http://poloclub.gatech.edu/cse6242

CSE6242: Data & Visual Analytics

Data Collection

Duen Horng (Polo) Chau

Associate Professor, College of Computing

Associate Director, MS Analytics

Georgia Tech

Mahdi Roozbahani

Lecturer, Computational Science & Engineering, Georgia Tech

Founder of Filio, a visual asset management platform

Partly based on materials by

Professors Guy Lebanon, Jeffrey Heer, John Stasko, Christos Faloutsos

How to Collect Data?

Method	Effort	
Download	Low	
API (Application program interface)	Medium	
Scrape/Crawl	High	

Data you can just download

NYC Taxi data: Trip (11GB), Fare (7.7GB)

StackOverflow (xml)

Wikipedia (data dump)

Atlanta crime data (csv)

Soccer statistics

Data.gov

…

Data you can just download

If you have leads, let us know on Piazza!

More datasets on course website:

Collect Data via APIs

Google Data API (e.g., Google Maps Directions API)
https://developers.google.com/gdata/docs/directory

Twitter (small subset)
https://dev.twitter.com/streaming/overview

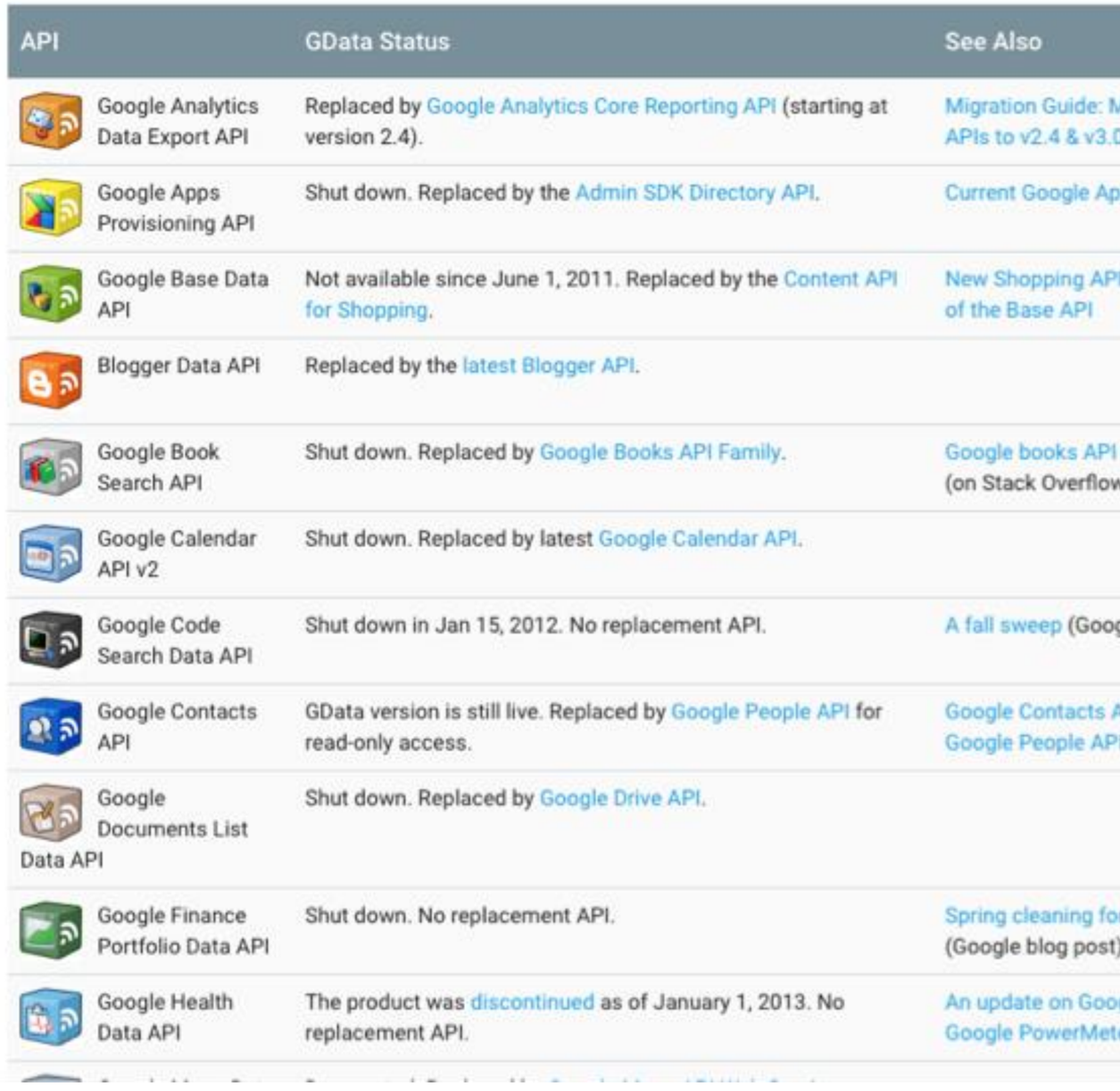
Last.fm (Pandora has unofficial API)

Flickr

data.nasa.gov

data.gov

Facebook (your friends only)



