

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

CSE6242: **Data** & **Visual** Analytics

# 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



Sign in

All Maps News Images Videos More Settings Tools

About 467,000,000 results (0.64 seconds)

### Atlanta, GA : Home

<https://www.atlantaga.gov/>

It is a great time to be in the City of Atlanta! Whether you're a native, first time visitor, business traveler who makes regular stops here, or one of the thousands of ...

### Atlanta - Wikipedia

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

Atlanta is the capital of and the most populous city in the U.S. state of Georgia, with an estimated 2016 population of 472,522. Atlanta is the cultural and ...

[Atlanta \(disambiguation\)](#) · [Hartsfield–Jackson Atlanta](#) · [Demographics of Atlanta](#)

### Top stories



Crash kills brothers returning to Georgia Southern after Thanksgiving br...

AJC.com

18 hours ago



Postal worker shot in the head, killed outside Atlanta

Fox News

20 hours ago



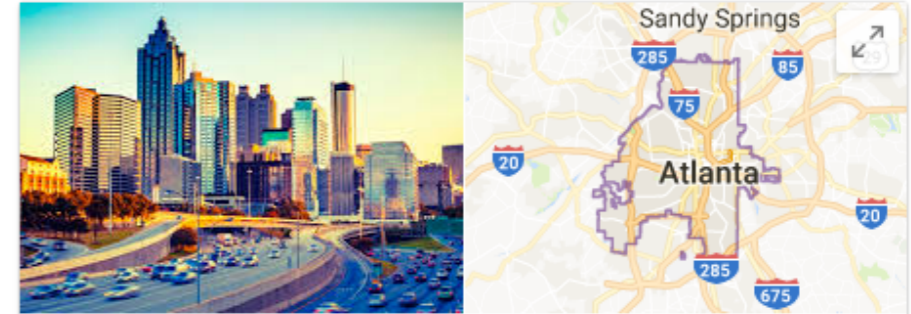
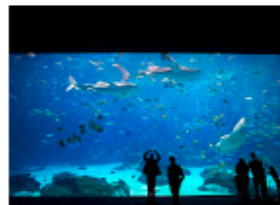
New FOX 5 poll: Virtual dead heat in Atlanta mayor's race

FOX 5 Atlanta

9 hours ago

→ More for atlanta

### Things to do in Atlanta



## 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.

Weather: 51°F (11°C), Wind NE at 3 mph (5 km/h), 96% Humidity

Local time: Wednesday 6:14 AM

Population: 472,522 (2016)

Area code: 404

### Plan a trip

Atlanta travel guide

3-star hotel averaging \$128, 5-star averaging \$366

Upcoming Events

Colleges and Universities: [Emory University](#), [MORE](#)

Did you know: Atlanta has the world's fourth-busiest city airport system by passenger traffic (101,491,106 total passengers). [wikipedia.org](https://www.wikipedia.org)

### People also search for

View 15+ more



# Apple Siri

Getting Answers



"Do I need an umbrella today?"



"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.



"Was that an earthquake?"

Search hundreds of travel sites at once.

HOTELS

FLIGHTS

CARS

PACKAGES

ROUND-TRIP

ONE-WAY

MULTI-CITY

EXPLORE

Atlanta (ATL)

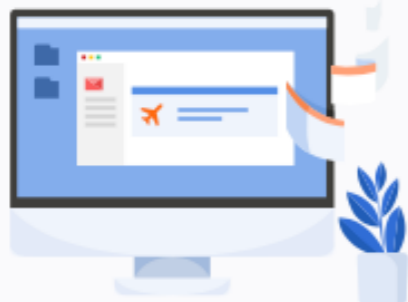


San Francisco (SFO)



Depart - Return

1 adult, Economy



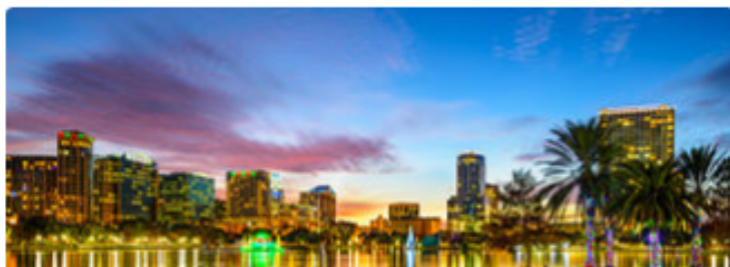
Stay up-to-date

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

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.



# The Knowledge Graph

Learn more about one of the key breakthroughs behind the future of search.



## See it in

Discover answers to questions you thought to ask, and explore



Ginevra de' Benci  
1478



The Virgin Mary  
1508

### Leonardo da Vinci



Leonardo di ser Piero da Vinci  
Renaissance polymath: painter, architect, musician, scientist, engineer, inventor, anatomist, geologist, cartographer, botanist, and writer

Born: April 15, 1452, Anchiano

Died: May 2, 1519, Clos Lucé

Buried: Château d'Amboise

Parents: Caterina da Vinci, Piero da Vinci

Structures: Vebjem Sand Dunes

# 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/)



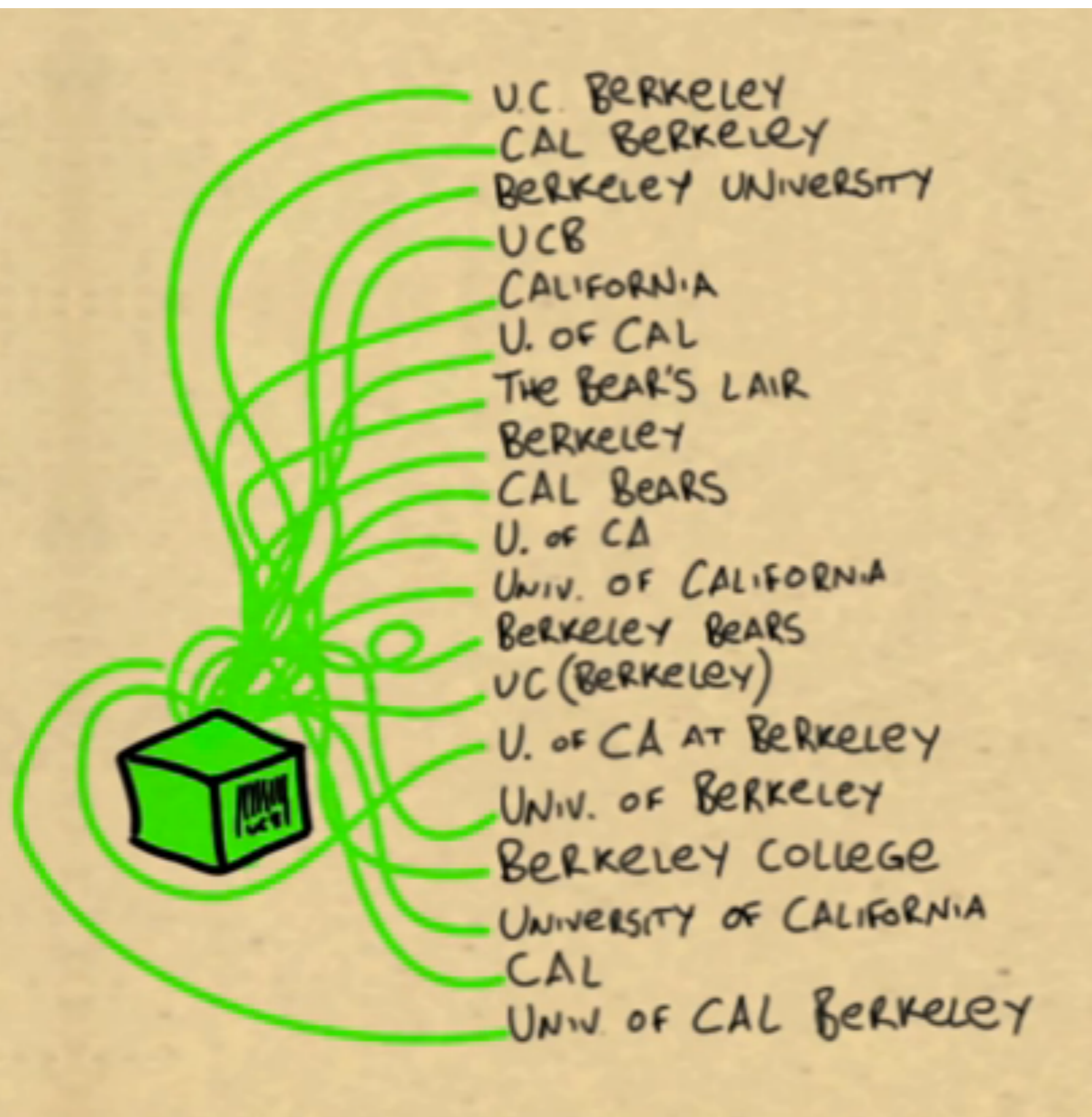
# What does Google know about Taylor Swift?

