SCOTT BAILEY
DATA AND VISUALIZATION LIBRARIAN

SOCIAL MEDIA: COLLECTION, SURVEILLANCE, AND ANALYSIS

Image c/o Electronic Frontier Foundation

WHAT IS SOCIAL MEDIA?

Posts created by users of different online platforms that may include text, images, videos, and other media.

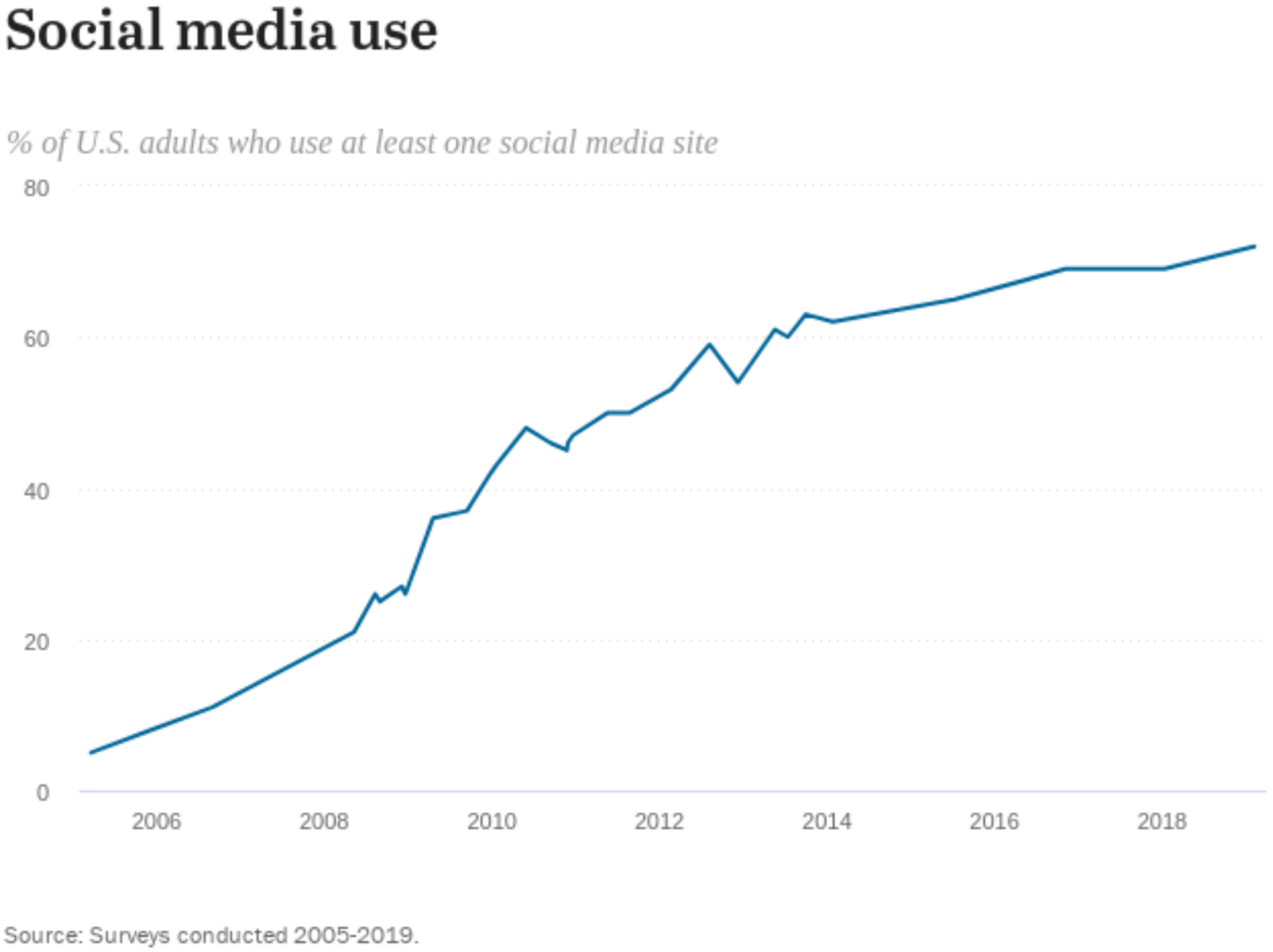
- Some of it is public, some is more or less private.

- Some of it is published independently of other users or previous content, and some is responsive to other content.

EXAMPLES OF SOCIAL MEDIA PLATFORMS

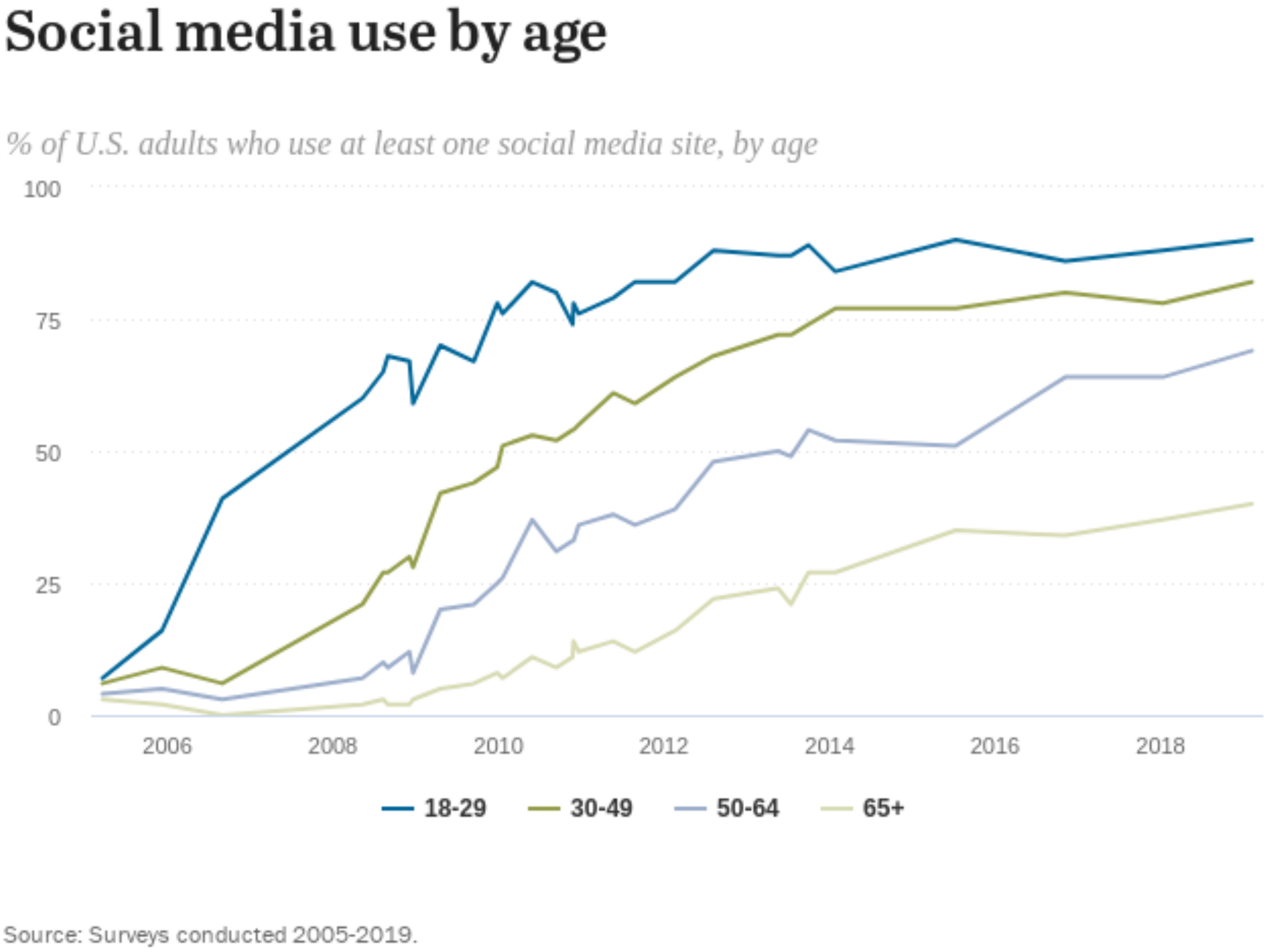
- Facebook
- Twitter
- Instagram
- LinkedIn
- Youtube
- Twitch
- Yelp

AMERICAN SOCIAL MEDIA USE OVER TIME



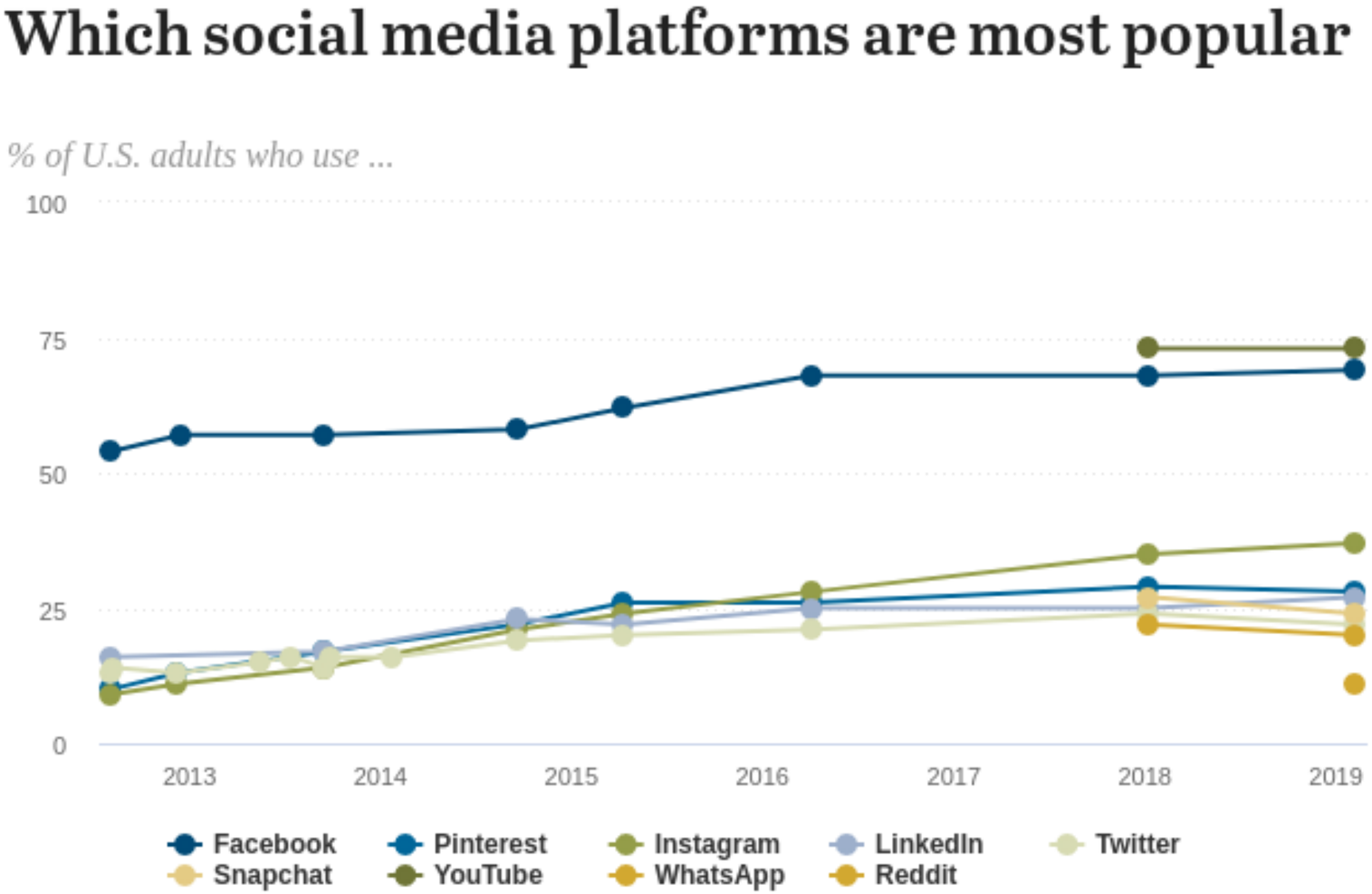
HTTPS://WWW.PEWRESEARCH.ORG/INTERNET/FACT-SHEET/SOCIAL-MEDIA/

AMERICAN SOCIAL MEDIA USE BY AGE



HTTPS://WWW.PEWRESEARCH.ORG/INTERNET/FACT-SHEET/SOCIAL-MEDIA/

AMERICAN SOCIAL MEDIA USE BY PLATFORM



HTTPS://WWW.PEWRESEARCH.ORG/INTERNET/FACT-SHEET/SOCIAL-MEDIA/

HOW MUCH DATA?

ALL DATA FROM OMNICORE AGENCY - HTTPS://WWW.OMNICOREAGENCY.COM

Thousands of Terabytes are created every day across all of the platforms through billions of posts of different types.

Instagram contains over 50 billion uploaded photos.

Youtube has over 5 billion uploaded videos, with many more comments than that.

Twitter has over 500 million tweets sent per day.

WHO COLLECTS SOCIAL MEDIA?

AND WHY DO THEY COLLECT IT?

- Companies
- Academics
- Non-profits
- Governments
- Journalists
- Individuals

HOW DO YOU COLLECT SOCIAL MEDIA DATA?

After the Cambridge Analytica scandal, it’s a lot harder than it used to be, but there are three main ways that exist.

- Official partnerships and research agreements - examples: Facebook, Yelp, Reddit

- Paid services - examples: Twitter and Brandwatch/Crimson Hexagon

- Web scraping or API use - examples: Twitter, Reddit, Youtube

- Use existing datasets that others have collected, which may need to be re-hydrated using APIs - examples: DocNow Catalog - https://catalog.docnow.io/

SOME USE CASES

ACADEMICS

Academics from many disciplines use social media data. Here's a very non-exhaustive list:

- Public health professionals and epidemiologists - heart disease, flu spread, etc.
  - https://softcloudtech.com/twitter-posts-predict-rates-of-heart-disease/
  - https://web.archive.org/web/20160309220440/http://web.natur.cuni.cz/~houdek3/papers/Eichstaedt%20et%20al%202015.pdf
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6615001/
- Environmental scientists - detecting spread of bird and plant diseases
- Linguists studying changing character of language and difference between language in different media
- Political scientists
- Sociologists and ethnographers

COMPANIES + GOVERNMENTS: PALANTIR

Palantir is a private company that aggregates data from many sources, including social media data, for use by governments and businesses, often for military and police action.

- https://www.vice.com/en_us/article/9kx4z8/revealed-this-is-palantirs-top-secret-user-manual-for-cops

- https://www.theverge.com/2020/4/21/21230453/palantir-coronavirus-trump-contract-peter-thiel-tracking-hhs-protect-now

COMPANIES: BRANDWATCH/CRIMSON HEXAGON

CrimsonHexagon, recently merged with Brandwatch, offers rich access to social media data, such as Twitter data, and built in analytics to companies and academics.

NCSU does pay for access to brandwatch, and researchers from across the university use the media data to understand business patterns, branding, politics, and more.

Marketing firms and businesses also use the data to better tailor their marketing strategies and to target consumers.

https://www.brandwatch.com/

ETHICS OF SOCIAL MEDIA COLLECTION

- Terms of use don’t always align with the goals and ethical frameworks of users.
  - Ex. Twitter and deleted tweets; historians and archivists of the cultural record
- Privacy - do the people who post publicly understand what they are doing? Do companies sufficiently inform users of what could be done with their data?
- Reliability and authenticity - given how hard it can be to verify identities and information in social media platforms, how well can it be used for analysis and prediction? Should researchers put forward fast moving and potentially beneficial analysis when it depends on unreliable or unverified data?

CAN COLLECTING SOCIAL MEDIA DATA EVER BE NEUTRAL?

WHO SHOULD BE ABLE TO COLLECT DATA? WHO SHOULD BE ABLE TO CONTROL HOW IT IS USED?

OUR ACTIVITY TODAY: COVID-19 TWEETS

Activity goal:

In our activity, we're going to explore some Twitter data that has already been collected, with an eye toward what we can quickly and easily learn from the data using R.

Data source: https://zenodo.org/record/3842180

This is a dataset compiled by researchers at the University of Southern California (USC), and which is being updated regularly.

Why can't we just scrape Twitter data quickly and easily right now?

R

A FREE AND OPEN SOURCE LANGUAGE FOR STATISTICAL COMPUTING

R has become one of two main coding languages for “data science” work, both in the academy and in industry. It’s a language that was specifically built for statistical computing, but has really taken off in the last several years due to the development of several packages that made it easier for people with little to no coding experience to jump in and just do data analysis and visualization without becoming an expert developer.

The easiest way to work in R is using RStudio: https://rstudio.com/

We’ll use their entirely online environment, RStudio Cloud, to work with our data.