

# Data Integration and Management

Dr. Bo (Beth) Sun

\*Partly based on materials by Prof. Chau in Georgia Tec

# 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.

About 932,000,000 results (0.63 seconds)

## Top stories



Airbnb pulled out every trick to stop NYC from curbing rentals

CNET

2 hours ago



Airbnb's NYC Bookings Could Be Cut in Half by New Rule

Bloomberg

5 hours ago



NYC's Best Restaurants of 2018 So Far, According to Ryan Sutton

Eater NY

25 mins ago

[➔ More for NYC](#)

## The Official Guide to New York City | nycgo.com

<https://www.nycgo.com/>

Find out what to do, where to go, where to stay and what to eat in NYC from the experts who know it best.

[Basic Information](#) · [Transportation](#) · [Weather](#) · [Official Visitor Centers](#)

## Welcome to NYC.gov | City of New York

<https://www1.nyc.gov/>

The official website of the City of New York. Find information about important alerts, 311 services, news, programs, events, government employment, the office of ...

[NYC Resources](#) · [NYC311](#) · [Office of the Mayor](#) · [Events](#)

## New York City - Wikipedia

[https://en.wikipedia.org/wiki/New\\_York\\_City](https://en.wikipedia.org/wiki/New_York_City)

The City of New York, often called **New York City** (**NYC**) or simply New York, is the most populous city in the United States. With an estimated 2017 population of ...

[Megacity](#) · [Boroughs of New York City](#) · [History of New York City](#) · [New York Harbor](#)

## Things to do in New York City



Statue of Liberty



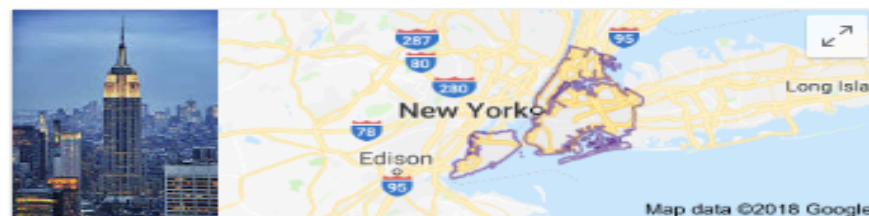
Central Park



Empire State



Times Square



## New York City

City in New York

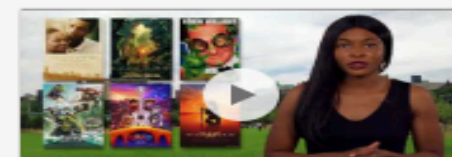
New York City comprises 5 boroughs sitting where the Hudson River meets the Atlantic Ocean. At its core is Manhattan, a densely populated borough that's among the world's major commercial, financial and cultural centers. Its iconic sites include skyscrapers such as the Empire State Building and sprawling Central Park. Broadway theater is staged in neon-lit Times Square.

**Land area:** 304.6 mi<sup>2</sup>**Weather:** 80°F (27°C), Wind SE at 14 mph (23 km/h), 79% Humidity**Local time:** Tuesday 10:40 AM**Population:** 8.538 million (2016)**Mayor:** [Bill de Blasio](#)

## Plan a trip

[New York City travel guide](#)[3-star hotel averaging \\$207, 5-star averaging \\$448](#)[1 h flight, from \\$305](#)**Colleges and Universities:** [New York University](#), [MORE](#)

**Did you know:** New York City is the second-most-populous North American city by population (8,622,698). [wikipedia.org](#)

NYC 311  
on Google

# 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 PACKAGESROUND-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

 Find...

BrowseQueryHelp

Sign In or Sign UpEnglish ▾

Important! Freebase is read-only and will be shut-down. [More.](#)3,179,263,202Facts  
(and counting)

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

DataSchemaQueriesAppsLoadsReview TasksUsers

Explore Freebase Data

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 banner features a dark background with a network of nodes and connecting lines, representing the Knowledge Graph. Nodes include a large portrait of Leonardo da Vinci, a Canadian flag, a green globe, and various smaller images of historical figures and landmarks. The Google logo is in the top left, and navigation links are in a blue bar at the top. The main title 'The Knowledge Graph' is in large white text, with a blue arrow pointing right. Below it is a subtitle. On the right, a sidebar shows a detailed view of Leonardo da Vinci's entry, including his portrait and biographical details.