<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:ANd9GcQmVDAhjhWnN2OWys2ZM03PGAhu",
        "url": "https://en.wikipedia.org/wiki/Taylor_Swift",
        "license": "http://creativecommons.org/licenses/by-sa/4.0/"
      },
      "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 ?



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**

[Mac](#)[iPad](#)[iPhone](#)[Watch](#)[TV](#)[Music](#)[Support](#)

# iPhone X

[Overview](#)[iOS](#)[Tech Specs](#)

If you're part of the iPhone Upgrade Program, you may be eligible for a new iPhone. [Find out now >](#)

[Model](#)[Carrier](#)[Finish](#)[Capacity](#)

## 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.<sup>1</sup>  
Or pay in full from \$999.

Show results for

- Any Category
- Cell Phones & Accessories
- Cell Phones
  - Unlocked Cell Phones
  - Carrier Cell Phones

Refine by

- Delivery Day**
- Get It by Tomorrow
- Amazon Prime**
- prime
- Eligible for Free Shipping**
- Free Shipping by Amazon
- Brand**
- Apple
  - ASA
  - GooPhone
  - iHarbort
  - EcoSale777
  - CRAFIC
  - AGRIGLE
  - TechCode
  - without brand
- See more

Cell Phone Display Size

- 3.9 Inches & Under
- 4.0 to 4.4 Inches
- 4.5 to 4.9 Inches
- 5.0 to 5.4 Inches
- 5.5 Inches & Over

Phone Color



Cell Phone Internal Storage Memory

- Under 4 GB
- 4 GB
- 8 GB
- 16 GB
- 32 GB
- 64 GB
- 128 GB
- 256 GB & Above

Cell Phone Features

Showing most relevant results. See all results for iPhone x.

**SPONSORED BY ZIZO**  
**Making a difference in protecting your phone**  
[Shop now >](#)

Sponsored ⓘ  
**RETINA iPhone Wide Angle Lens**  
 0.6 x | 60% Wider Picture With Every Snap | Bundle With 10x Macro Lens Pro | Clip-On Cell...  
**\$34<sup>99</sup>** ✓prime  
 ★★★★★ 7

Sponsored ⓘ  
**SPEATE iPhone Charger Cable**  
 Nylon Braided Lightning to USB Cord with 3PACK 3FT 6FT 10FT Fast Syncing and Charging for...  
**\$11<sup>99</sup>** ✓prime  
 ★★★★★ 113

Sponsored ⓘ  
**Lightning Cable, NiocTech Nylon Braided USB A to Lightning**  
 Compatible Cable for iPhone X / 8 / 8 Plus / 7 / 7 Plus / 6 / 6...  
**\$9<sup>99</sup>** ✓prime  
 ★★★★★ 30

**Apple iPhone X, GSM Unlocked 5.8", 256 GB - Space Gray**  
 More Choices from \$1,260.00  
 ★★★★★ 69  
 Price may vary by color

**Apple iPhone X, Fully Unlocked 5.8", 64 GB - Silver**  
**\$1,148<sup>99</sup>**  
 ★★★★★ 36  
 Price may vary by color

