http://poloclub.gatech.edu/cse6242

CSE6242 / CX4242: Data & Visual Analytics

Data Collection

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

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

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:
http://poloclub.gatech.edu/cse6242/2017fall/#datasets

Schedule Homework Project Datasets Readings & Resources Prerequisites

All students must first review prerequisites & course expectation.

CSE6242 / CX4242, Fall 2017

Data and Visual Analytics

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)

iTunes

Data that needs scraping

Amazon (reviews, product info)

ESPN

eBay

Google Play

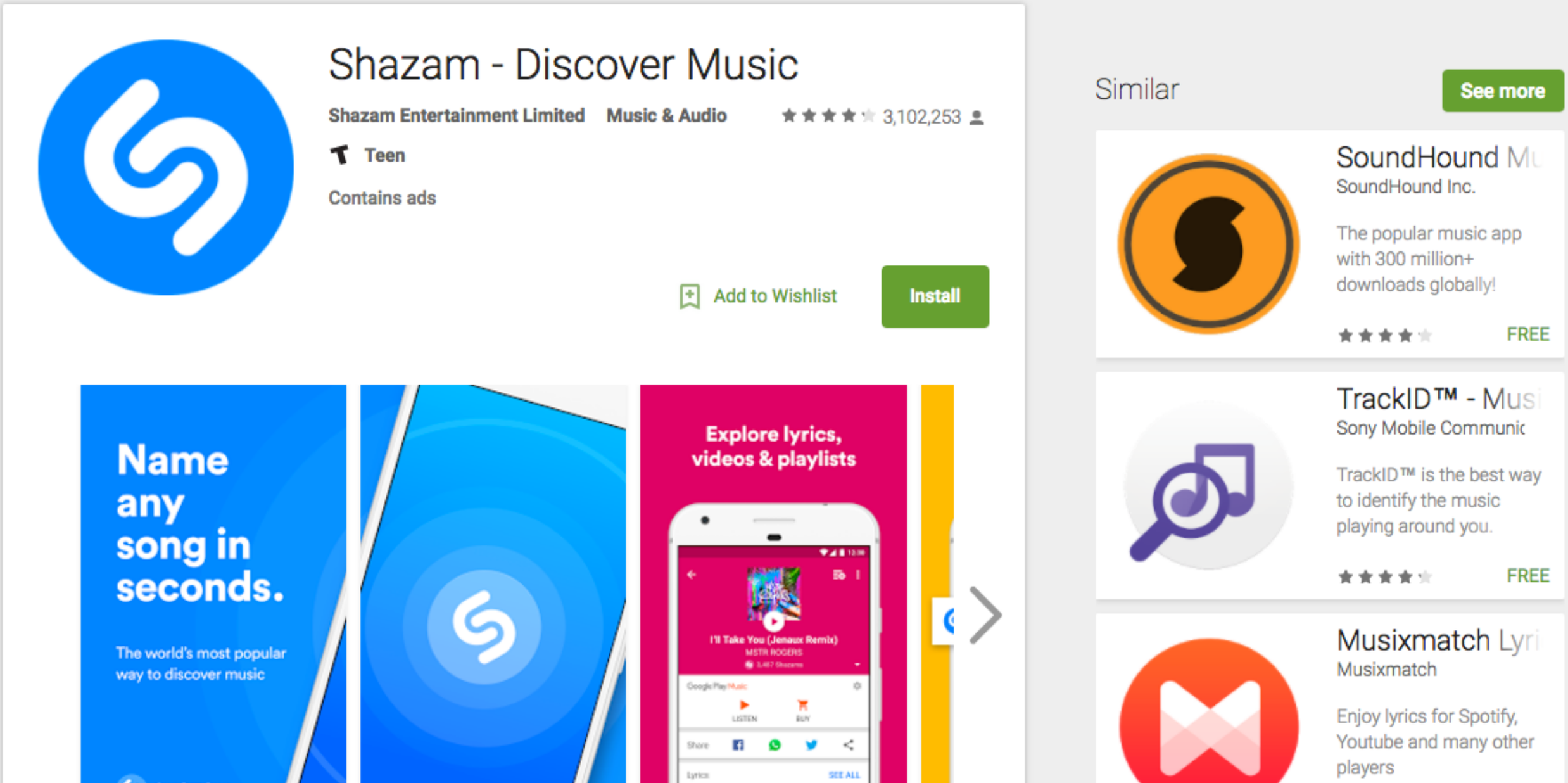
Google Scholar

…

How to Scrape?