API	GData Status	See Also
Google Analytics Data Export API	Replaced by Google Analytics Core Reporting API (starting at version 2.4).	Migration Guide: M APIs to v2.4 & v3.0
Google Apps Provisioning API	Shut down. Replaced by the Admin SDK Directory API.	Current Google Ap
Google Base Data API	Not available since June 1, 2011. Replaced by the Content API for Shopping.	New Shopping API of the Base API
Blogger Data API	Replaced by the latest Blogger API.	
Google Book Search API	Shut down. Replaced by Google Books API Family.	Google books API (on Stack Overflow
Google Calendar API v2	Shut down. Replaced by latest Google Calendar API.	
Google Code Search Data API	Shut down in Jan 15, 2012. No replacement API.	A fall sweep (Goog
Google Contacts API	GData version is still live. Replaced by Google People API for read-only access.	Google Contacts A Google People API
Google Documents List Data API	Shut down. Replaced by Google Drive API.	
Google Finance Portfolio Data API	Shut down. No replacement API.	Spring cleaning fo (Google blog post)
Google Health Data API	The product was discontinued as of January 1, 2013. No replacement API.	An update on Goo Google PowerMet

Data that needs scraping

Amazon (reviews, product info)

ESPN

eBay

Google Play

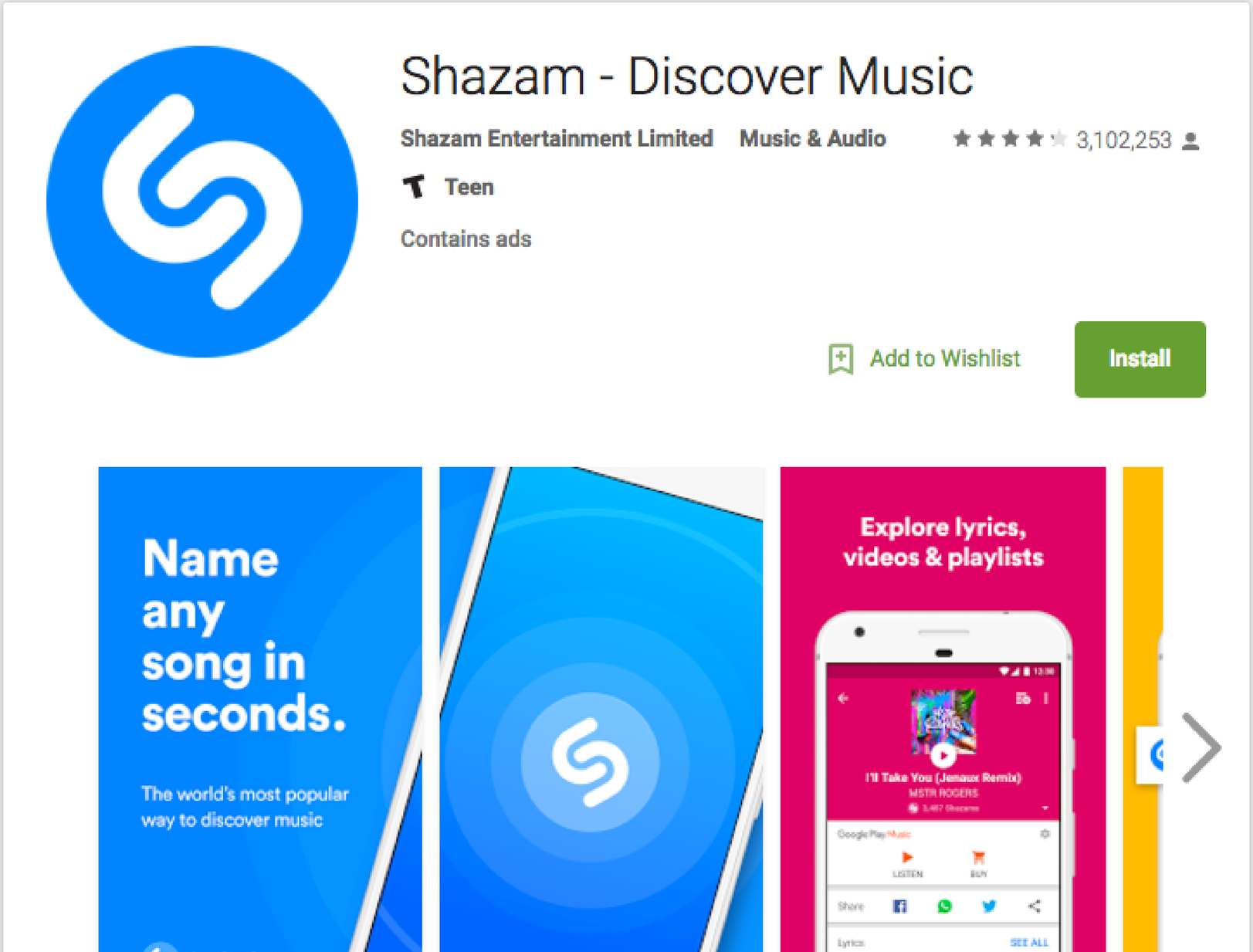
Google Scholar

...