**Apple iPhone X 256 GB T-Mobile - Space Gray, Locked to T-Mobile**  
**\$1,315<sup>00</sup>**

**New Apple iPhone X, GSM Unlocked 5.8", 256 GB - Silver**  
 More Choices from \$1,150.00

**24K Gold Plated iPhone X 256 GB Silver - Unlocked Custom**  
 More Choices from \$2,199.00

**Apple iPhone X AT&T 64GB (Space Gray) Locked to AT&T**  
**\$1,170<sup>00</sup>** ~~\$1,350.00~~

**iPhone X Case, Crystal Clear Shock Absorption Technology Bumper Transparent TPU+Acrylic Cover vase for...**  
**\$8<sup>99</sup>** ✓prime  
 ★★★★★ 3

**iPhone X Case - Zizo [Static Series] Shoc...** ✓prime ★★★★★ 110  
**iPhone X Case - Zizo [ION Series] with F...** ✓prime ★★★★★ 123  
**iPhone X Case - Zizo [Bolt Series] with S...** ✓prime ★★★★★ 677

Ad feedback

Refine your search for **iphone x**

Include description

Find deals and best selling products for Apple iPhone X Cell Phones & Smartphones

Shop Now →

All Listings Auction Buy It Now

Sort: Best Match

View: [Grid Icon]

Group Similar Listings

1,475 results for **iphone x** Save this search

Guaranteed 3 day delivery

Shop by Model

iPhone X

iPhone 8 Plus

iPhone 8

iPhone 7 Plus

iPhone 7



**Apple iPhone X 64GB - GSM & CDMA Unlocked -USA Model -**  
**\$990.00** Buy It Now  
 Free Shipping  
**914+ Sold**



**Apple iPhone X 256GB - GSM&CDMA Unlocked-USA**  
 ★★★★★  
**\$1,145.00** Buy It Now  
 Free Shipping  
**1523+ Sold**



**Apple iPhone X - 64GB - Space Gray (Factory Unlocked) - Brand**  
 ★★★★★  
**\$1,350.00** or Best Offer  
 Free Shipping  
**256 Sold**



SPONSORED  
**24K Gold Plated Apple iPhone X 256GB - Silver Unlocked Custom**  
**\$1,999.00** or Best Offer  
 Free Shipping  
**15 Watching**



SPONSORED  
**Brand New Apple iPhone X - 256GB - Space Gray (Unlocked)**



**Brand New - Sealed Apple iPhone X 256GB SILVER A1901**  
 ★★★★★



**Apple iPhone X - 256GB - Space Gray AT&T A1901 (GSM)**  
 ★★★★★



**BRAND NEW, Apple iPhone X MQCL2LL/A A1865 64GB Silver**  
 ★★★★★

Categories

- All
- Cell Phones & Accessories
  - Cell Phone Accessories
  - Cell Phones & Smartphones
  - Cell Phone & Smartphone Parts
  - Other Cell Phones & Accs
  - More
- Computers/Tablets & Networking
- Consumer Electronics
- Clothing, Shoes & Accessories
- Business & Industrial
- Sporting Goods
- Show more

Features see all

- 3G Data Capable (134)
- 4G Data Capable (151)
- 4K Video Recording (140)
- Accelerometer (108)
- Bluetooth Enabled (154)
- Dual Rear Cameras (87)
- Fingerprint Sensor (69)
- Wireless Charging (45)

Network see all

- AT&T (263)
- Sprint (105)
- T-Mobile (79)
- Verizon (118)
- Unlocked (554)

Color see all

- Black (85)
- Clear (71)
- Gold (33)
- Gray (696)
- Silver (411)

Storage Capacity see all

- 256GB (627)
- 64GB (553)

Processor see all

- Dual Core (92)
- Hexa Core (986)

Model Number see all

**Save \$200 instantly**  
 on iPhone 7 256 GB with Sprint Flex.

Switch now

\$22.92/mo. for 18 mo. for well-qualified customers with new line activ. For a limited time only. While supplies last.

