http://poloclub.gatech.edu/cse6242

CSE6242 / CX4242: Data & Visual Analytics

Data Integration

Duen Horng (Polo) Chau

Assistant Professor

Associate Director, MS Analytics

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.

2 personal results. 106,000,000 other results.

City of **Atlanta**, GA : Home
www.**atlanta**ga.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.: /ətˈlæntə/, stressed /ætˈlæntə/, locally /ætˈlænə/) 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 ...

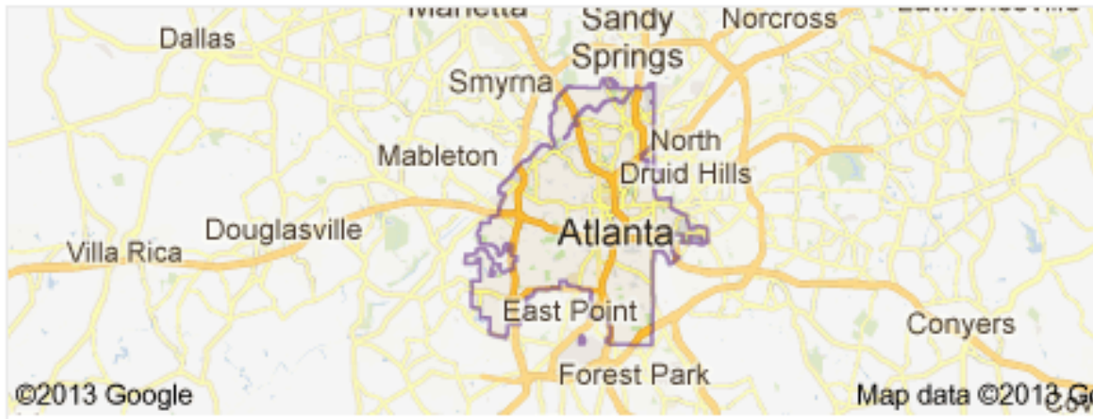
50 Fun Things to Do in **Atlanta** - **Atlanta** Convention and Visitor's ...
www.**atlanta**.net/50fun/
Check out our guide to the top 50 Fun Things to Do in **Atlanta** by activity or neighborhood. The **Atlanta** Convention & Visitors Bureau is your guide to finding fun ...

Things to do in **Atlanta** | www.accessatlanta.com
www.access**atlanta**.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

Atlanta is the capital of and the most populous city in the U.S. state of Georgia, with an estimated 2011 population of 432,427. Wikipedia