Google Play example

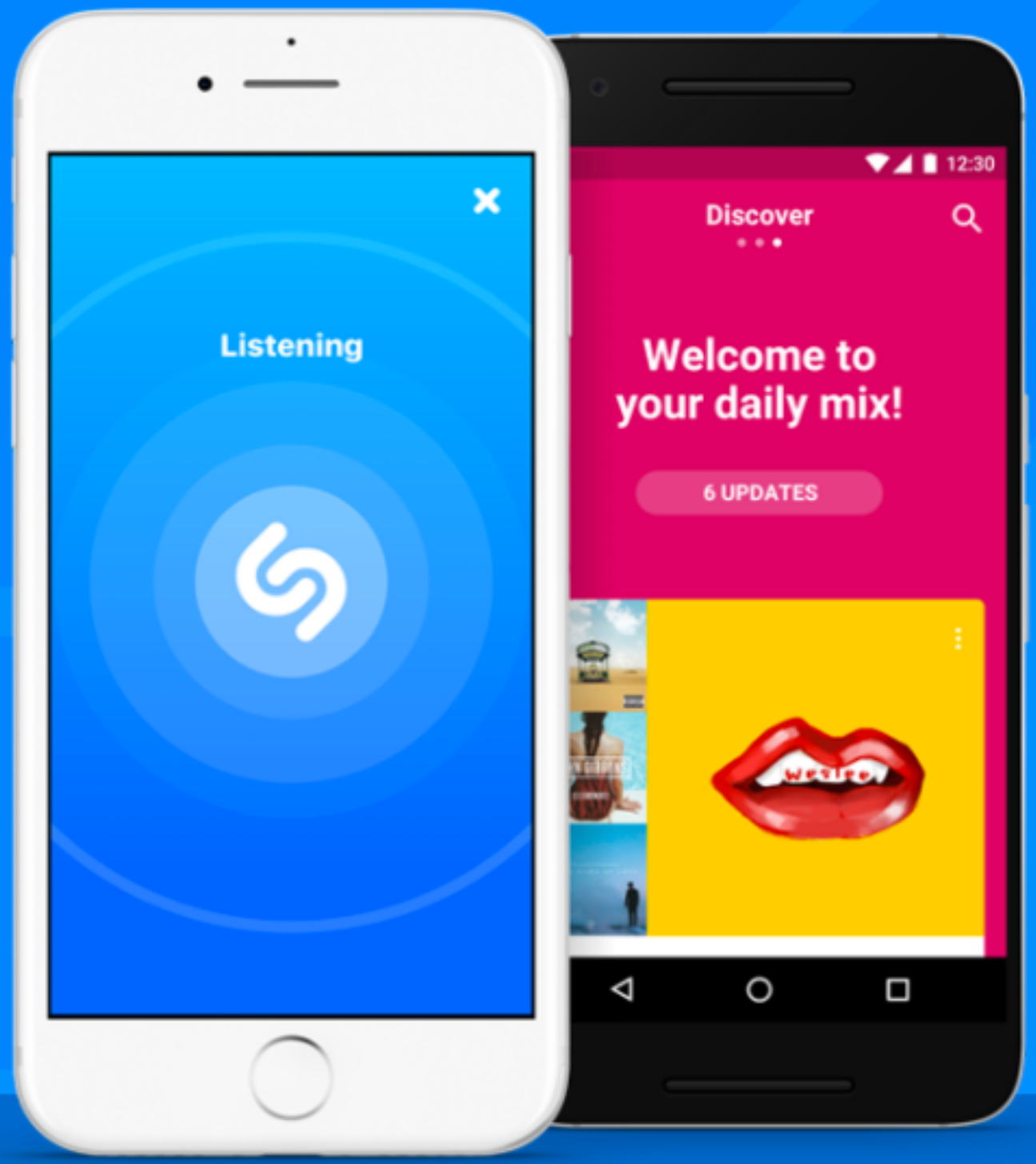
Goal: collect the network of similar apps



