

poloclub.github.io/#cse6242

CSE6242/CX4242: **Data** & **Visual** Analytics

Data Collection

Duen Horng (Polo) Chau

Professor, College of Computing
Associate Director, MS Analytics
Georgia Tech

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 Ed Discussion!

More datasets on course website:

CSE6242A/CX4242A

Schedule

Homework

Project

Warnings

Policy

Datasets

Resources

There are [multiple CSE6242 sections](#). This is the course homepage for **campus CSE6242A/**

CSE6242A/CX4242A Fall 2024

Data and Visual Analytics

Georgia Tech, College of Computing

Tue & Thu, 3:30-4:45pm, [Clough 152](#)

Collect Data via APIs

Google Data API

(e.g., Google Maps Directions API)

<https://developers.google.com/gdata/docs/directory>

Last.fm (Pandora has unofficial API)

Flickr

data.nasa.gov

data.gov










Facebook (your friends only)

Home > Products > Google Data APIs > Guides

GData API Directory

Warning: Several of the APIs listed on this page are deprecated or obsolete, and some have not use the Google Data Protocol.

The following Google services provide APIs that implement, or used to implement,

API	GData Status
 Google Analytics Data Export API	Replaced by Google Analytics Core Reporting API (starting at version 2.4).
 G Suite Provisioning API	Shut down. Replaced by the Admin SDK Directory API .
 Google Base Data API	Not available since June 1, 2011. Replaced by the Content API for Shopping .
 Blogger Data API	Replaced by the latest Blogger API .
 Google Book Search API	Shut down. Replaced by Google Books API Family .
 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.
 Google Contacts API	Contacts API is deprecated and is schedule to be shut down on June 15, 2021. Replaced by Google People API .
 Google Documents List Data API	Shut down. Replaced by Google Drive API .

Data that needs scraping

Amazon (reviews, product info)