Google Inside Search

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

# The Knowledge Graph

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

**Leonardo da Vinci**

Leonardo di ser Piero da Vinci  
Renaissance polymath: painter, architect, musician, scientist, engineer, inventor, anatomist, 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 Da Vinci

See it in

Discover answers to questions you thought to ask, and explore new ones.

Learn more about Google Knowledge Graph at <https://goo.gl/mkCKMg>

Google has the Knowledge Graph.

Facebook has...



facebook

Email or Phone

☐ Keep me logged in

Password

Log In

Sign Up

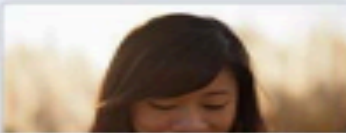
Connect and share with the people in your life.

# Introducing Graph Search

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



at Facebook



**Sharon Hwang**

Product Designer at Facebook

📍 Lives in San Francisco, California

👤 Relationship with Mike Matas

👥 13 mutual friends including Matt Brown

➕ Add Friend    📧 Message



**Merin Okunola**

Business Lead to VP, Global Marketing So...



**Russ Maschmeyer**

Interaction & User Experience Designer a...



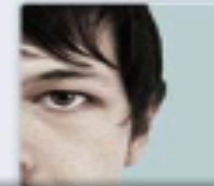
**Peter Jordan**

Film Producer at Facebook



**Anish Bhasin**

Graphic Designer at Facebook

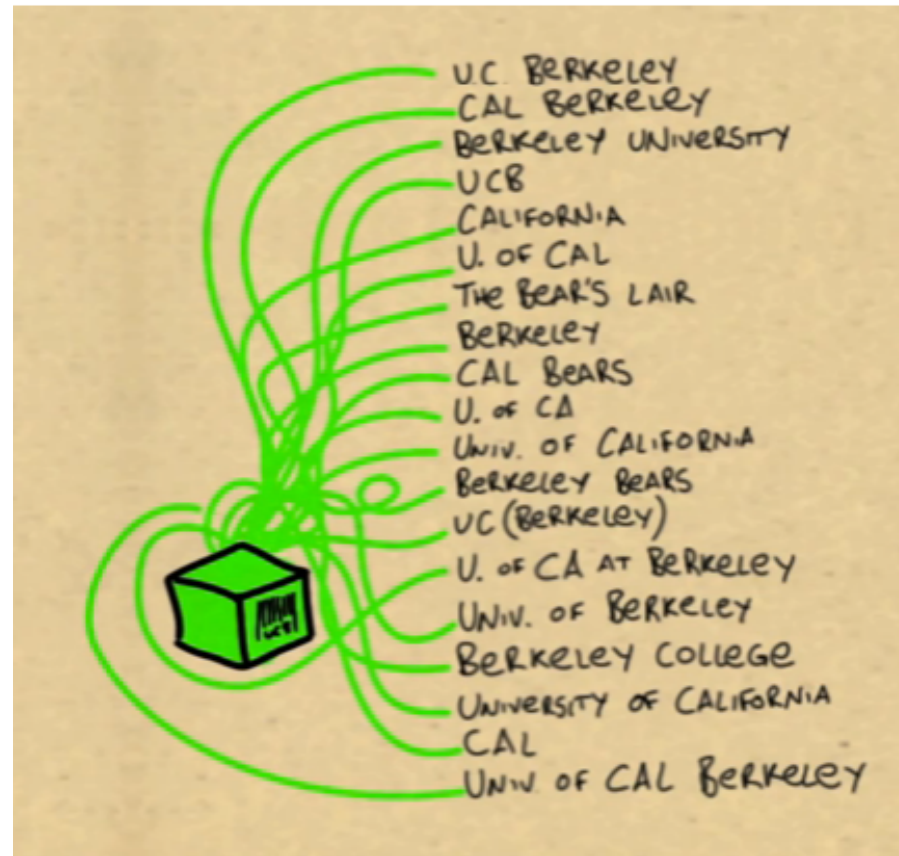


## Find people who share your interests

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

Graph Search intro video: <https://youtu.be/W3k1USQbq80>

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



(Screenshot from FreeBase video)

Then you need to do...

# **Entity Resolution**

(A hard problem in data integration)

# What is VIS?



[https://www.ted.com/talks/hans\\_rosling\\_shows\\_the\\_best\\_stats\\_you\\_ve\\_ever\\_seen](https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen)