Data Cleaning & Data Integration

Duen Horng (Polo) Chau
Associate Director, MS Analytics
Assistant Professor, College of Computing
Georgia Tech

Google “Polo Chau”

Georgia Tech College of Computing Follow

CV (PDF) Bio Publications Design Students Blog Address

POLO CHAU

Legal name:
Duen Horng Chau

Associate Director, MS in Analytics
Assistant Professor, School of Computational Science & Engineering
Adjunct Assistant Professor, School of Interactive Computing
College of Computing
Georgia Tech

Admin: Carolyn Young Financial Manager: Arlene Washington
polo@gatech.edu www.cc.gatech.edu/~dchau
Office: Klaus 1324 404-385-7682
Google Scholar (h-index: 20) YouTube videos

LinkedIn profile Follow @PoloChau Follow @PoloChau

POSITIONS

May 2014 - Associate Director
MS in Analytics, Georgia Tech

Aug 2012 - Assistant Professor
School of Computational Science & Engineering, Georgia Tech

Dec 2012 - Adjunct Assistant Professor
School of Interactive Computing, Georgia Tech

EDUCATION

Aug 2012 **Ph.D. Machine Learning** Carnegie Mellon University
Thesis: Data Mining Meets HCI: Making Sense of Large Graphs

Polo Club of DATA SCIENCE