ESPN

eBay

Google Play

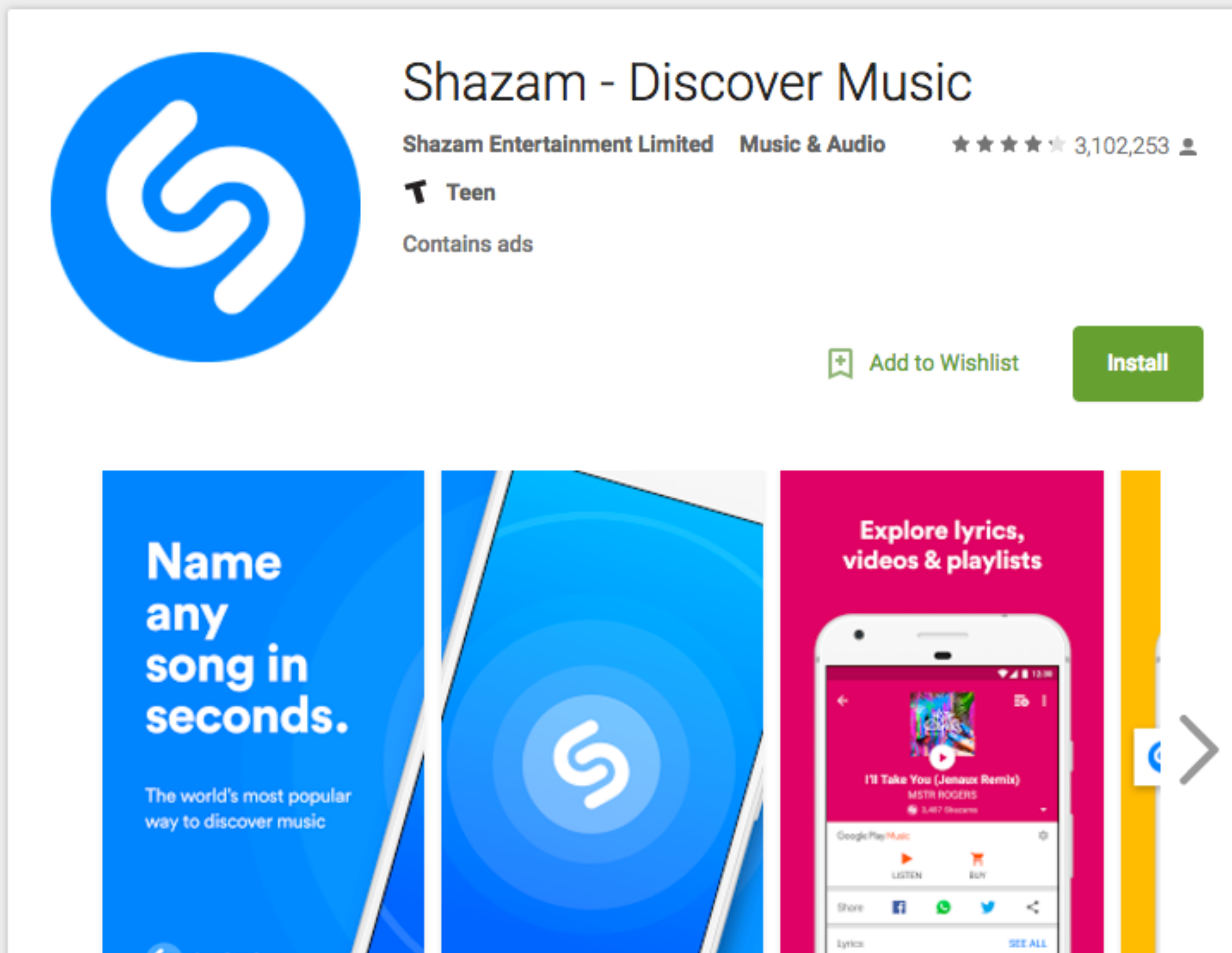
Google Scholar

...

How to Scrape?

Google Play example

Goal: collect the network of similar apps



Shazam - Discover Music
Shazam Entertainment Limited Music & Audio ★★★★★ 3,102,253
Teen
Contains ads

Add to Wishlist Install

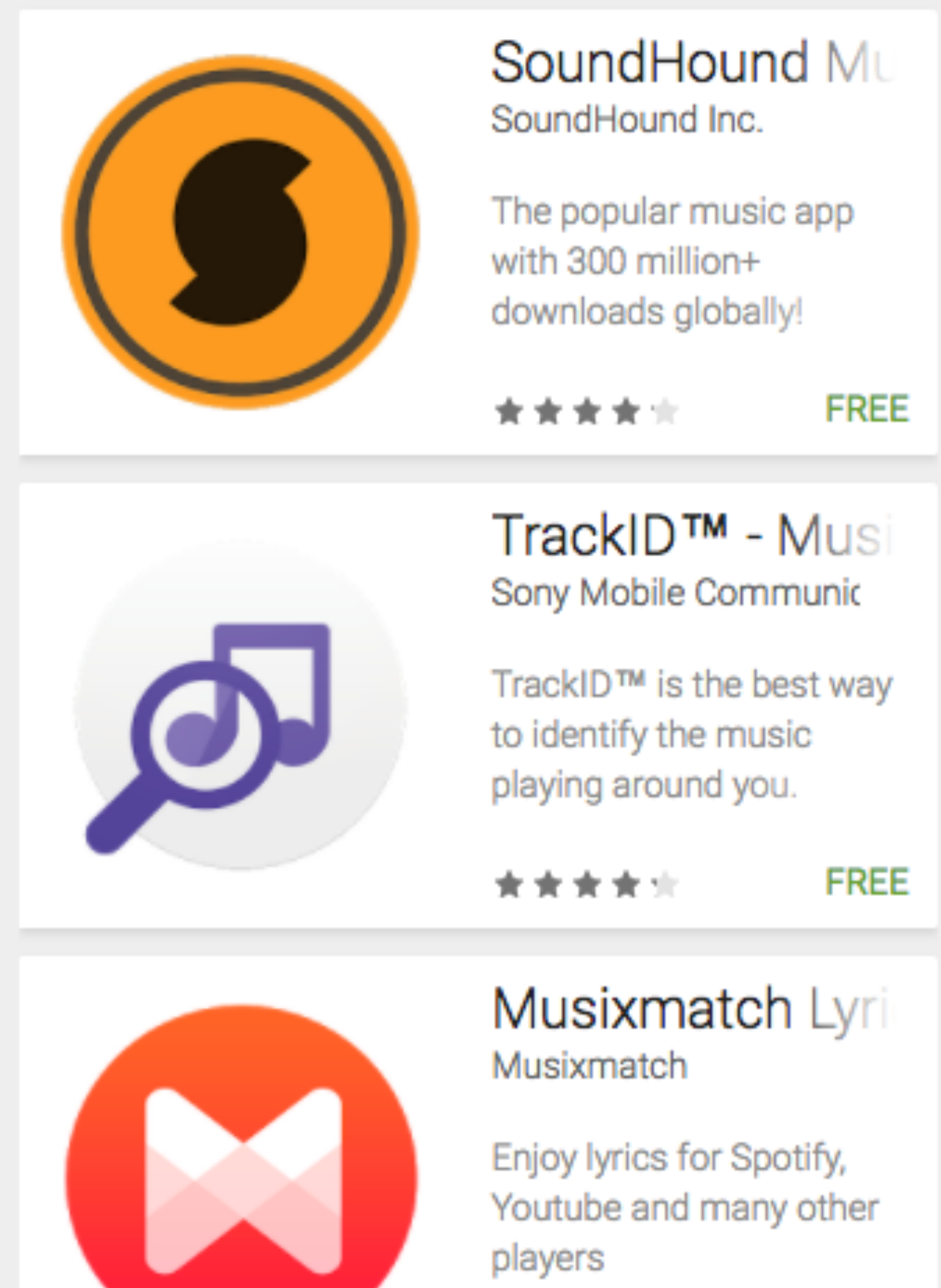
Name any song in seconds.
The world's most popular way to discover music