Population: 432,427 (2011) United States Census Bureau

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

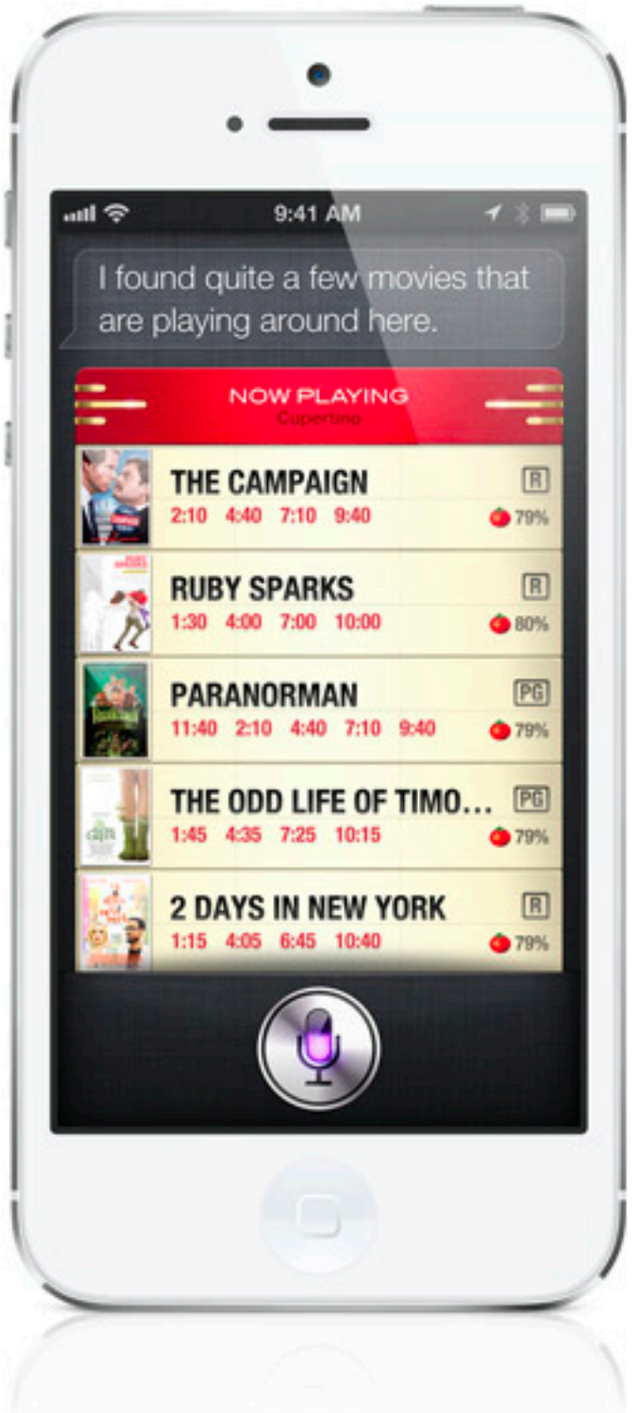
Upcoming events

Jan 17 Thu	Blue Man Group Fox Theatre Atlanta
Jan 17 Thu	Purity Ring at Variety Playhouse on Jan 17, 2013 Variety Playhouse
Jan 18 Fri	Ellie Goulding w/ St. Lucia The Tabernacle

Points of interest

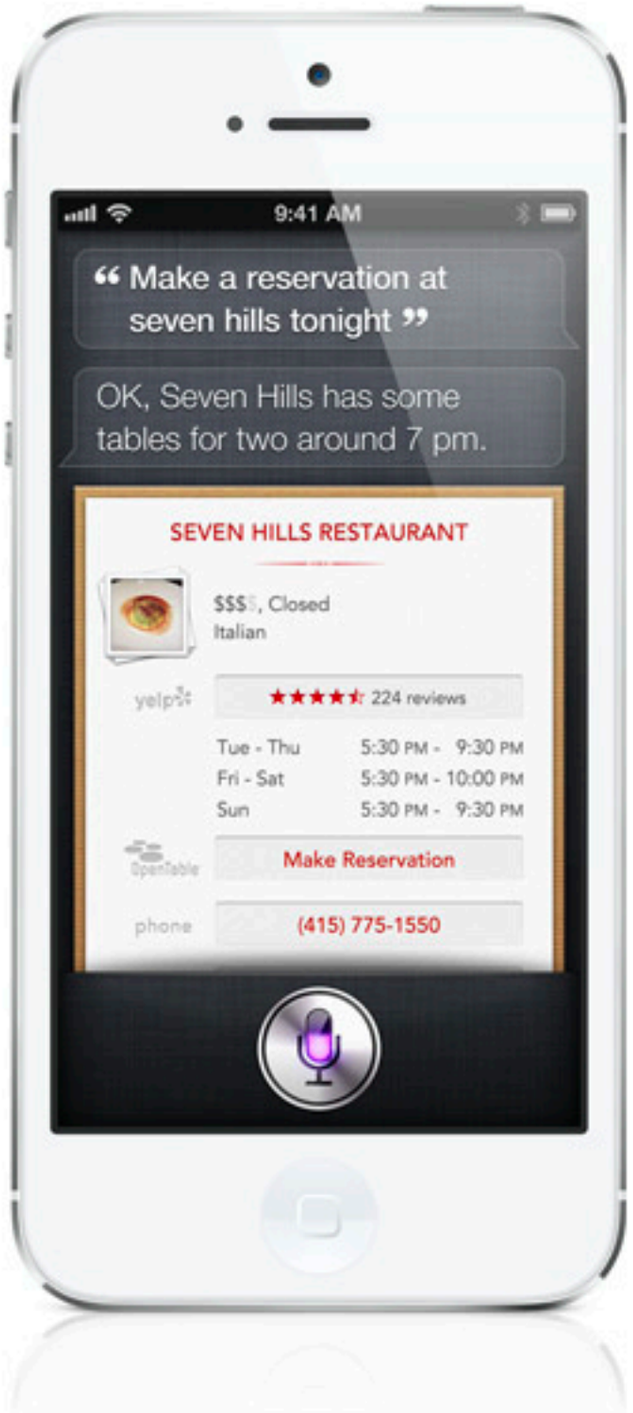
Know the score.

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.

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 | Wed 9/13 – Fri 9/22 | 1 adult, Economy

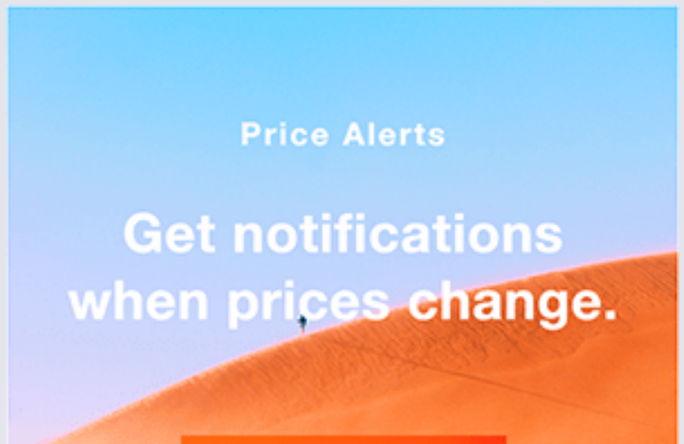
Compare vs. KAYAK all | none ☐ Priceline ☐ Orbitz ☐ Hotwire ☐ Travelocity ☐ JustFly