NIH MD2K Center of Excellence, Co-PI

MD2K Center of Excellence for Mobile Sensor Data-to-Knowledge

Visualization at Georgia Tech

IDEA workshop at KDD 2013, 2014, 2015

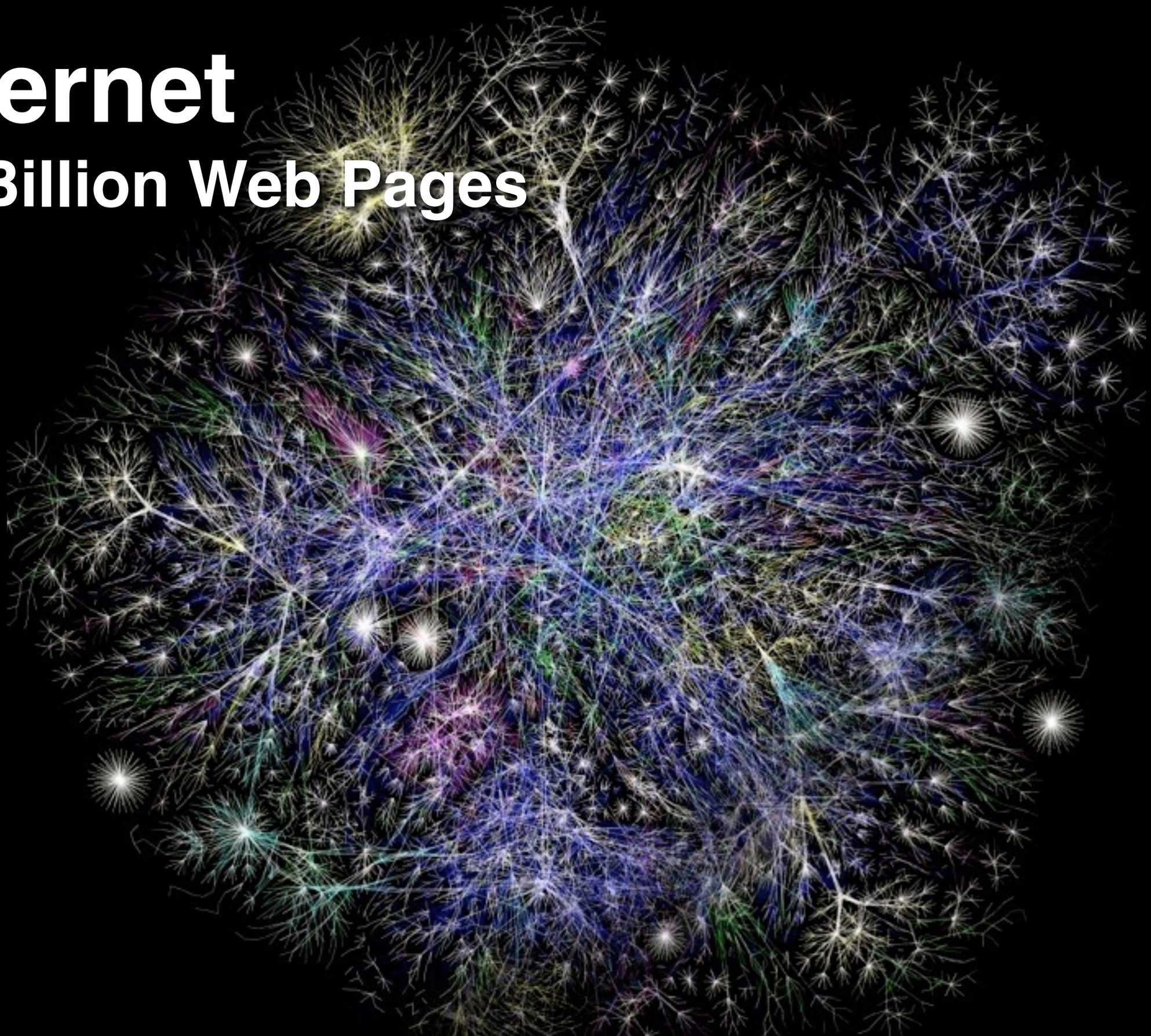
WSDM'16 Publicity Chair

Polo Club
of
DATA SCIENCE

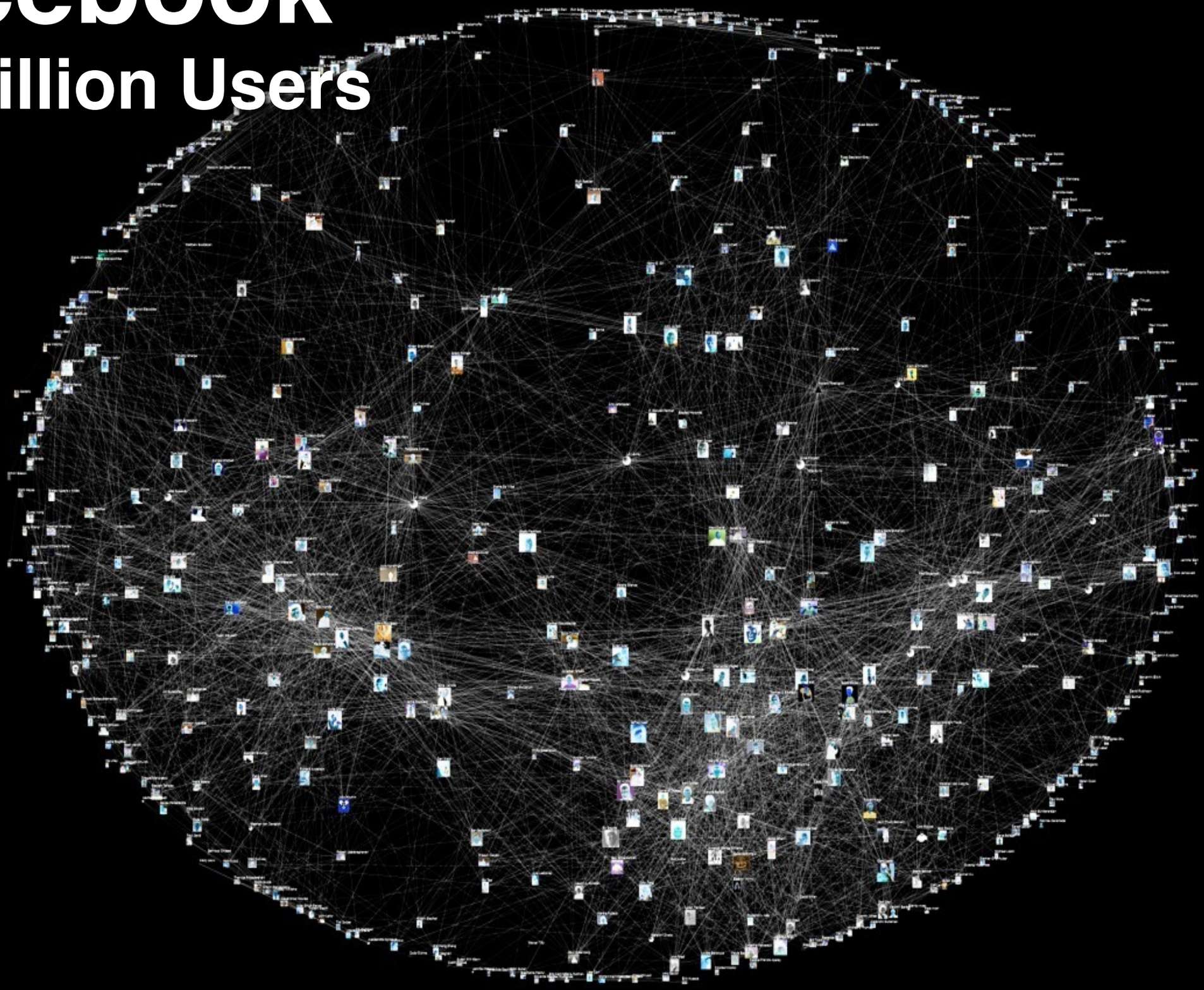
We work with large graphs

Internet

50 Billion Web Pages



Facebook
1.2 Billion Users



Modified from Marc_Smith, flickr

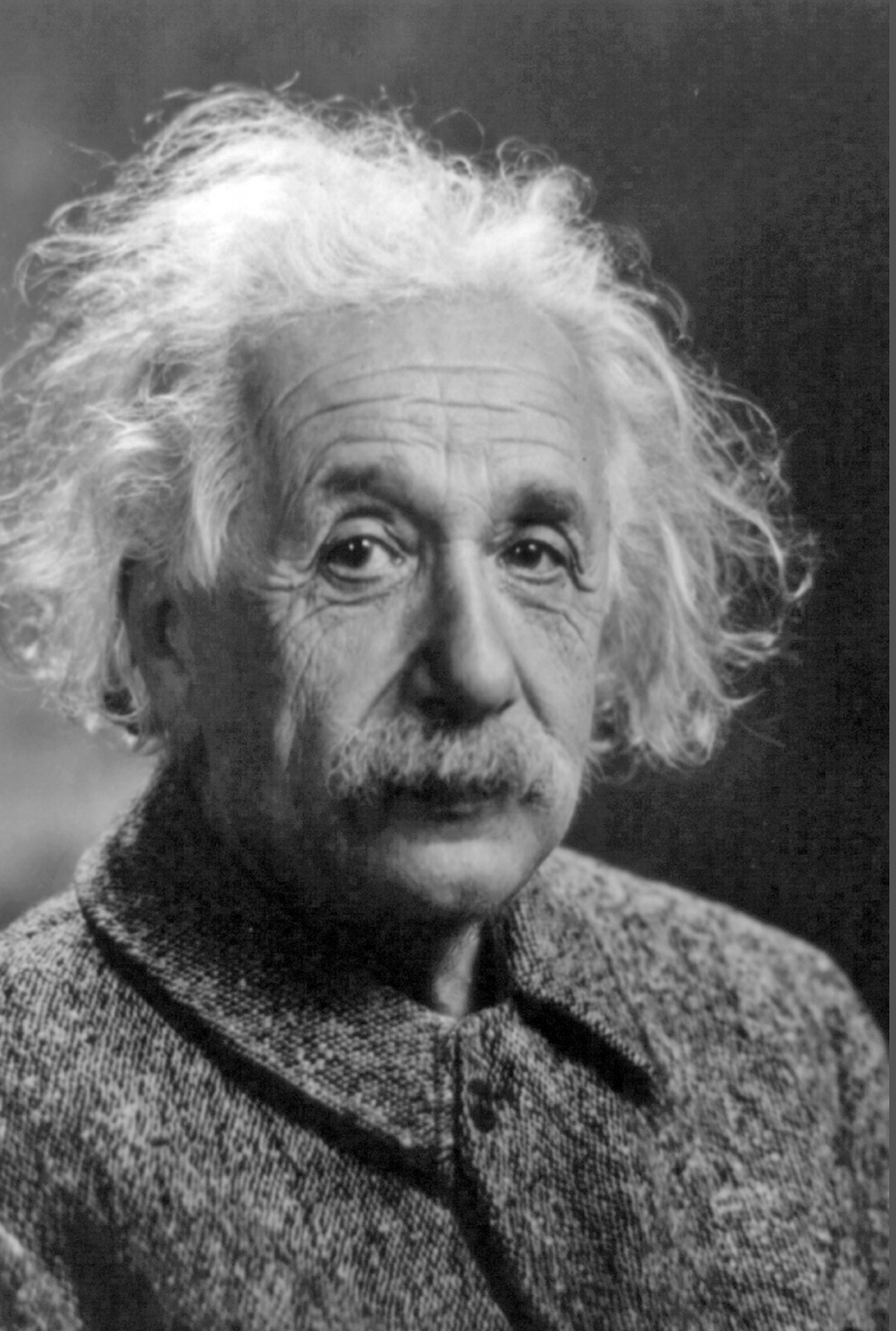
Large Networks We Analyzed

Graph	Nodes	Edges
YahooWeb	1.4 Billion	6 Billion
Symantec Machine-File Graph	1 Billion	**37 Billion**
Twitter	104 Million	3.7 Billion
Phone call network	30 Million	260 Million

Our Approach for Making Sense of Data

DATA MINING	HCI (Human-Computer Interaction)
Automatic	User-driven; iterative
Summarization, clustering, classification	Interaction, visualization
>Millions of items	Thousands of items

Our research combines the
Best of Both Worlds



"Computers are incredibly fast, accurate, and stupid.

Human beings are incredibly slow, inaccurate, and brilliant.

Together they are powerful beyond imagination."

(Einstein might or might not have said this.)