Explore lyrics, videos & playlists

I'll Take You (Jenaux Remix)
MSTR ROGERS
3,487 streams

Similar

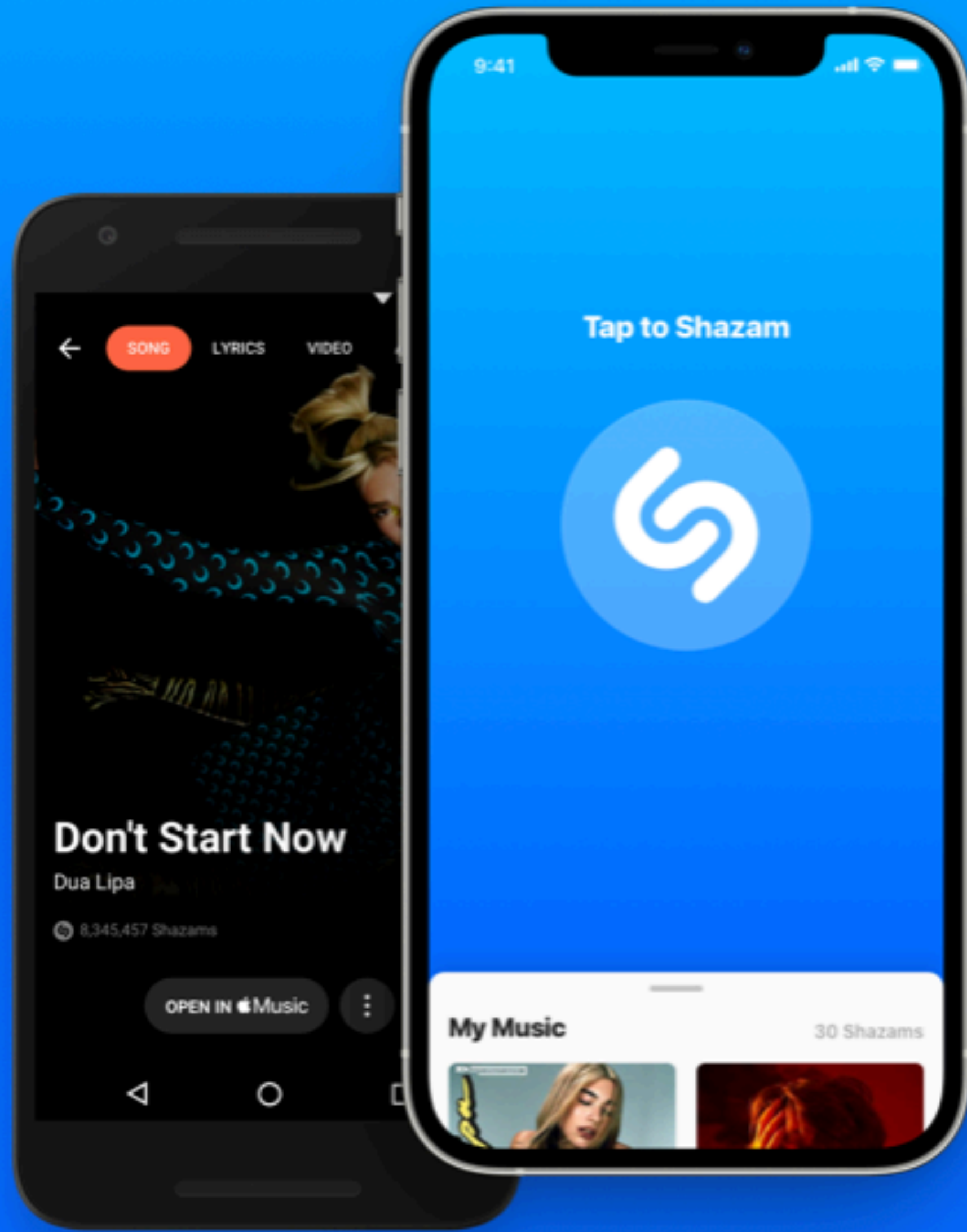
See more



SoundHound Music
SoundHound Inc.
The popular music app with 300 million+ downloads globally!
★★★★★ FREE