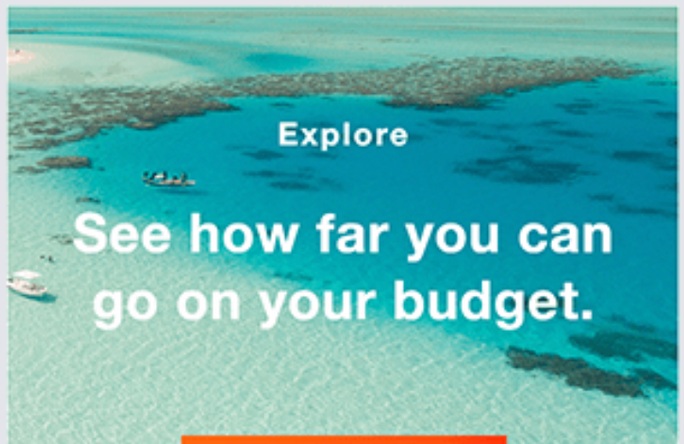
☐ I am a student

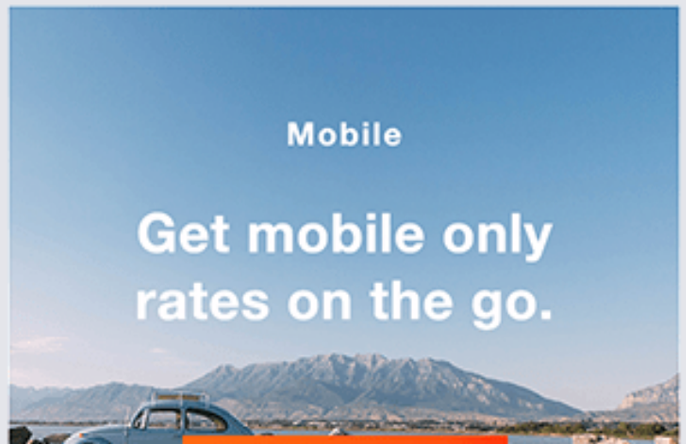
Stay up-to-date

Subscribe now and receive the latest travel news.

Your email address SIGN UP







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	Obama

id	state
111	GA
222	NY
333	CA

id	name	state
111	Smith	GA
222	Johnson	NY
333	Obama	CA

2. **Open Refine**

http://openrefine.org (video #3)

So **IDs** are really important!

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: https://www.youtube.com/watch?v=TJfrNo3Z-DU

Freebase to moved to Wikidata in July (2015): http://goo.gl/3ZDTg7

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.

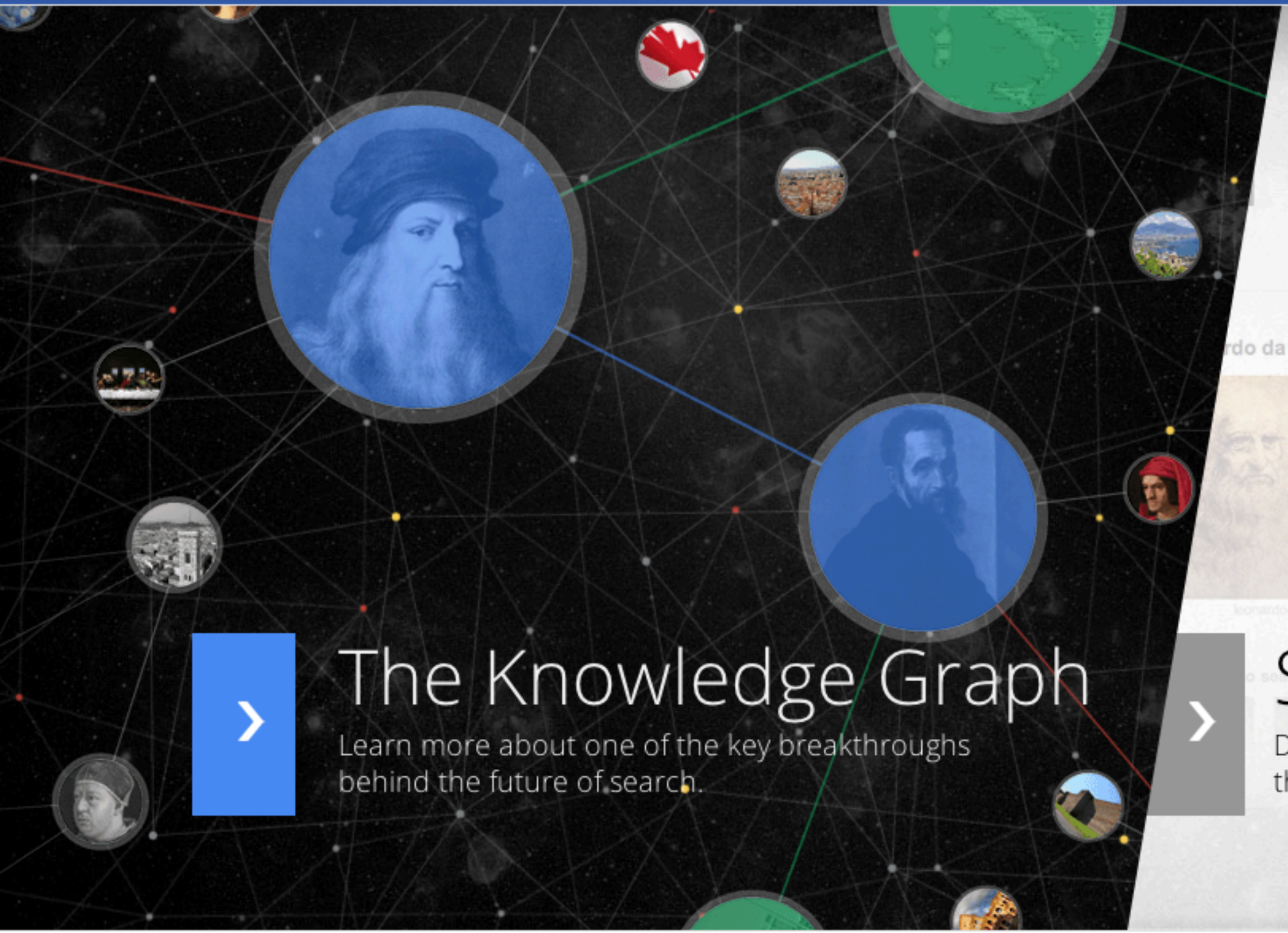
So what?