How to Scrape?

Google Play example

Goal: collect the network of similar apps



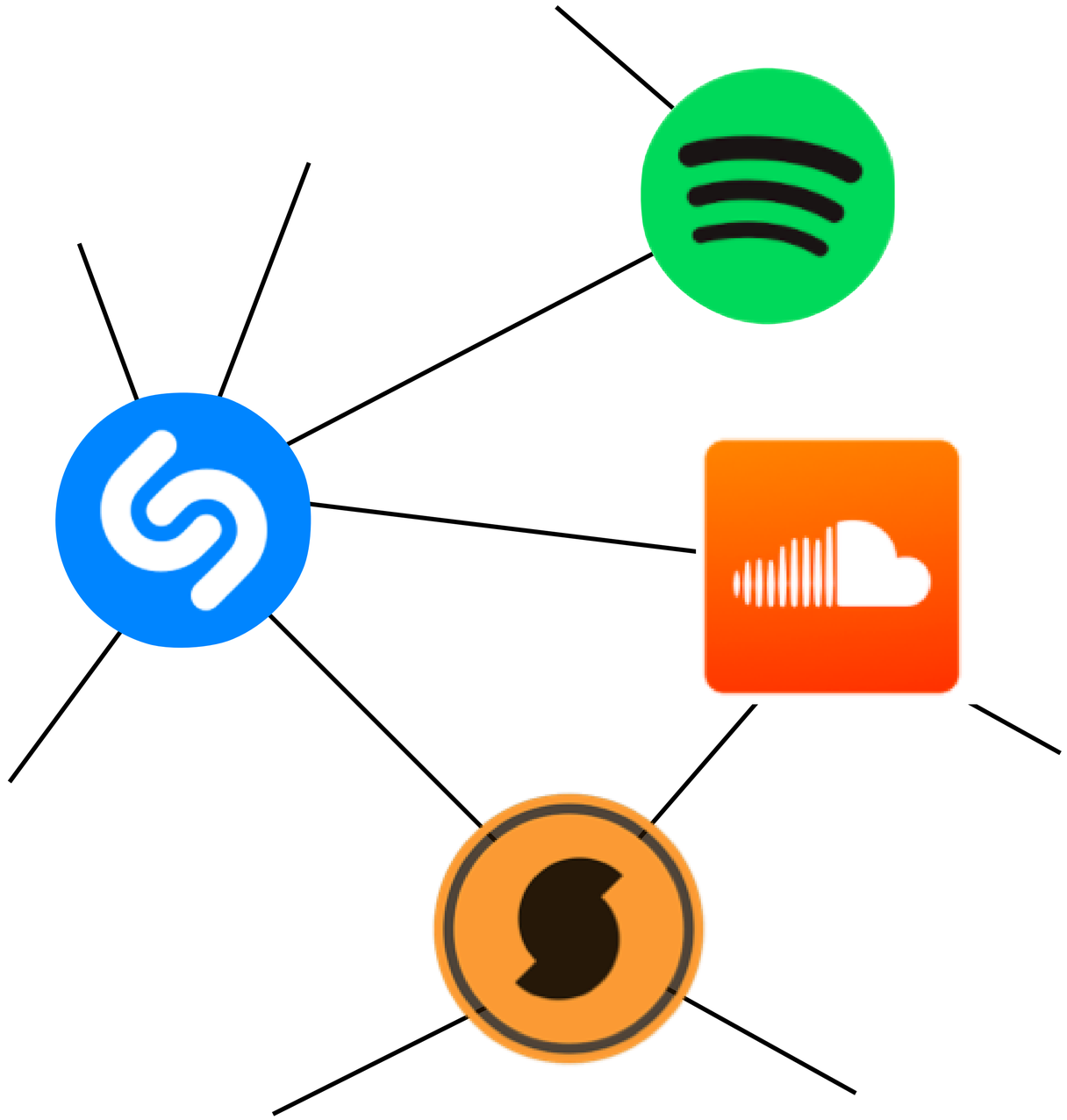
Name any song in seconds

Shazam will identify any music playing around you.

GET IT NOW

How to Scrape?

Goal: Write a **program/algorithm** to scrape Google Play to **collect a million-node network** of similar apps



Each **node** is an app

An **edge** connects two similar apps

Hint: start with some apps (e.g., Shazam), and go from there.

How to Scrape?

Google Play example

Goal: collect the network of similar apps

https://play.google.com/store/apps/details?id=**com.shazam.android**

https://play.google.com/store/apps/details?id=**com.spotify.music**

Popular Scraping Libraries

Selenium. Supports multiple languages. http://www.seleniumhq.org

Beautiful Soup. Python. https://www.crummy.com/software/BeautifulSoup

Scrapy. Python. https://scrapy.org

JSoup. Java. https://jsoup.org

Important considerations:

Different web content shows up depending on web browsers used

Scraper may need different “web driver” (e.g., in Selenium), or browser “user agent”

Data may show up after certain user interaction (e.g., click a button)

- Scraper may need to simulate the actions.
- Selenium supports more actions than beautiful soup: http://www.discoversdk.com/blog/web-scraping-with-selenium