Name any song in seconds

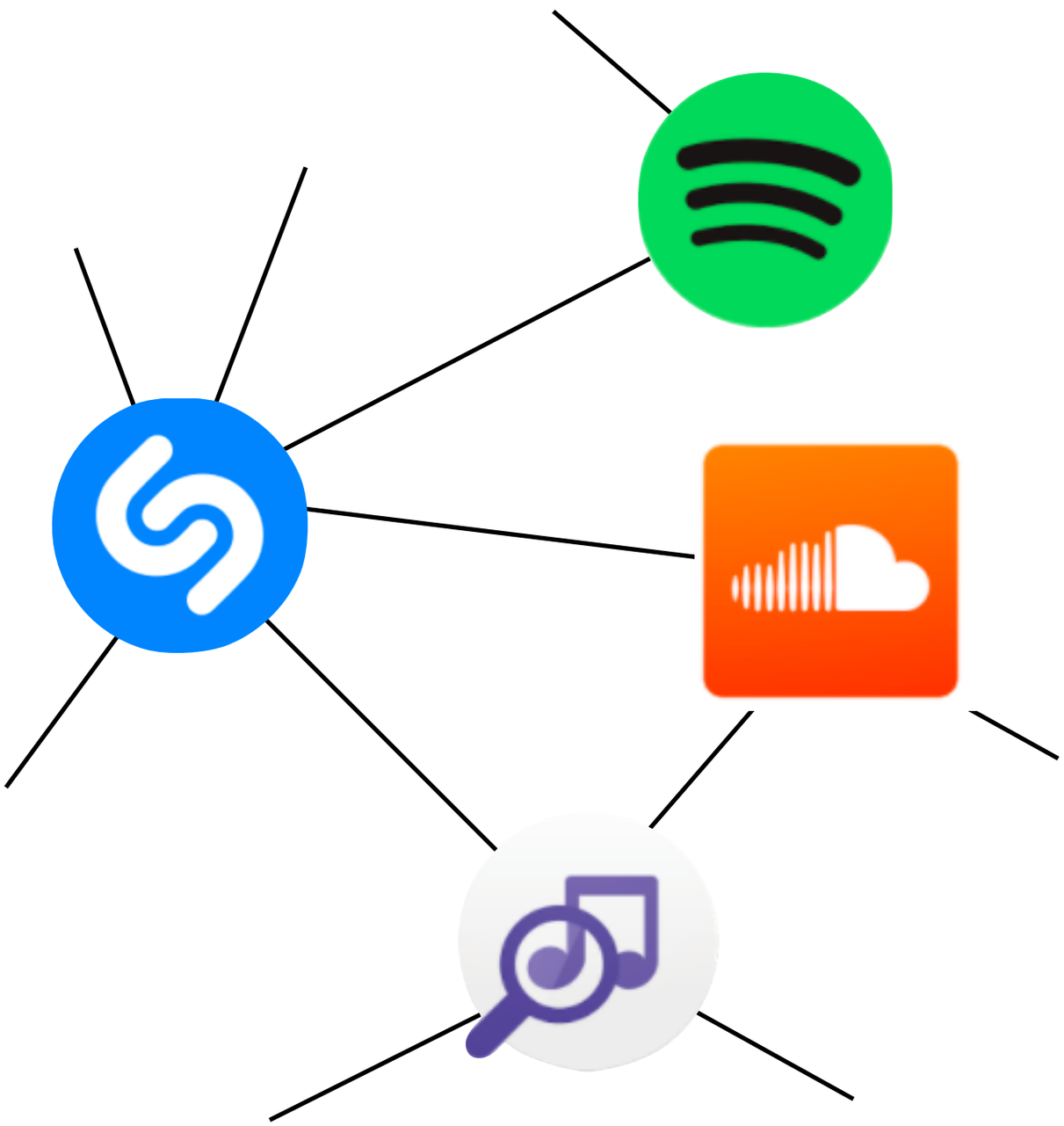
Shazam will identify any music playing around you.

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:
http://www.discoversdk.com/blog/web-scraping-with-selenium
Beautiful Soup supports some.