What can you do with Freebase?

Hint: Google acquired it in 2010

Google | Inside Search

Home Tips & Tricks **Features** Search Stories Playground Blog Help



https://www.google.com/intl/bn/insidesearch/features/search/knowledge.html

Freebase replaced by
Google Knowledge Graph API

Example:
What does Google know about Taylor Swift?

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

Google has the Knowledge Graph.

Facebook has…

facebook

Email or Phone

Password

Log In

Keep me logged in

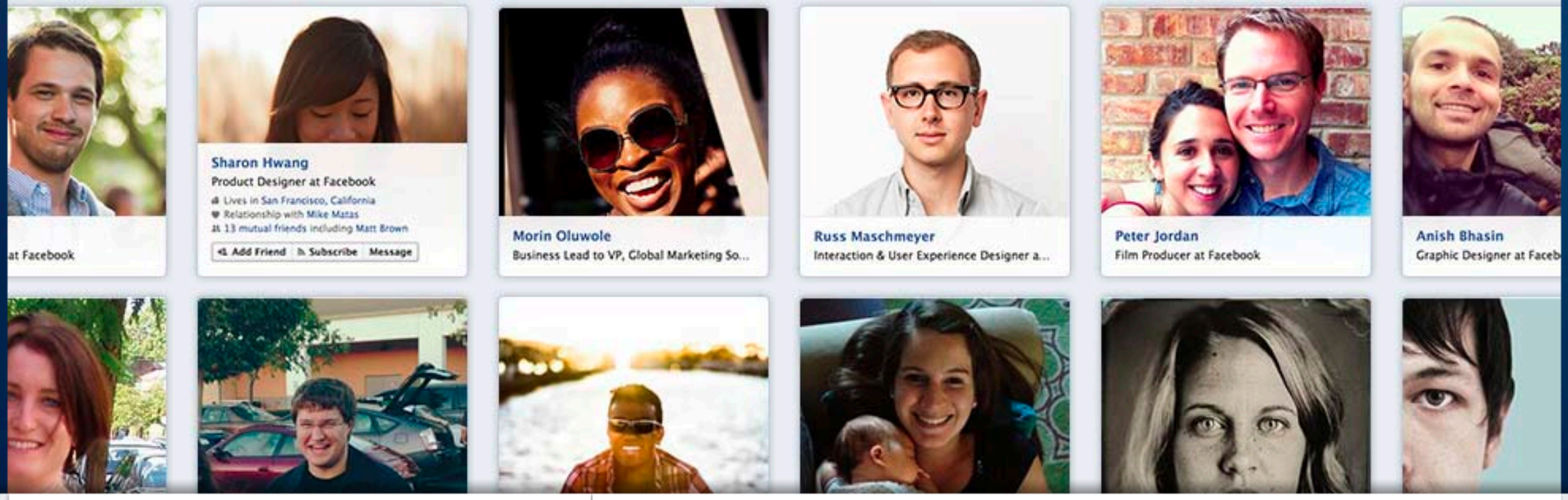
Forgot your password?

Sign Up

Connect and share with the people in your life.