Data Cleaning & Data Integration

- Data “cleaners”: tools for cleaning data
- Data Integration: why you should care, and how to do it

At Georgia Tech, I teach…

http://poloclub.gatech.edu/cse6242

CSE6242 / CX4242

Data & Visual Analytics

Data is Dirty!!

Data Janitor

For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights

Monica Rogati, Jawbone’s vice president for data science, with Brian Wilt, a senior data scientist. Peter

Before we can do all those exciting analysis/machine learning/deep learning…

Data scientists… spend from **50-80%** of their time mired in this more mundane labor of collecting and preparing unruly digital data…

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
Data Wrangler: http://vis.stanford.edu/wrangler/

OpenRefine

A free, open source, powerful tool for working with messy data

Home

Download

Documentation

Community

Post archive

A Governance Model for OpenRefine

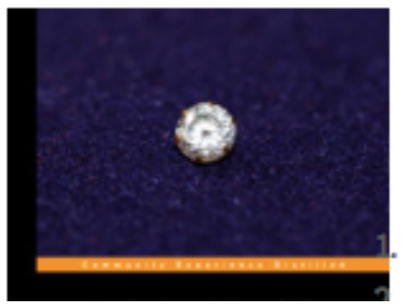
Using OpenRefine: a manual

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

DataWrangler alpha

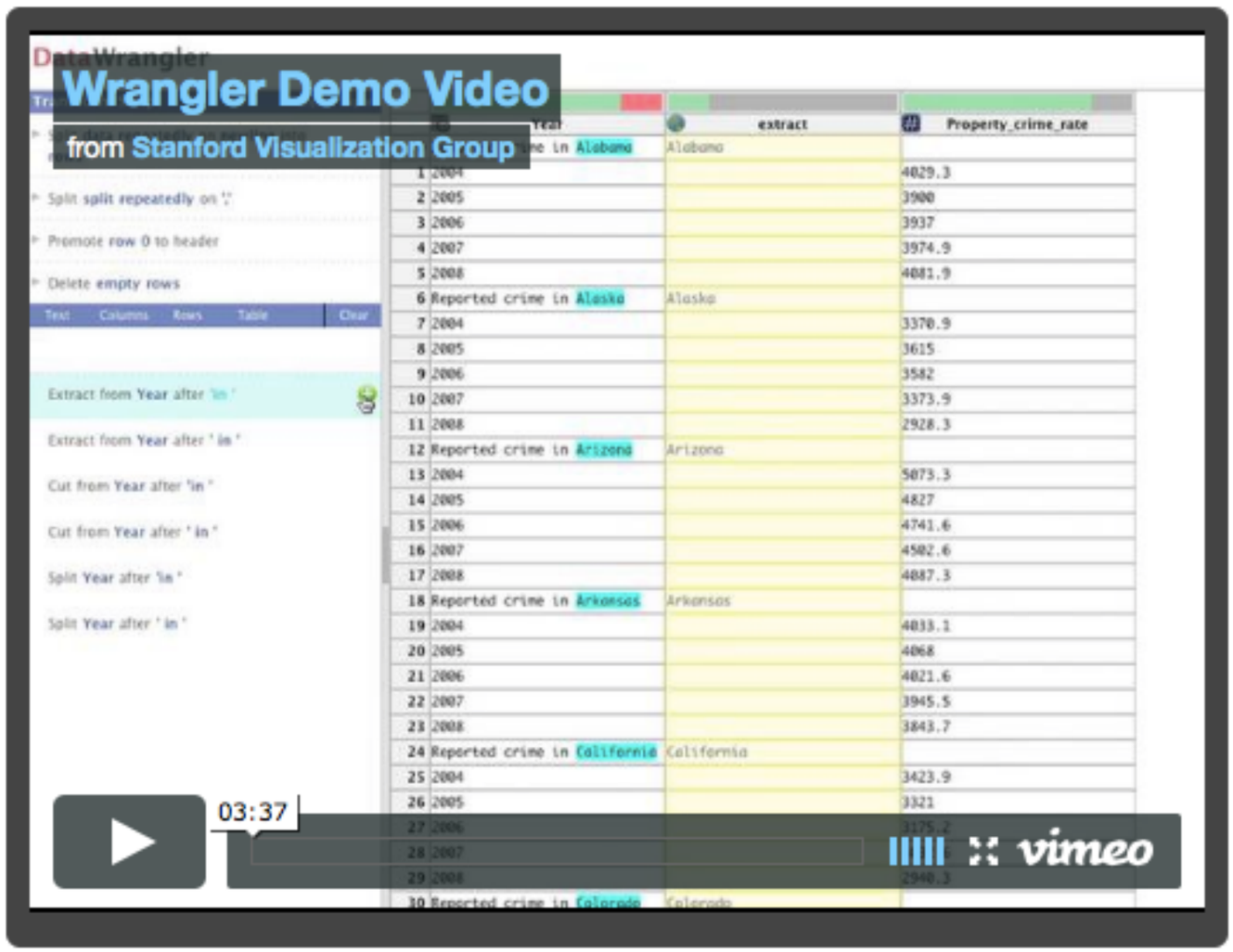
Wrangler is an interactive tool for data cleaning and transformation. Spend less time formatting and more time analyzing your data.

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**?

- [W] replacing null values
- [G, W] reduplication
- [trillium] verify integrity (based on business rule)
- [G] clustering
- [W] unfolding
- [G: integration] pull location info
- [G] offline
- [W] online?
- [W] non-structured -> more structured
- [W, G] output script

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

What is **Data Integration**?

Why is it Important?

Data Integration

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

How to help people effectively leverage multiple data sources?

(People: analysts, researchers, practitioners, etc.)

Examples businesses that derive value via 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 ...

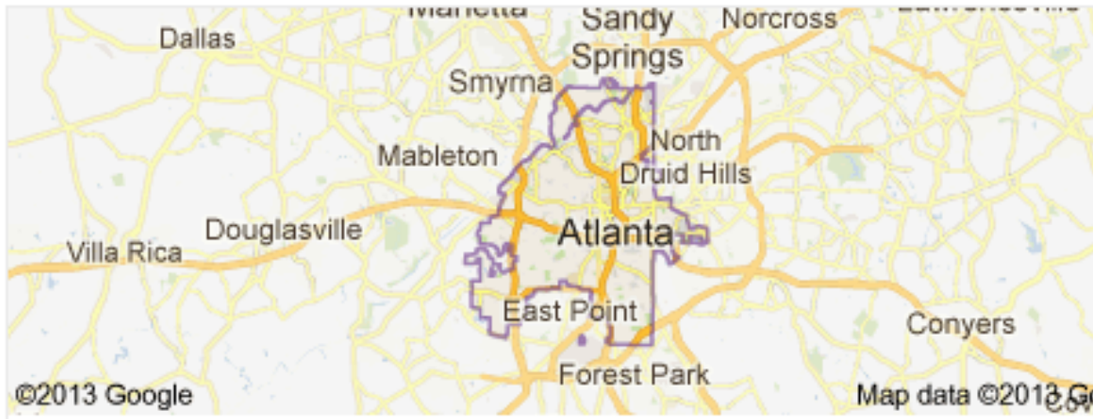
50 Fun Things to Do in **Atlanta** - **Atlanta** Convention and Visitor's ...
www.**atlanta**.net/50fun/
Check out our guide to the top 50 Fun Things to Do in **Atlanta** by activity or neighborhood. The **Atlanta** Convention & Visitors Bureau is your guide to finding fun ...

