

<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:ANd9GcQmVDAhjhWnN20Wys2ZM03PGAh", ←
        "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.⁶



Pay as low as \$0 after
trade-in.⁶

Pay as low as \$0 after
trade-in.⁶⁶

New

Buy iPhone 15

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

Get \$30-\$620 for your trade-in.⁶⁰

Get 3% Daily Cash back with Apple Card.



Model. Which is best for you?

iPhone 15

6.1-inch display¹

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

iPhone 15 Plus

6.7-inch display¹

From \$899
or \$37.45/mo.
for 24 mo.*



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](#) [Q](#) [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)**Custom Products by FSC25**

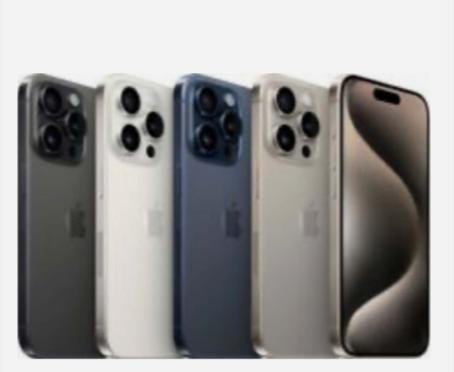
64K items sold

[Shop Store on eBay](#)

Sponsored

170,000+ results for **iphone 15**[Save this search](#)

Shipping to: 30349 v

[Auction](#)[Buy It Now](#)[Condition](#) v[Shipping](#) v[Local](#) v[Sort: Best Match](#) v[grid](#)[iPhone 15 Charger OEM Original](#)
[Genuine Apple iPhone 15 Charger...](#)

Brand New

\$4.87 to \$10.87

Was: \$a.&() 13% off

[Buy It Now](#)

Free shipping

Free returns

Buy 1, get 120% off with coupon

Sponsored

[Apple - iPhone 15 PRO MAX - 256gb -](#)
[Unlocked - Factory Warranty - All...](#)

Brand New

\$1,194.99[Buy It Now](#)

Free shipping

Free returns

116+ sold

[Apple iPhone 15 128GB Factory](#)
[Unlocked AT&T-Mobile Verizon ...](#)

Very Good - Refurbished

\$664.95[Buy It Now](#)**Free 3 day shipping**

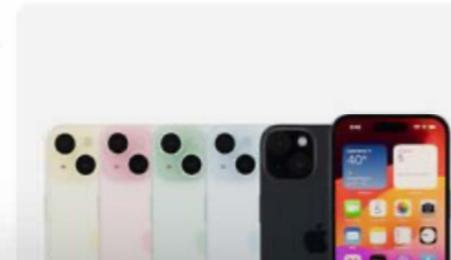
Free returns

\$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

E 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

Potential Duplicate Pairs by Similarity Metric

Similarity	Left Node	Right Node
0.982	Elizabeth Churchill	Elizabeth F. Churchill
0.981	Kristian Simsarian	Kristian T. Simsarian
0.981	Gregg Vanderheiden	Gregg C. Vanderheiden
0.981	Christine Neuwirth	Christine M. Neuwirth
0.981	IGargtW,Fin-ricc	Gaarge
0.981	Catherine R. Marshall	Catherine C. Marshall
0.980	Pamela K. Schraedley	Pamela Schraedley
0.980	Katherine M. Everett	Katherine Everett

Potential duplicate viewer

0.980	Mja Van Der Wege	Mja M. Van Der Wege
0.980	Elizabeth Veinott	Elizabeth S. Veinott
0.979	Timothy Bickmore	Timothy W. Bickmore

Search Algorithm Blocking Algorithm - Sample Clustering By Name

Search Potential Duplicates Both Within and Across Data Source

Number of Potential Duplicate Pairs (1 ~ 300) 200

Search Potential Duplicate Pairs

Relational context viewer

Potential Duplicates Viewer

person_id	full_name	last_name	first_name	middle_name	suffix	affiliation
P95459	George W. Fitzmaurice	Fitzmaurice	George	W.		
P95460	George Fitzmaurice	Fitzmaurice	George			Alias/wavefront, Toronto, Ontario, Canada and University