Introducing Graph Search

People who like **Cycling** and are from my hometown



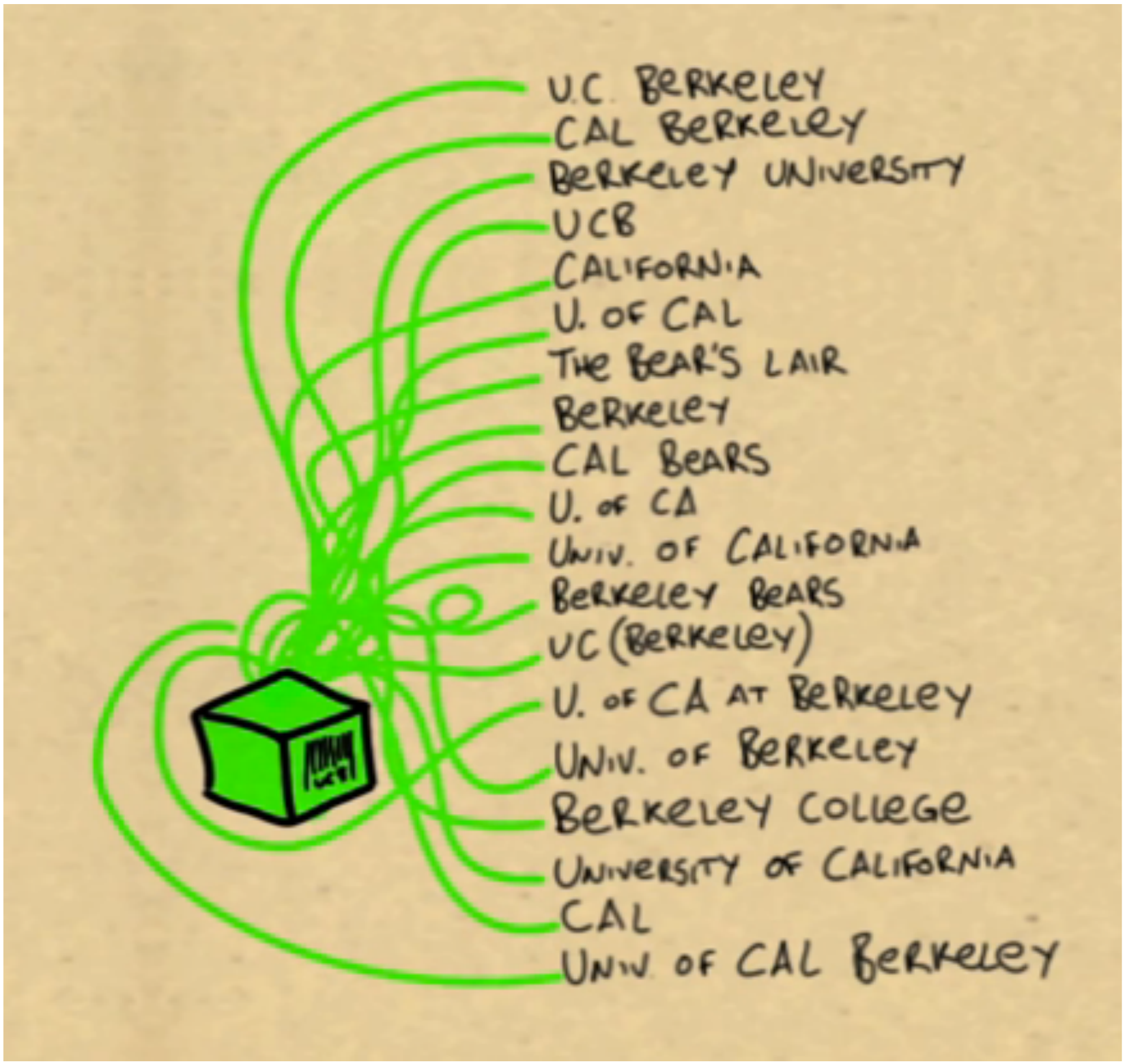
Find people who share your interests

Want to start a book club or find a gym buddy? Connect with friends who like the same activities—and meet new people, too.

https://www.youtube.com/watch?v=W3k1USQbq80

https://www.facebook.com/about/graphsearch

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 6 on Apple, Amazon and eBay

iPhone 7 (PRODUCT)RED™ Special Edition is available now. Buy ›

Model Carrier Finish Capacity

Choose an iPhone 7 model.

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

iPhone 7
4.7-inch display

Select

From $649

iPhone 7 Plus
5.5-inch display

Select

From $769

Get help buying. Chat now or call 1-800-MY-APPLE.