Browse related



Cables & Adapters for iPhone X



Tool Kits for iPhone X

# 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/>

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	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 Nam

Search Potential Duplicates: Both Within and Across Data Source

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

Search Potential Duplicate Pairs

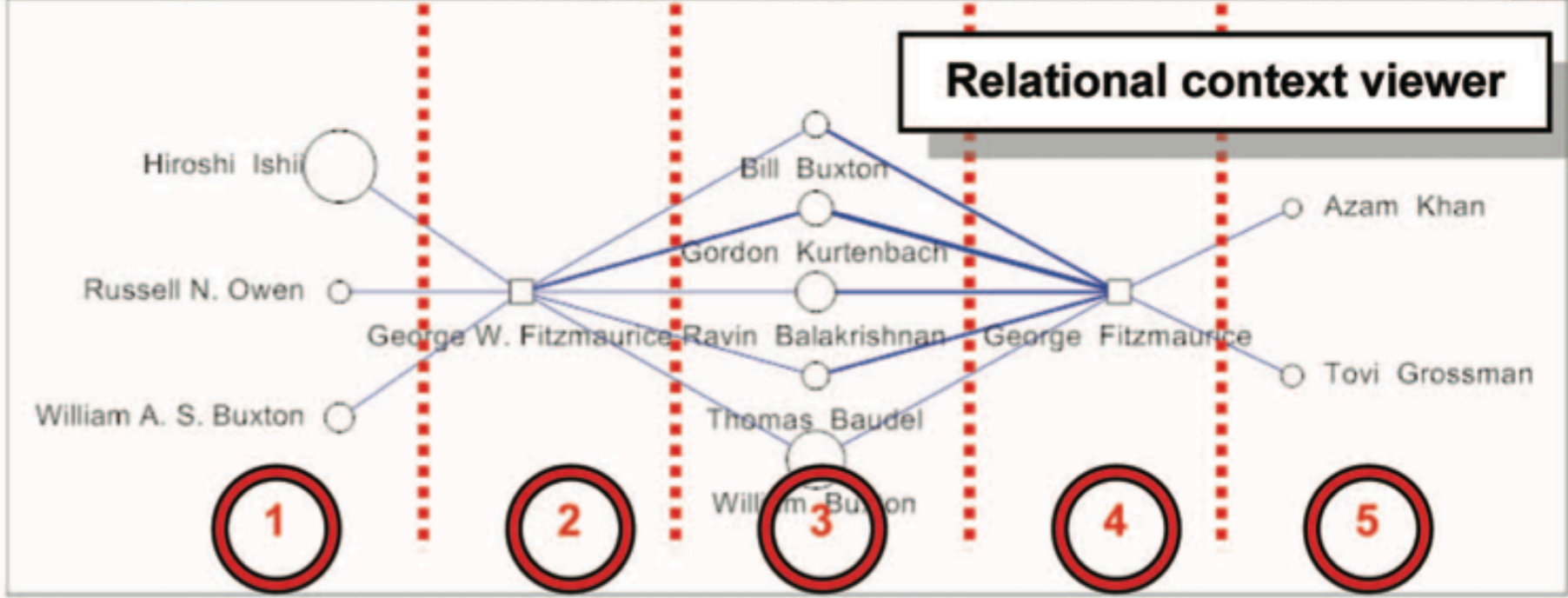
Potential duplicate viewer