Merge Duplicates **Mark Distinct**

Node Detail Viewer (10 items)

person_id	full_name	last_name	first_name	middle_name
P110925	Hiroshi Ishii	Ishii	Hiroshi	
P298693	William A. S. Buxton	Buxton	William	A. S.
P250512	Russell N. Owen	Owen	Russell	N.
P284951	Tovi Grossman	Grossman	Tovi	
P23365	Azam Khan	Khan	Azam	

Edge Data

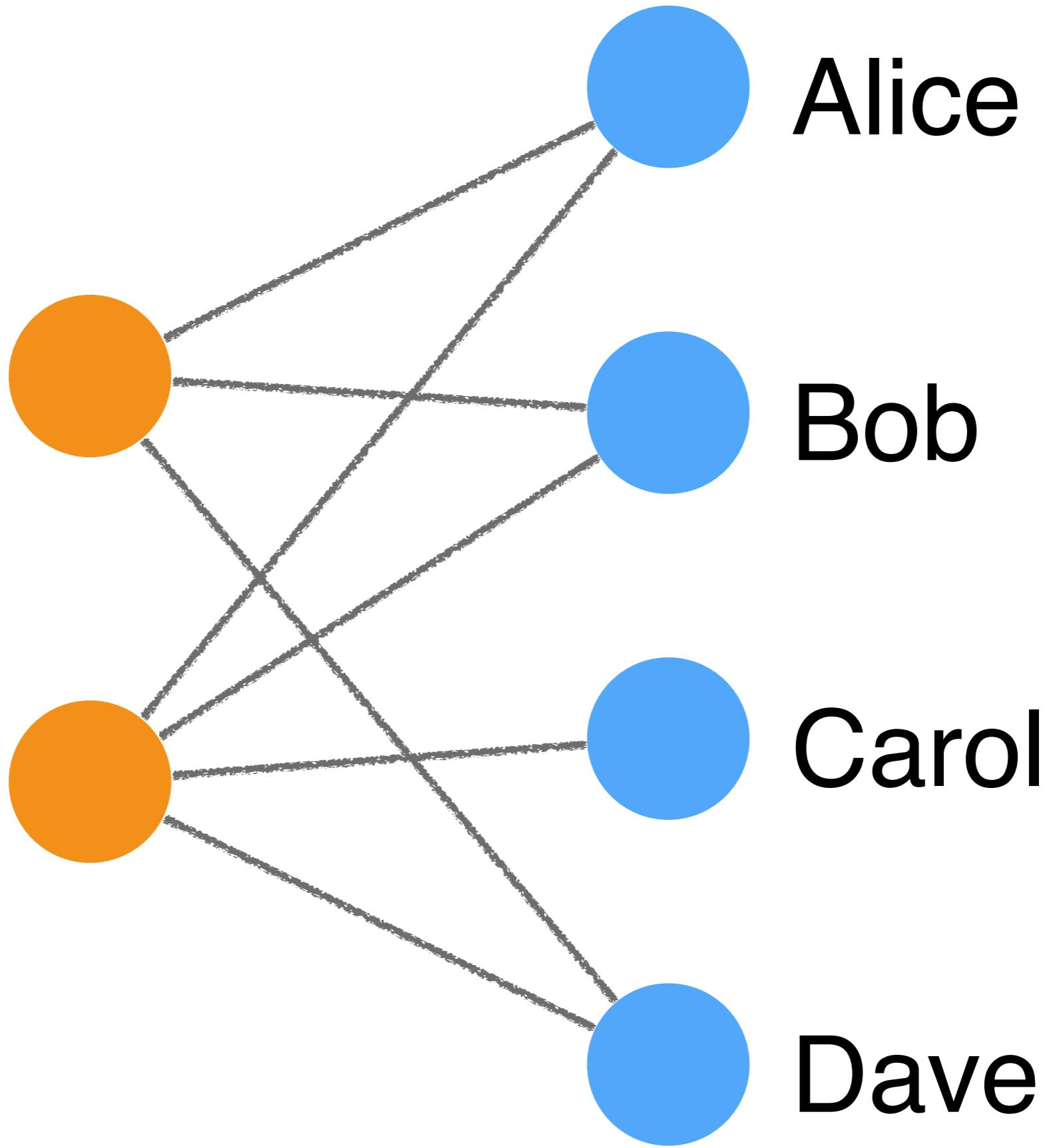
article	
223964	Bricks
303047	The Hotbox
503398	Creating principal 3D curves with digital tape drawing
303033	An exploration into supporting artwork orientation in the user
258578	An empirical evaluation of erasable user interfaces