Results for **Cell Phones & Accessories : "iPhone 7"**

Sort by Relevance

Show results for

- Any Category
- **Cell Phones & Accessories**
  - Unlocked Cell Phones
  - Cell Phone Cases
  - Screen Protectors
  - Carrier Cell Phones
  - Cell Phone Cables
  - Stands
  - Cases, Holsters & Clips
  - Chargers & Power Adapters
  - Armbands
  - Car Chargers
  - See more

Refine by

Amazon Prime

- prime

Eligible for Free Shipping

- Free Shipping by Amazon

Brand

- Apple
- New Tech Junkies
- amFilm
- Supcase
- Caseology
- Winstion
- Sakula
- iKits
- Azusa
- Allytech
- Ted Baker
- CHOETECH
- Cineyo

Sponsored

Apple iPhone 6s Unlocked Cellphone, Gold, 16 GB (Refurbished)

$324.99 prime

1,436

Apple iPhone 7 Factory Unlocked GSM Smartphone - 32GB, Rose Gold (Certified Refurbished)

$519.99 prime

119

Apple iPhone 7 Unlocked Phone 128 GB - US Version (Black)

$739.00

Only 3 left in stock - order soon.

548

Apple iPhone 7 - 32GB - T-Mobile - Black (Certified Refurbished)

$499.99 prime

3

Apple iPhone 7 - 32GB - AT&T Locked - Rose Gold (Certified Refurbished)

$499.99 prime

Only 18 left in stock - order soon.

Apple iPhone 6 16GB Factory Unlocked GSM 4G LTE Smartphone, Silver (Certified Refurbished)

$288.35 prime

OEM Apple iPhone 7 Earpod Wired Headphones with Lightning Connector - White/MMTN2AM/A (Certifie...

$25.95 ~~$44.99~~ prime

iPhone 7 Case, SUPCASE Unicorn Beetle Style Premium Hybrid Protective Clear Bumper Case [Scratch...

$14.99 prime

iPhone 7 128GB AT&T Smartphones

Shop by Category

Cell Phones, Smart Watches & Accessories

- **Cell Phones & Smartphones**
- Smart Watches
- Smart Watch Accessories
- Cell Phone Accessories
- Cell Phone Displays
- Cell Phone Cards & SIM Cards
- Cell Phone & Smartphone Parts
- Vintage Cell Phones
- Cell Phone Wholesale Lots
- Other Cell Phones & Accessories

Model see all

- [x] iPhone 7

Network see all

- [x] AT&T

Storage Capacity see all

- [x] 128GB

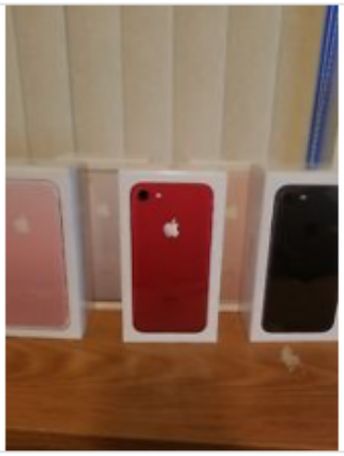
Features see all

- [ ] 3G Data Capable
- [ ] 4G Data Capable
- [ ] 4K Video Recording
- [ ] Bluetooth Enabled
- [ ] Dual SIM
- [ ] Fingerprint Sensor
- [ ] Touchscreen
- [ ] Wi-Fi Capable

All Listings | Auction | Buy It Now

Sort: Best Match | View:

1-48 of 274 Results



Apple iPhone 7 - 32GB and 128GB - (AT&T) Factory Unlocked New - 4 colors

$699.99

FAST 'N FREE

13 watching

Top Rated Plus

Brand: Apple

NEW LISTING **iPhone 7 128GB ATT Black Empty Box w/Original Accessories! Over $80 Value!**

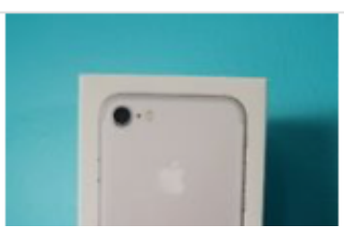
New box with new accessories..new never used..original paper work and stickers

$39.99

or Best Offer

FAST 'N FREE

Brand: Apple



NEW LISTING **Apple iPhone 7 - 128GB - Silver (AT&T) Smartphone**

8 product ratings

$550.00

Brand: Apple

See more like this

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

Back Forward

Search Potential Duplicate Pairs by Similarity Metric

Potential Duplicate Pairs | 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	George W. Fitzmaurice	George Fitzmaurice
0.981	Catherine R. Marshall	Catherine C. Marshall
0.980	Pamela K. Schraedley	Pamela Schraedley
0.980	Katherine M. Everitt	Katherine Everitt
0.980	Mija Van Der Wege	Mija M. Van Der Wege
0.980	Elizabeth Veinott	Elizabeth S. Veinott
0.979	Timothy Bickmore	Timothy W. Bickmore