Search Nodes by Keywords

Search

person_id	full_name	last_name	first_name	mid

Search Potential Duplicates of Selected Node



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 Data

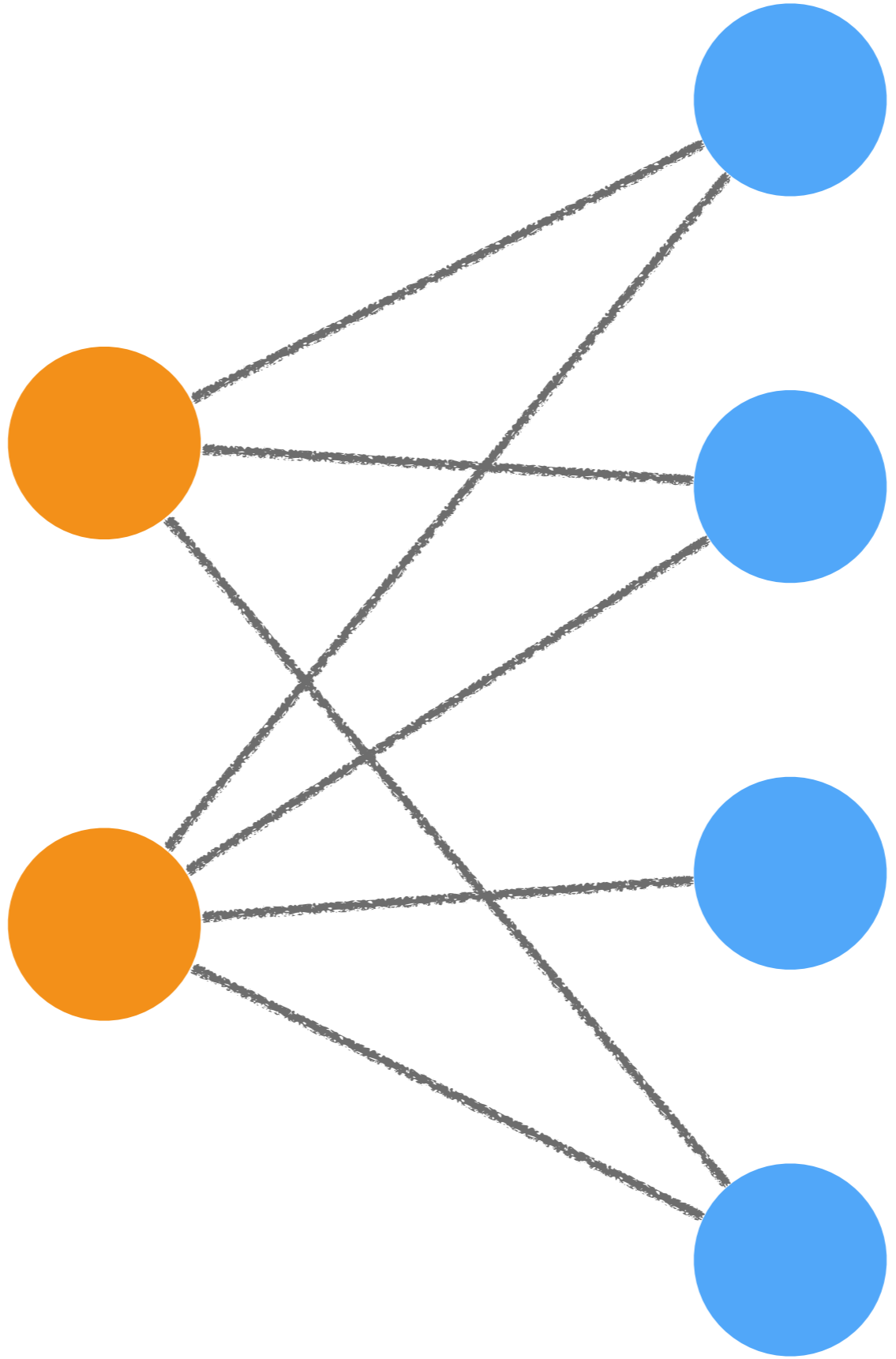
article	
223964	Brooks
303047	The Hotbox
503398	Creating principal 3D curves with digital tape drawing
303033	An exploration into supporting artwork orientation in the user i
258578	An emotional evaluation of orasable user interfaces