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

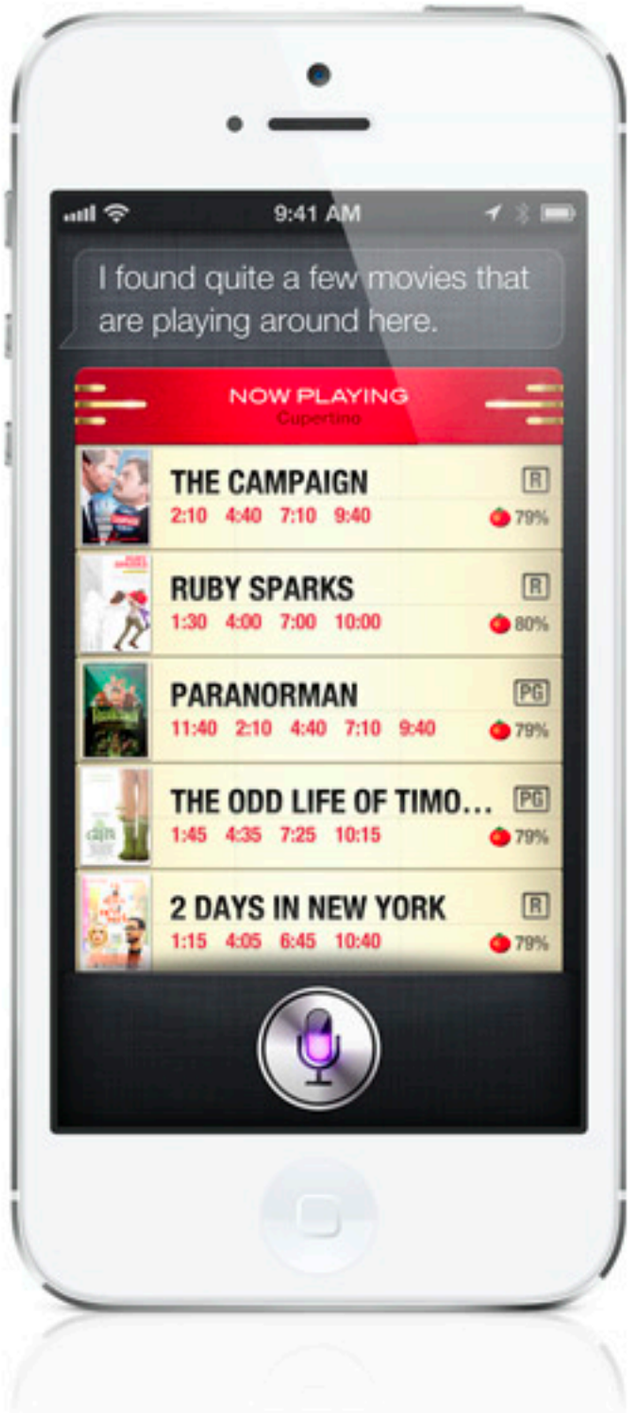
Upcoming events

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

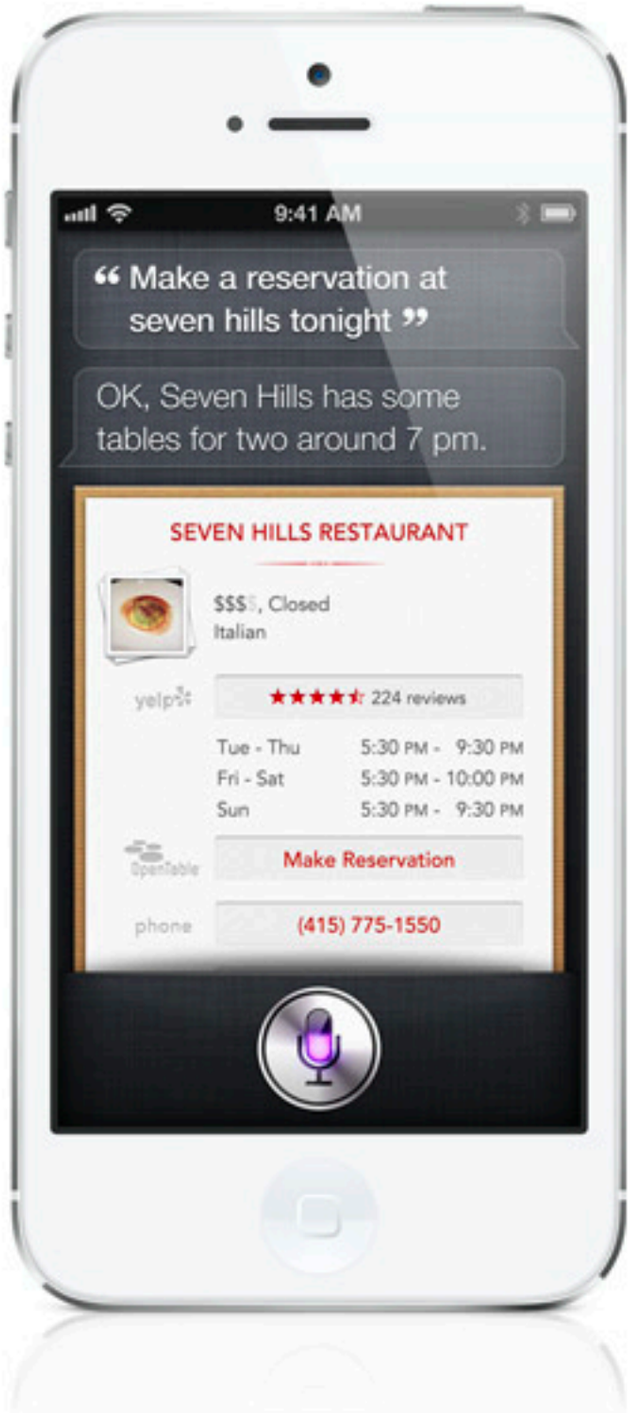
Know the score.

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.