Potential duplicate viewer

Search Algorithm: Blocking Algorithm - Sample Clustering By Nam

Search Potential Duplicates: Both Within and Across Data Source I

Number of Potential Duplicate Pairs (1 ~ 300): 200

Search Potential Duplicate Pairs

Search Nodes by Keywords

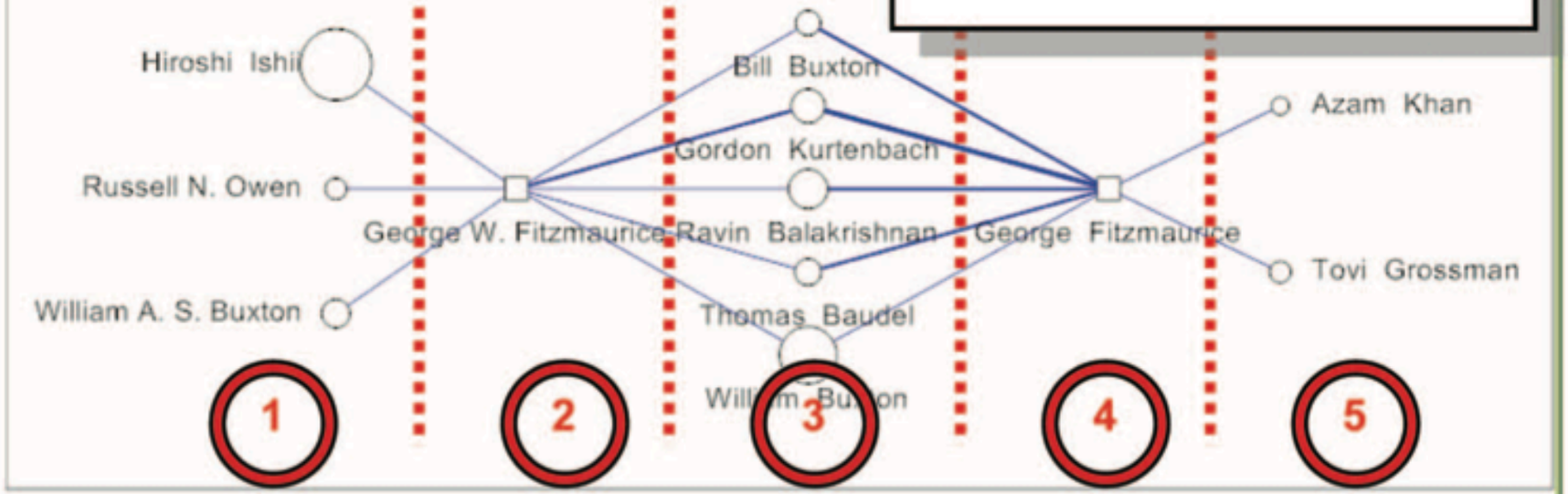
Search

person_id	full_name	last_name	first_name	mid

Search Potential Duplicates of Selected Node

Name Ascending | Number of Edge E | Show All Edges

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)

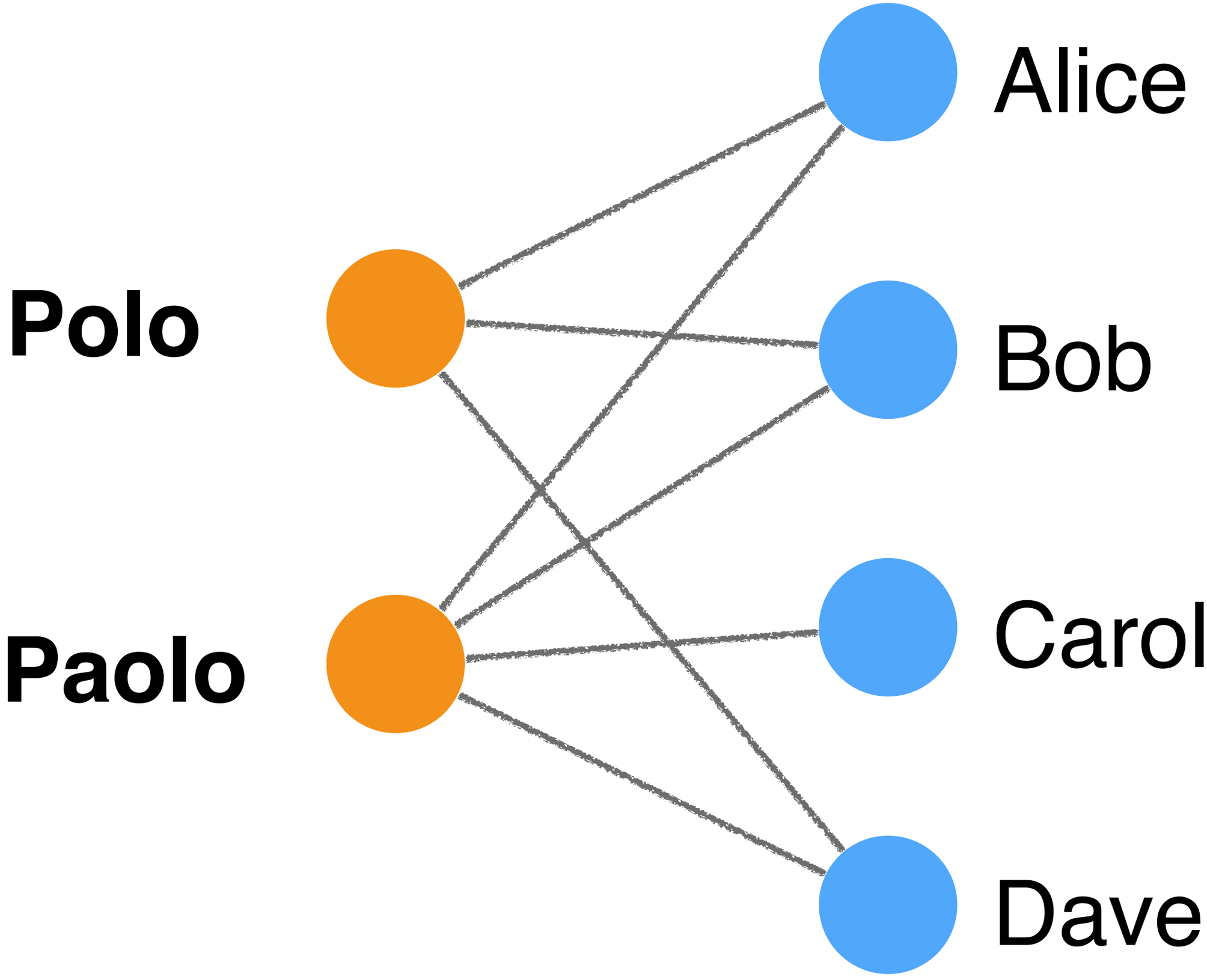
person_id	full_name	last_name	first_name	mid
P110925	Hiroshi Ishii	Ishii	Hiroshi	
P298693	William A. S. Buxton	Buxton	William	A. S
P250512	Russell N. Owen	Owen	Russell	N.
P284951	Tovi Grossman	Grossman	Tovi	
P23365	Azam Khan	Khan	Azam	