Data detail viewer



**Polo**

**Palo**



**Alice**

**Bob**

**Carol**

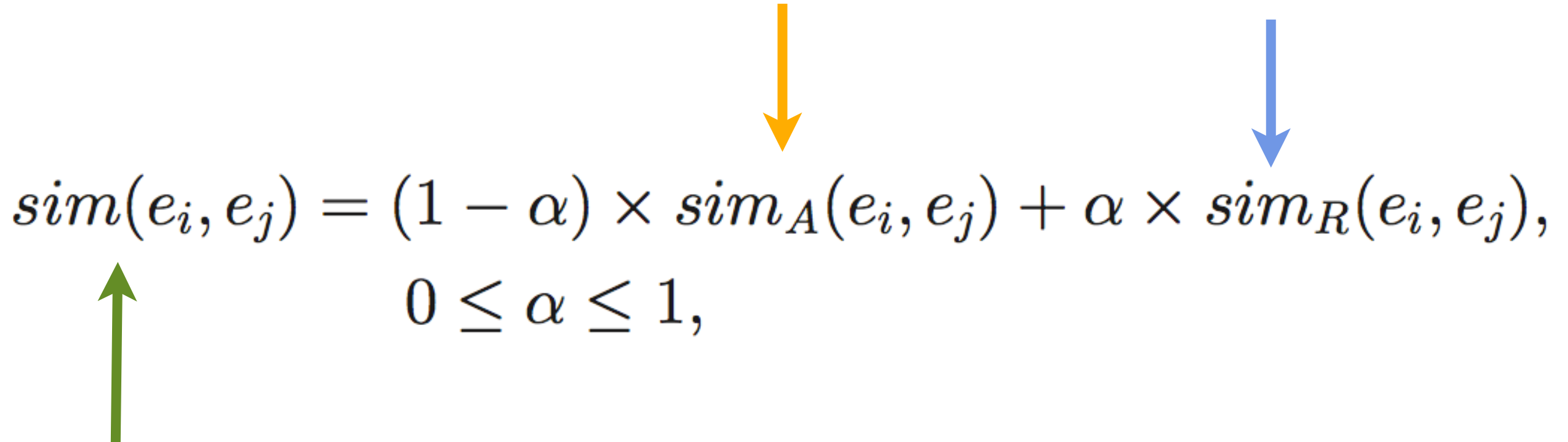
**Dave**

# 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)

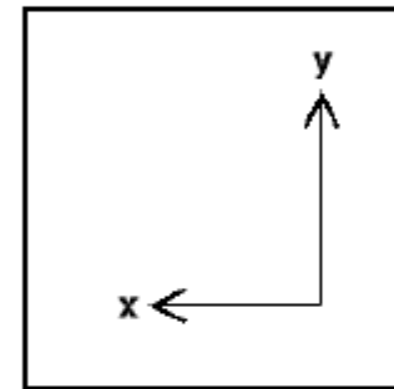


$$sim_A(e_i, e_j) = \sum_{k=1}^n w_k \times sim\_fun_k(e_i \cdot a_k, e_j \cdot a_k),$$
$$-1 \leq w_k \leq 1 \quad \text{and} \quad \sum_{k=1}^n |w_k| = 1,$$

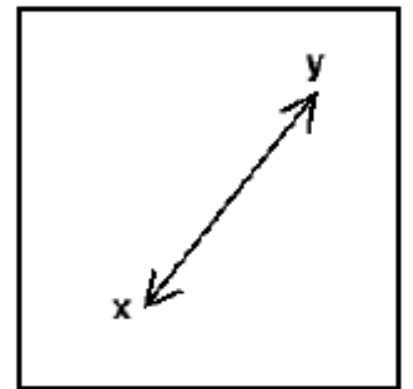
# Numerous **similarity** functions

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

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



Manhattan



Euclidean

- 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”

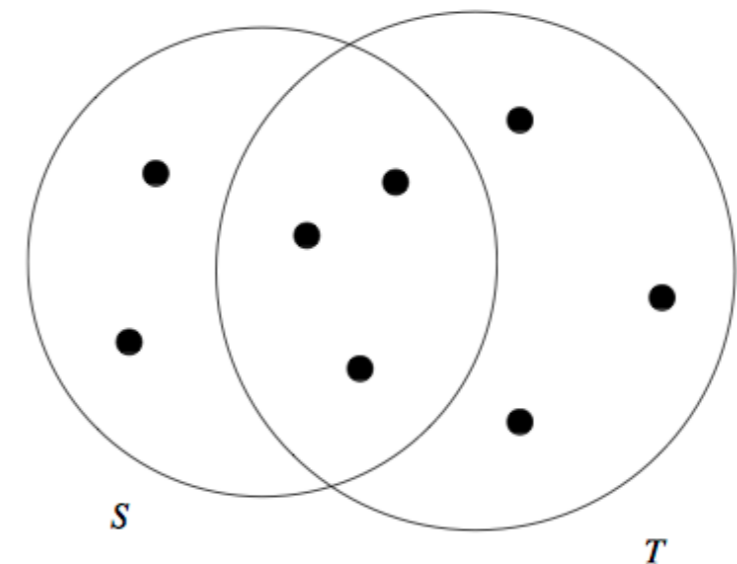


Figure 3.1: Two sets with Jaccard similarity 3/8

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

by Lise Getoor and Ashwin Machanavajjhala