CL atlanta > all atlanta > housing > all [account] post

housing, apartments, real estate, etc

+ show 12 categories

search titles only
has image
posted today
search nearby areas

PRICE
min max

all bedrooms
all bathrooms

FT²
min max

cats ok
dogs ok
furnished
no smoking
wheelchair access

+ housing type
+ laundry
+ parking

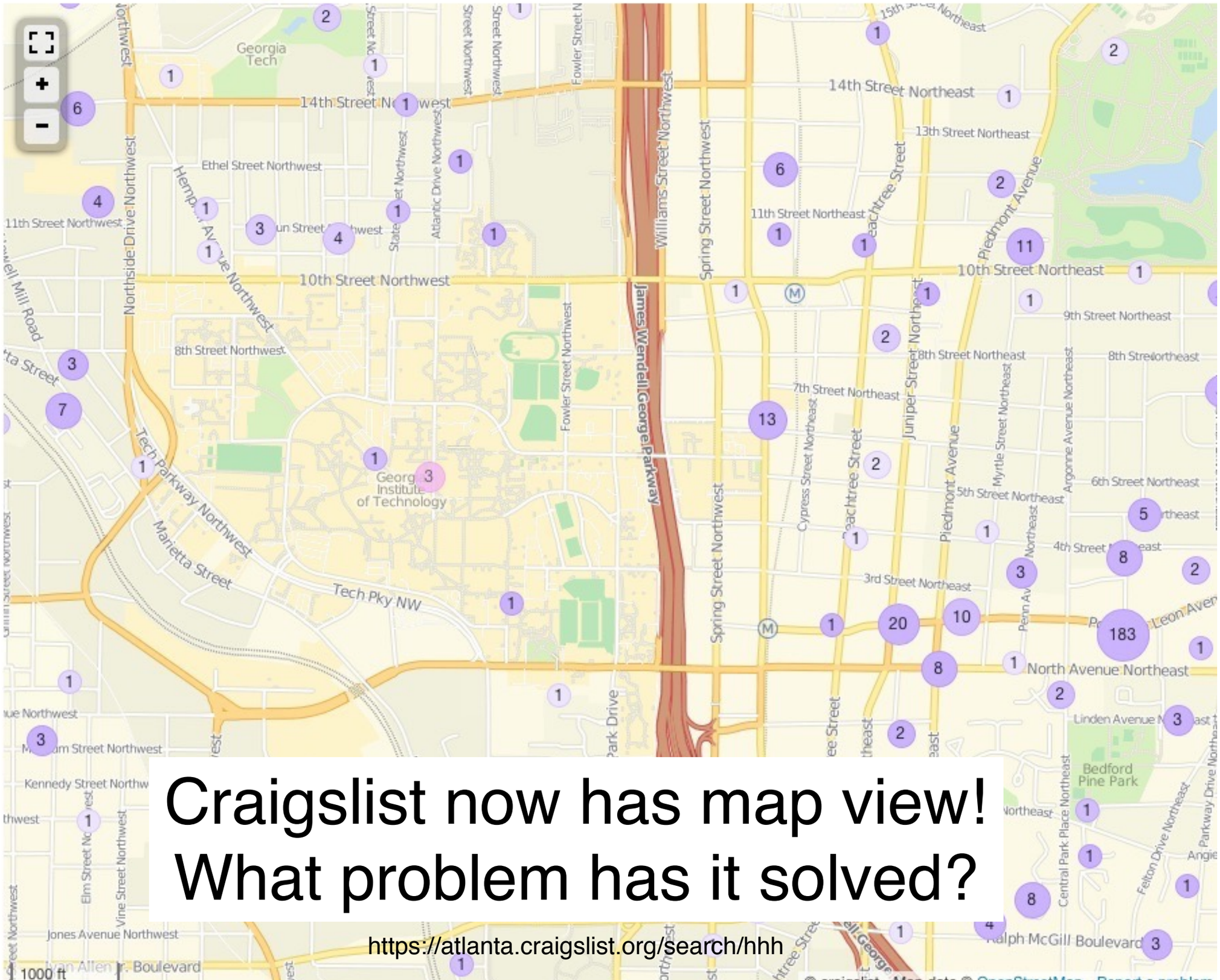
open house date

reset search

search housing, apartments, real estate, etc

email alert save search

list thumb gallery map displaying 2670 of 2500 postings



Craigslist now has map view!
What problem has it solved?

https://atlanta.craigslist.org/search/hhh

KAYAK HOTELS **FLIGHTS** CARS DEALS VACATIONS MORE MY TRIPS LOGIN

Compare hundreds of travel sites at once.
Find the **best deals** faster.

SEARCH ONE AND DONE.

Round-trip | One-way | Multi-city

ATL | To | Depart – Return | 1 adult | Find Flights

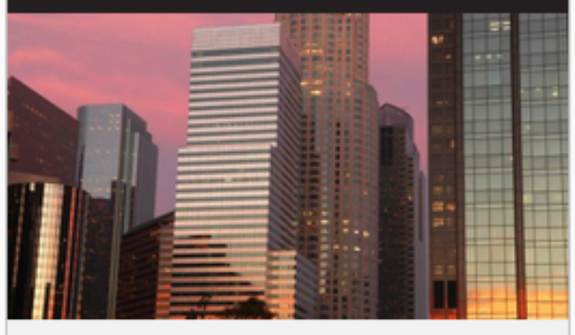
add nearby airports | add nearby airports | find hotels | find car rentals | More search options

✓ I am a student change

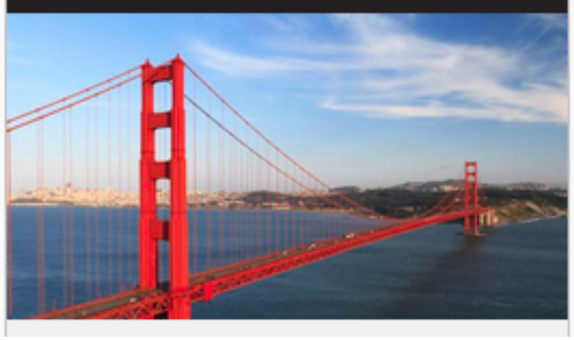
New York

Las Vegas

Los Angeles



San Francisco



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. **Google Refine**

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

So, it’s great to assign an **ID** to everything!

But how?

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 move over 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.