TrackID™ - Music
Sony Mobile Communications
TrackID™ is the best way to identify the music playing around you.
★★★★★ FREE

Musixmatch Lyrics
Musixmatch
Enjoy lyrics for Spotify, Youtube and many other players



Name songs in seconds

Find music, concerts and more with Shazam



Get the app

Scan the code with your smart phone camera to download the free app

Available on [iOS](#), [Android](#), and [more devices](#)

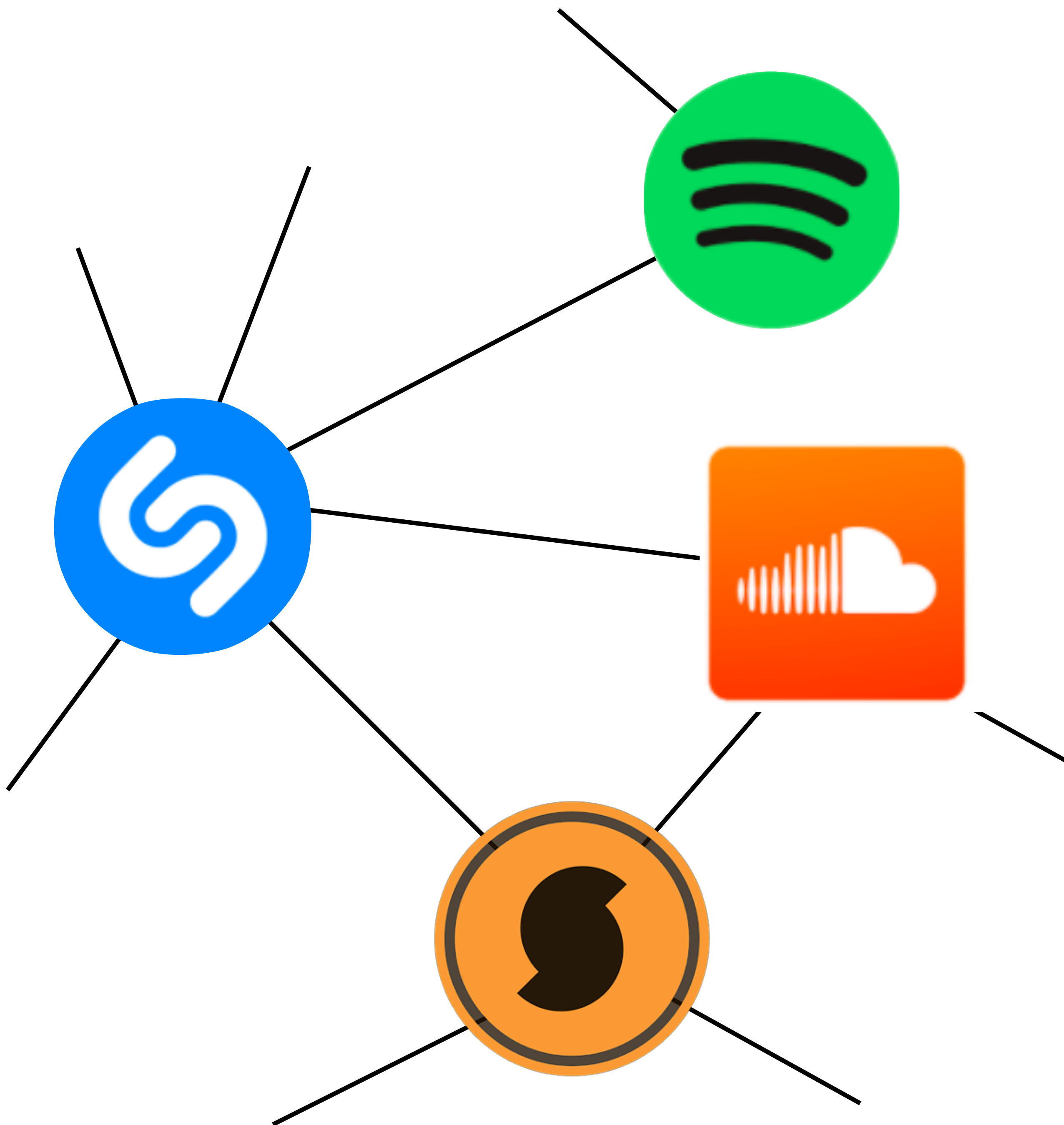


Get up to 2 months free of Apple Music

TRY 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>