Data detail viewer

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)



$$sim_A(e_i, e_j) = \sum_{k=1}^n w_k \times sim_funk(e_i \cdot a_k, e_j \cdot a_k),$$
$$-1 \leq w_k \leq 1 \quad \text{and} \quad \sum_{k=1}^n |w_k| = 1,$$

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”

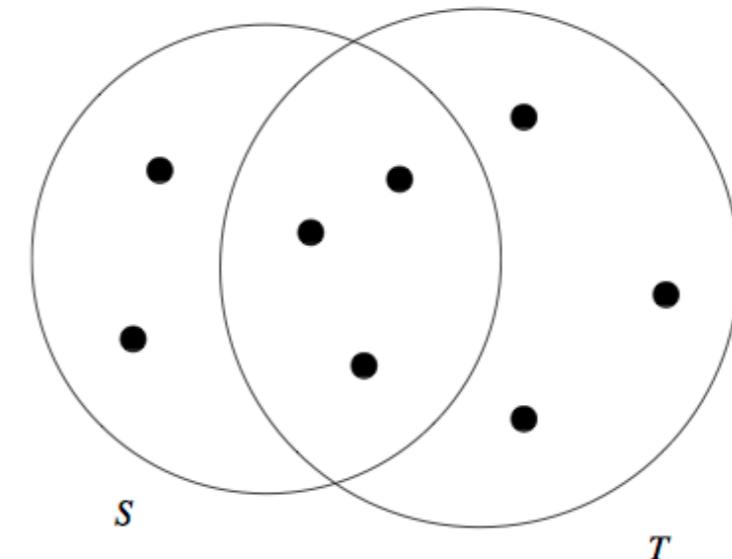
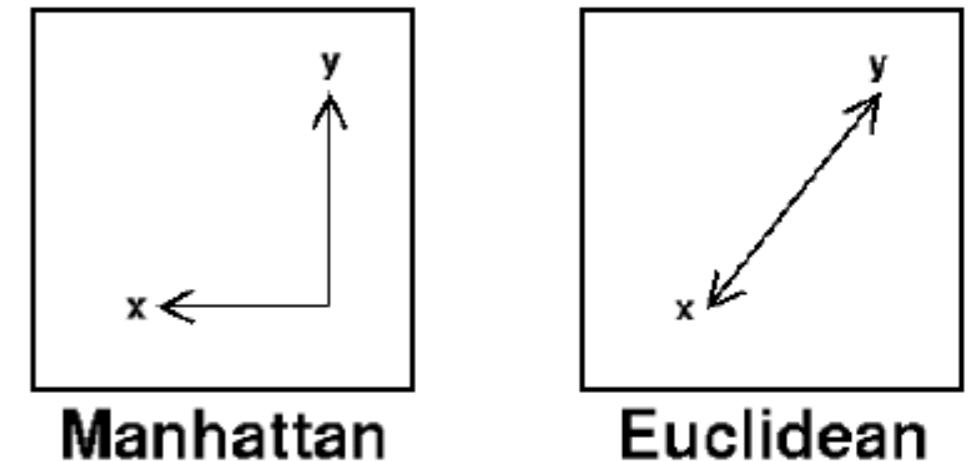


Figure 3.1: Two sets with Jaccard similarity 3/8

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](#)

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

Geospatial & Temporal Data

[GeoDistance](#) • [DateDifference](#)

Excellent Tutorial on Entity Resolution

[http://www.umiacs.umd.edu/~getoor/Tutorials/
ER_KDD2013.pdf](http://www.umiacs.umd.edu/~getoor/Tutorials/ER_KDD2013.pdf)

by Lise Getoor and Ashwin Machanavajjhala