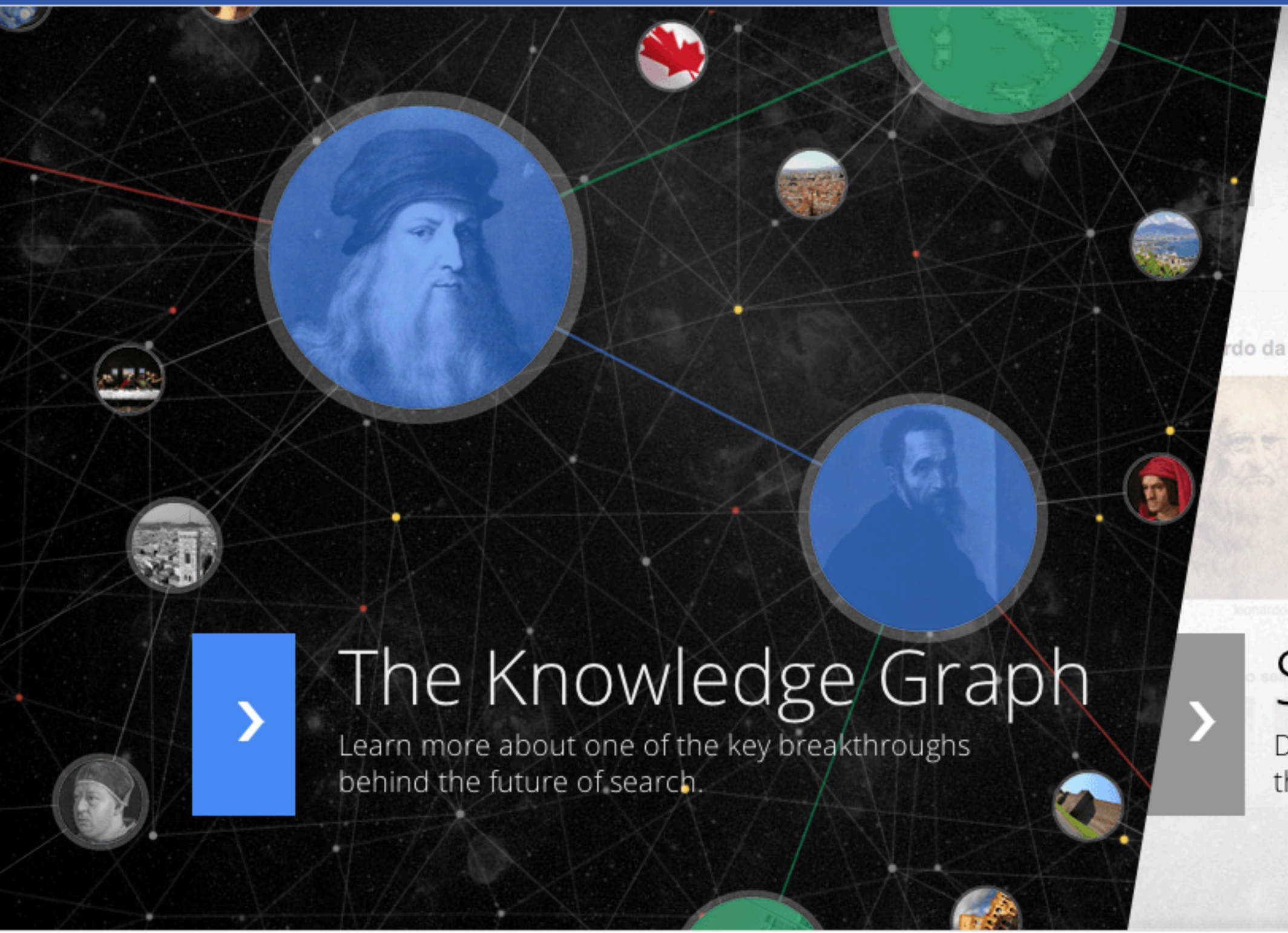
So what?

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.youtube.com/watch?v=mmQl6VGvX-c

facebook

Email or Phone

Password

Log In

Keep me logged in

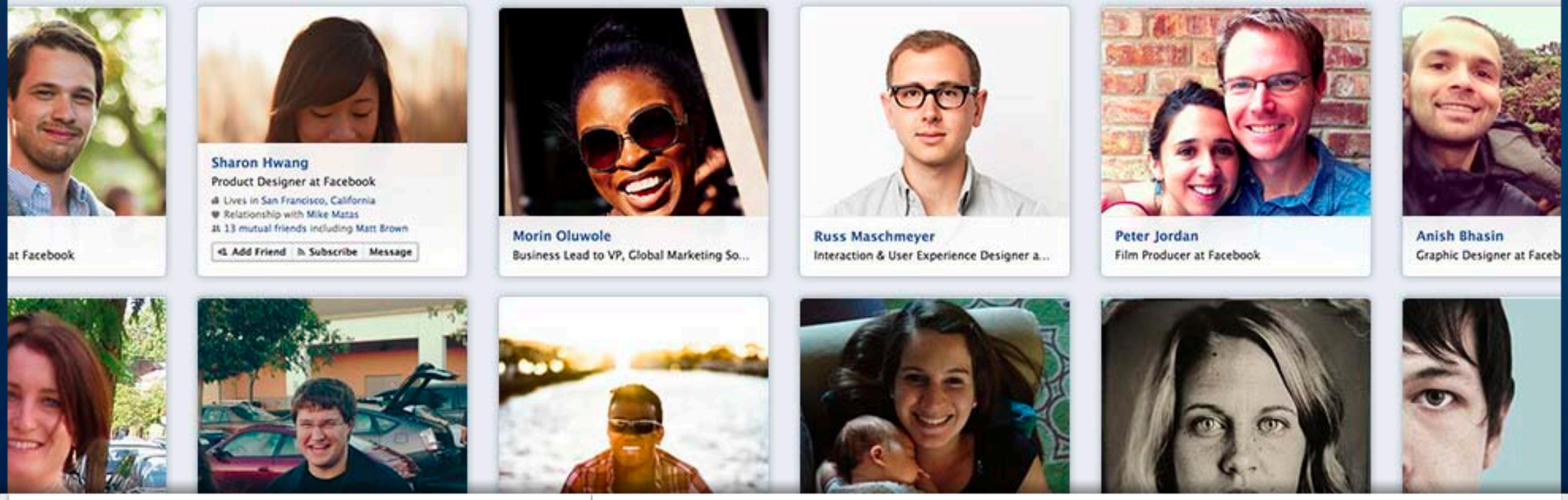
Forgot your password?

Sign Up

Connect and share with the people in your life.