Edge Deta

Data detail viewer

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 i
258578	An emoirical evaluation of orasoable user interfaces

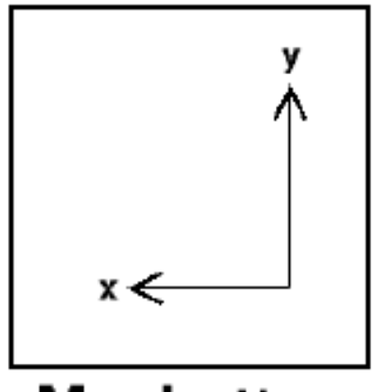
Finding possible duplicates completed!


Polo
Alice
Bob
Paolo
Carol
Dave

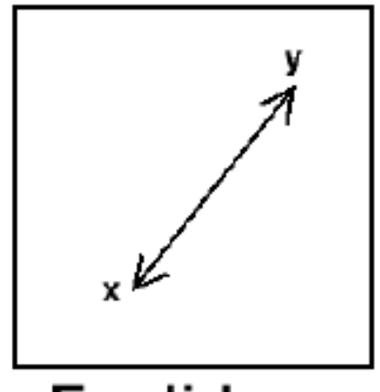
Numerous **similarity functions**

Excellent read: http://infolab.stanford.edu/~ullman/mmds/ch3a.pdf

- Euclidean distance
  Euclidean norm / L2 norm
- TaxiCab/Manhattan distance



Manhattan



Euclidean

- Jaccard Similarity (e.g., used with w-shingles)
  e.g., overlap of nodes' #neighbors

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

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

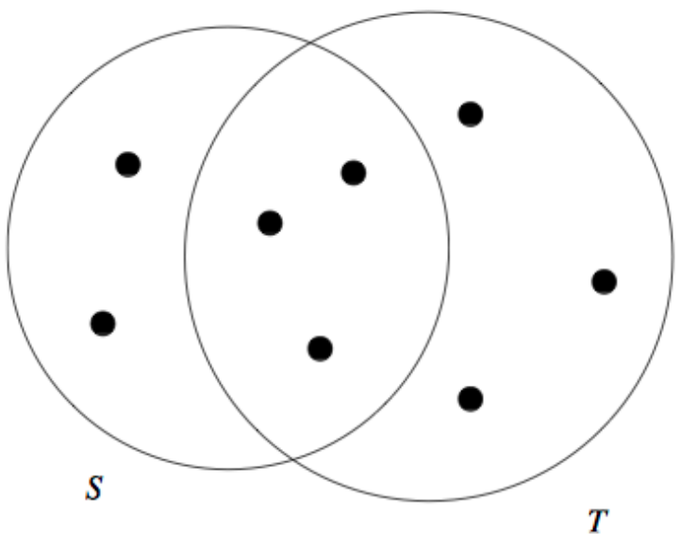


Figure 3.1: Two sets with Jaccard similarity 3/8

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

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)

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

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

Excellent Tutorial on Entity Resolution

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

by Lise Getoor and Ashwin Machanavajjhala