# Getting Data From the Web

Will Doyle

# Getting Data From the Web

- Most large websites (Facebook, Twitter, Instagram) can best be thought of as huge databases.
- They collect data from user interactions, then generate models to predict consumer activity.
- These models are then used to deliver ads to the users most likely to respond to them.
- This is of course very simplified.
- Bottom line: Data are the lifeblood of many organizations.

# Getting Data From the Web

- Access to data is controlled by its owner.
- You're not allowed to just go and collect data from any website.
- You need permission.
- Use these carefully, and always ask for permission and/or establish that you've been granted permission to the data.

# Sources of Web Data

- Direct download
- Web scraping
- Application Programming Interfaces (APIs)

# Sources of Web Data

- **Direct download**
- **Web scraping**
- Application Programming Interfaces (APIs)

**You are responsible for establishing that you are within the terms of agreement when using these**

**(Check robots.txt)**

# Sources of Web Data

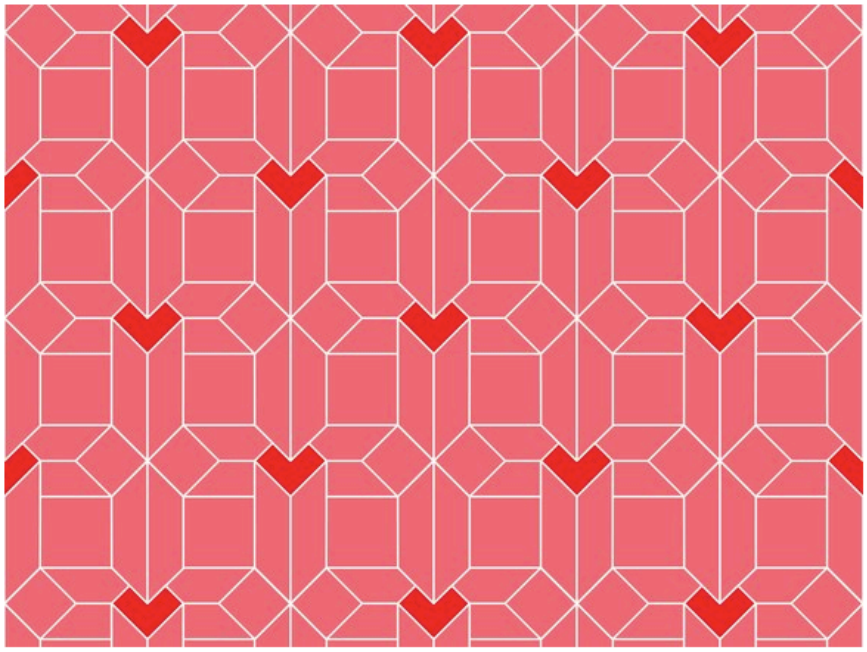
- Direct download
- Web scraping
- **Application Programming Interfaces (APIs)**

**While an API may have some built-in restrictions, you still must read and stay within the terms of service**

# What Not to Do

MICHAEL ZIMMER OPINION 05.14.16 07:00 AM

## OKCUPID STUDY REVEALS THE PERILS OF BIG-DATA SCIENCE



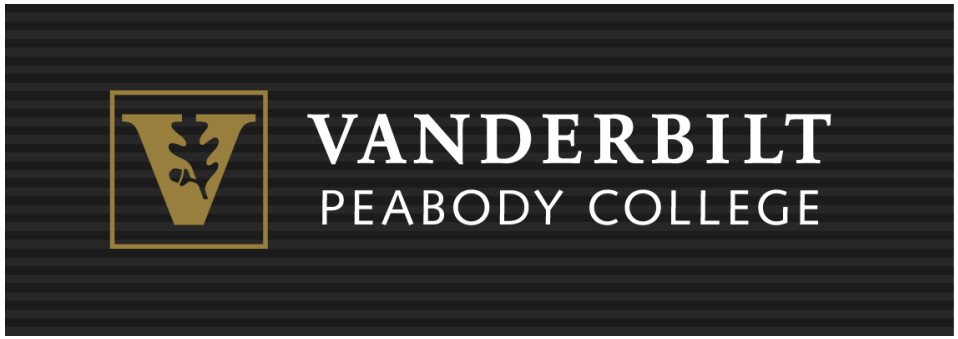
GETTY IMAGES

Source: https://www.wired.com/2016/05/okcupid-study-reveals-perils-big-data-science/

# What Not to Do

"When asked whether the researchers attempted to anonymize the dataset, Aarhus University graduate student Emil O. W. Kirkegaard, who was lead on the work, replied bluntly: 'No. Data is already public.'"

Source: https://www.wired.com/2016/05/okcupid-study-reveals-perils-big-data-science/


VANDERBILT
PEABODY COLLEGE

# APIs

## Concepts

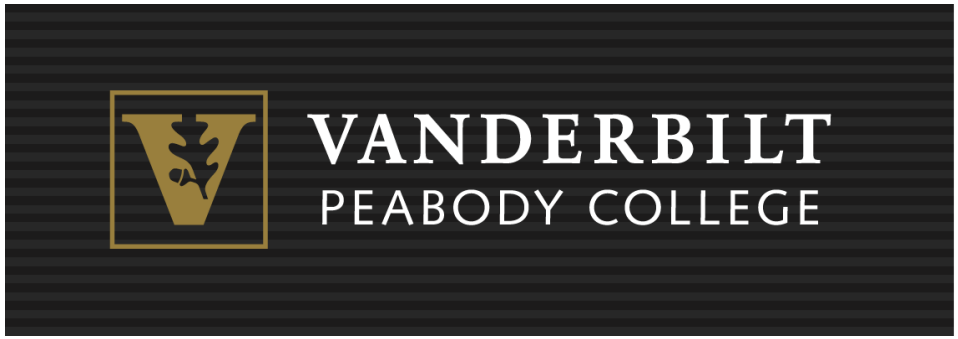
Will Doyle

# Suggested Process for Working With Web Data

- If you're aware of a resource with a large amount of data, look for an API first.
- If it does have an API, then look for a "wrapper" package, like the ACS package for R.
- If it doesn't have a wrapper package, then you'll need to write the request yourself.

# Process Suggestions

- If no API is available, then check robots.txt and terms of service for whether or not you can scrape.
- If it's okay, then you can create a program to scrape and structure data.
- All normal ethics requirements apply (IRB, etc.).


VANDERBILT
PEABODY COLLEGE