Introducing Graph Search

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



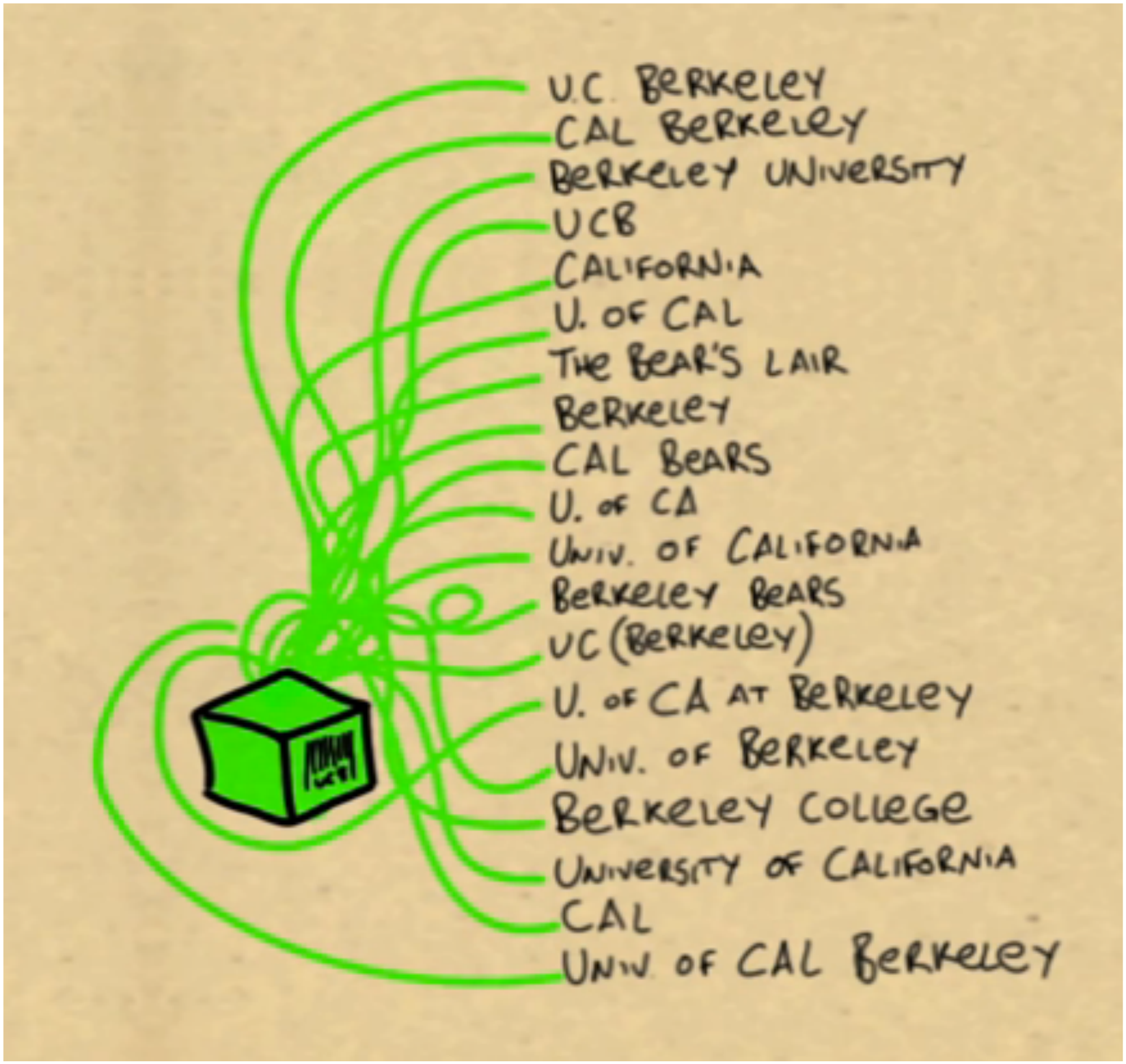
Find people who share your interests

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** important?

Case Study

Let’s shop for an iPhone 6 on Apple, Amazon and eBay

View Gallery

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[1] From $199

1-24 of 12,517,781 results for **Cell Phones & Accessories** : **"iPhone 6"**

Sort by Relevance

Narrow your choices

Any Category
Cell Phones & Accessories
- Cell Phone Cases (11,768,864)
- Accessories (459,701)
- Cases, Holsters & Clips (12,139,858)
- 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
- i-Blason
- Rinbers
- DandyCase
- JETech
- Luvvitt
- Bear Motion
- Samsung
- Beluga
- Nupro
- MXX

Showing results in **Cell Phones & Accessories**. Show instead results in All Departments.

Related Searches: iphone 5, iphone 5s, iphone 6 plus.

See Color & Size Options

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

1,078

See Color & Size Options

Apple iPhone 6 16GB (4.7-inch) 4G LTE Factory Unlocked GSM Dual-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

1,078

See Color Options

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
$1.25 new (7 offers)

FREE Shipping on orders over $35

169

Sponsored

Bolt MFI Certified 3ft Lightning Cab
$13.79 Prime
(25)

iPhone 6 Case, E.C.L USA ® iPho
$19.99 ~~$59.99~~ Prime
(46)

Coosh Mfi-Certified Rapid Heavy-D
$19.99 ~~$69.95~~ Prime

Categories

Cell Phones & Accessories

Cell Phones & Smartphones

Cell Phone Cases, Covers & Skins

More

See all categories

Storage Capacity see all

- 128GB (965)
- 64GB (2,021)
- 32GB (14)
- 16GB (3,326)
- Not Specified (519)

Model see all

- iPhone 6 (3,514)
- iPhone 6 Plus (2,091)
- iPhone 5s (15)

Network see all

- AT&T (1,391)
- Sprint (953)
- T-Mobile (900)
- Verizon (943)

Color see all

- Black (221)
- Gold (2,001)
- Gray (2,096)
- Pink (10)
- Silver (1,301)
- White (205)

Screen Size see all

All Listings | Auction | Buy It Now

Sort: Best Match View:

All > Cell Phones & Accessories > Cell Phones & Smartphones

iphone 6 6,847 listings + Follow this search

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

Popular on eBay

Apple iPhone 5 - 16GB - (Factory Unlocked)...

$219.88

Buy It Now

Free shipping

Popular

Apple iPhone 5C - 16gb - Factory GSM...

$194.99

Buy It Now

Free shipping

Popular

D-Dupe

Interactive Data Deduplication and Integration
TVCG 2008

University of Maryland
Bilgic, Licamele, Getoor, Kang, Shneiderman

http://linqs.cs.umd.edu/basilic/web/Publications/2008/kang:tvcg08/kang-tvcg08.pdf

http://www.cs.umd.edu/projects/linqs/ddupe/ (skip to 0:55)

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

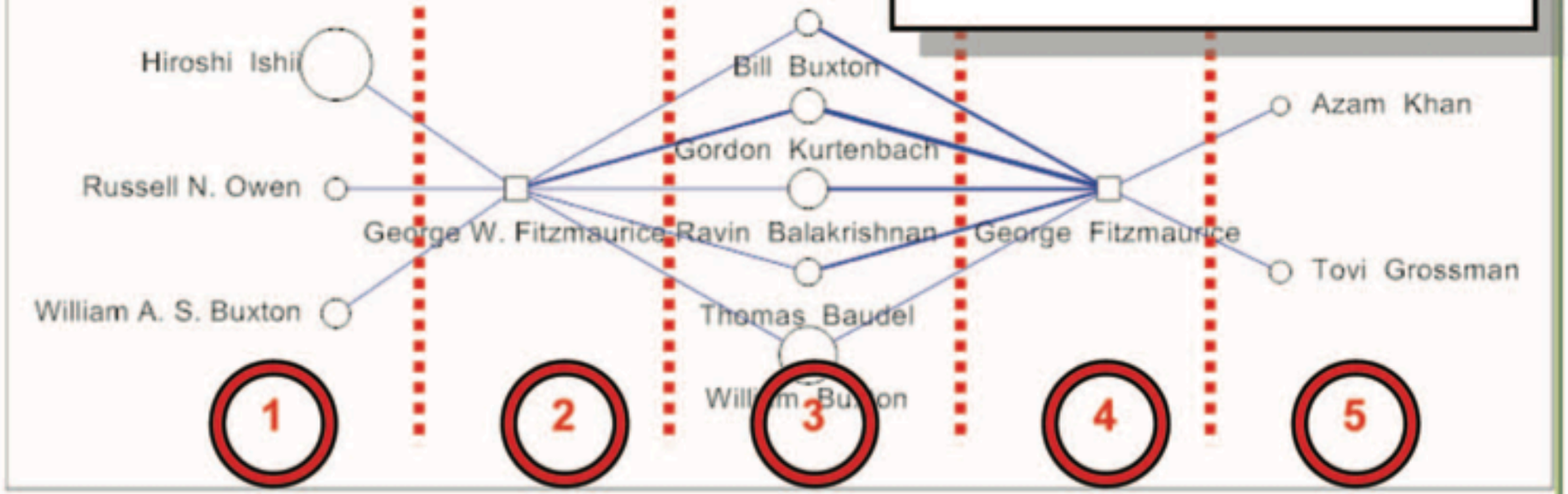
Search

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)

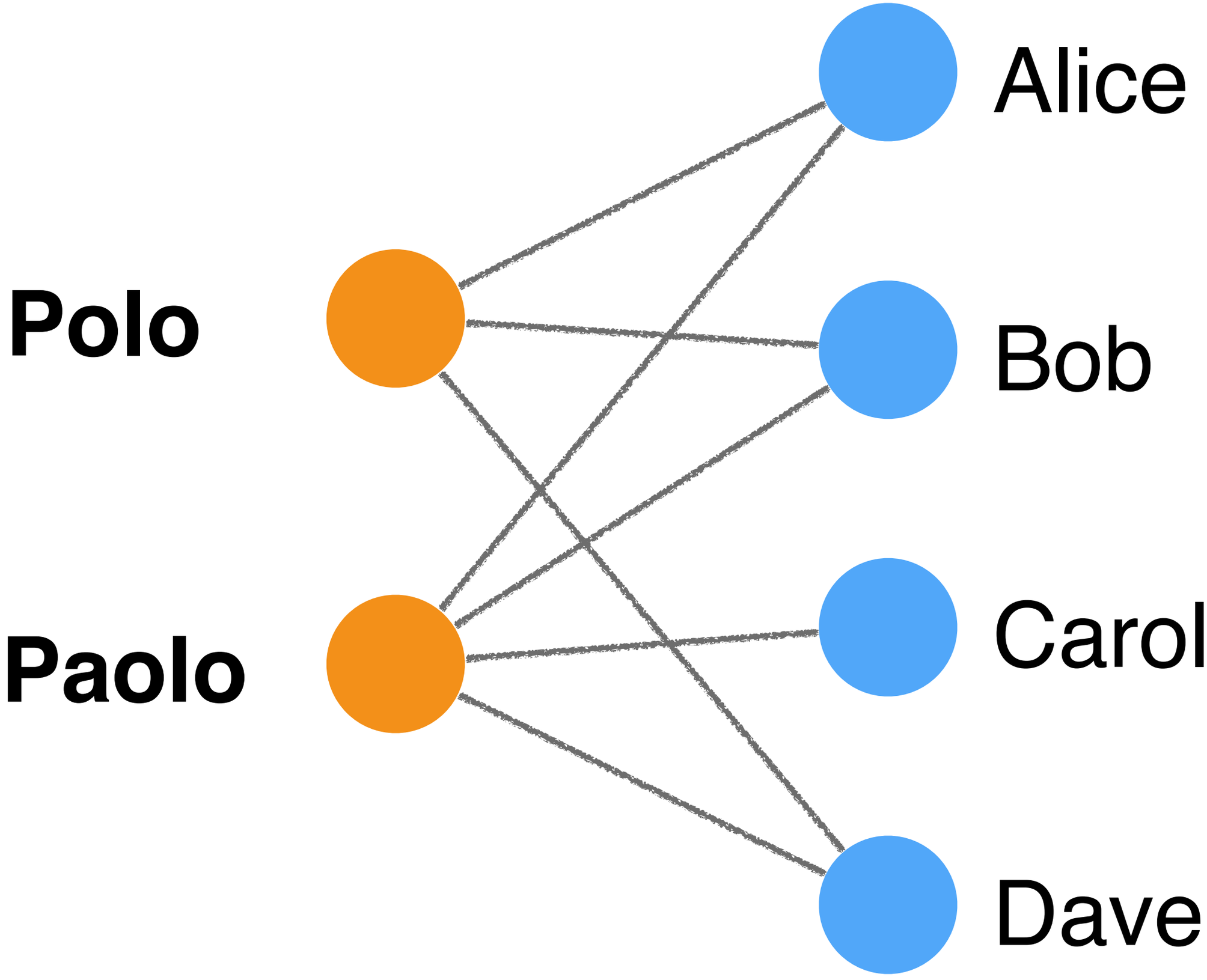
person_id	full_name	last_name	first_name	mid
P110925	Hiroshi Ishii	Ishii	Hiroshi	
P298693	William A. S. Buxton	Buxton	William	A. S
P250512	Russell N. Owen	Owen	Russell	N.
P284951	Tovi Grossman	Grossman	Tovi	
P23365	Azam Khan	Khan	Azam	

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

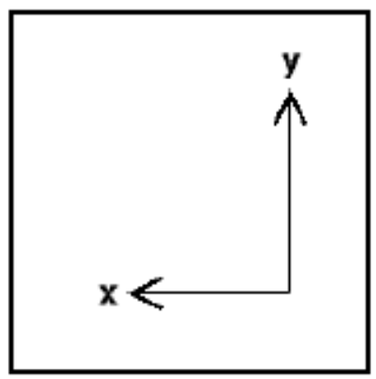
Finding possible duplicates completed!


Polo
Alice
Bob
Paolo
Carol
Dave

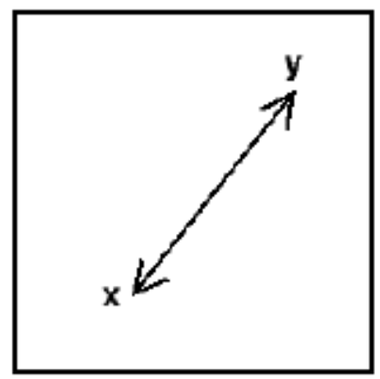
Numerous **similarity functions**

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"
- Canberra distance
  (compare ranked items)

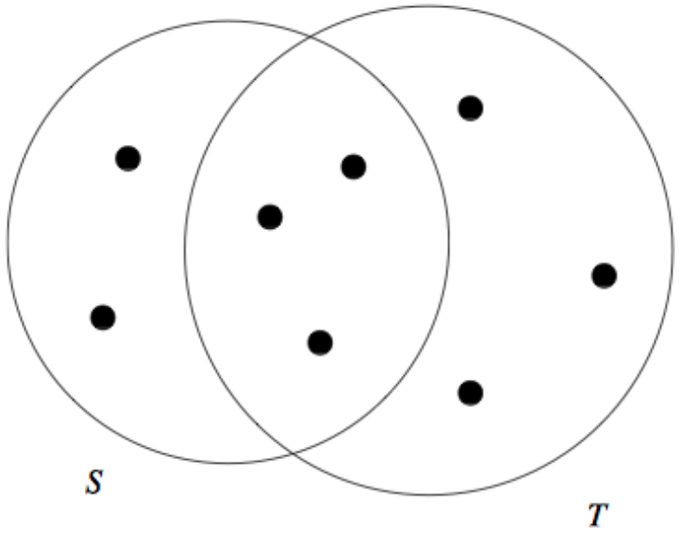


Figure 3.1: Two sets with Jaccard similarity 3/8

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)

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

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