

Introducing Digital Behavioral Data

Weiai Xu (Wayne), PhD

Assistant Professor

Department of Communication, UMass-Amherst

Email: weiaixu@umass.edu

curiositybits.cc



Weiai Wayne Xu

Assistant Professor in
Computational Communication
[University of Massachusetts - Amherst](#)



About Me

curiositybits.cc

Weiai Xu (call me Wayne)

Assistant Professor

Department of Communication

Office Hours: 11 am-1 pm Tues. & Thur. **Or by appointment**

Office: N334 ILC

weiaixu@umass.edu

Research: I write computer codes and crunch numbers to study online communities



Weiwei Wayne Xu

Assistant Professor in
Computational Communication

University of Massachusetts -
Amherst



About Me

curiositybits.cc



Two goals

Goal number one:

- **Practical data science & computer programming skills.** Use the R language for data mining, analytics, and building visualization apps.

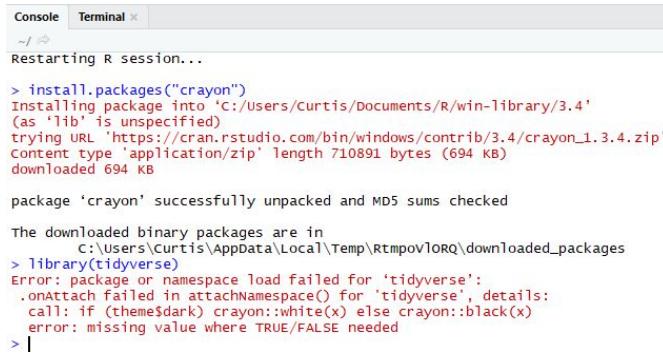
Goal number one:

- **Be a critical user of algorithms and digital behavioral data.** Understand the biases and undesirable consequences they may have on the civil society.

This course does not assume prior knowledge of computer programming. But you will learn to apply R codes.



Coding is 10% intelligence and 90% endurance.



The screenshot shows the RStudio interface with the 'Console' tab selected. The session starts by restarting the R session. It then attempts to install the 'crayon' package, which succeeds. However, when trying to load the 'tidyverse' package, it fails with an error message about a missing value where TRUE/FALSE is needed. The session ends with a single character '|'. The console output is as follows:

```
Console Terminal <br/>~ / ~<br/>Restarting R session...<br/>> install.packages("crayon")<br/>Installing package into 'C:/Users/Curtis/Documents/R/win-library/3.4'<br/>(as 'lib' is unspecified)<br/>trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.4/crayon_1.3.4.zip'<br/>Content type 'application/zip' length 710891 bytes (694 KB)<br/>downloaded 694 KB<br/>package 'crayon' successfully unpacked and MD5 sums checked<br/>The downloaded binary packages are in<br/>  C:\Users\Curts\AppData\Local\Temp\RtmpoVlORQ\downloaded_packages<br/>> library(tidyverse)<br/>Error: package or namespace load failed for 'tidyverse':<br/>.onattach failed in attachNamespace() for 'tidyverse', details:<br/>call: if (theme$dark) crayon::white(x) else crayon::black(x)<br/>error: missing value where TRUE/FALSE needed<br/>> |
```

Turn every error message into a learning opportunity

Four modules

1. Hands-on workshops
2. Online tutorials: <https://curiositybits.cc/tutorial/>
3. Peer-to-peer learning: find a class partner. You two will discuss any error and difficulty that either of you has encountered. You will learn to debug code collaboratively
4. In-class discussions (on topics related to the societal impacts of data and algorithms)

Four modules

CuriosityBits Creative Commons

Home Tutorials Posts Publications Contact CV •Light• ⚡ 🔍 ☾

COMM497DB Fall 2019

Access [the interactive tutorial](#)

- Download course syllabus
- Week1: [Slide: Introducing Digital Behavioral data](#)
- Week1: [Slide: An Introduction to R](#)
- Week1: [Practice code](#)
- Week2: [Slide: Application Programming Interfaces \(API\)](#)
- Week2: [Slide: Collect Twitter Data](#)
- Week2: [Practice code for API](#)
- Week2: [Practice code for Collecting Twitter Data part 1](#)
- Week3: [Slide: Twitter Data Explained](#)
- Week3: [Practice code for Collecting Twitter Data part 2](#)
- Week4: [Slide: Visualizing Twitter Virality](#)



An interactive tutorial for COMM 497DB

Weiwei Wayne Xu

Libraries/packages

Data frames

Connecting to the Twitter API

Collect tweets by keywords/hashtags

Collect Twitter user timeline

Collect Twitter user info

Make Wordclouds

Predict Ideology (in progress)

Start Over

Using R for Digital Behavior Analytics

Libraries/packages

What is a library/package? Think of R as an operating system (e.g., iOS, Windows) and a library/package as an app running on the system. Each library is designed to accomplish specific tasks. For example, the library *ggplot2*—which is a library we will use throughout the semester—is for visualizing data, and the library *rtweet* is used for collecting Twitter data.

Use **installed.packages()** to install libraries. Use **library()**, or **require()** to load an installed library.

Next, we will install a fun library called *cowsay*.

Code

⟳ Start Over

▶ Run Code

```
1 installed.packages("cowsay") #make sure the library name is wrapped by quotation.  
2  
3 library(cowsay) #load the library, alternatively, you can use require(cowsay)
```

This tutorial is hosted on a cloud server, running the above code won't have an effect on your local machine. Put the code in RStudio and run it on your local machine. Keep an eye on what is happening in the Console.

Let's have some fun with cowsay.

Run the code and see what happens.

Code

⟳ Start Over

▶ Run Code

```
1 library(cowsay)  
2 say("Hello! Welcome to COMM 497DB")  
3
```

Evaluation

1. Preparation and debugging (15%)
2. Assignments (35%)
3. Final project (40%): an example

<https://carl-macdonald.shinyapps.io/2016RussianMeddlingABiggerPicture/>

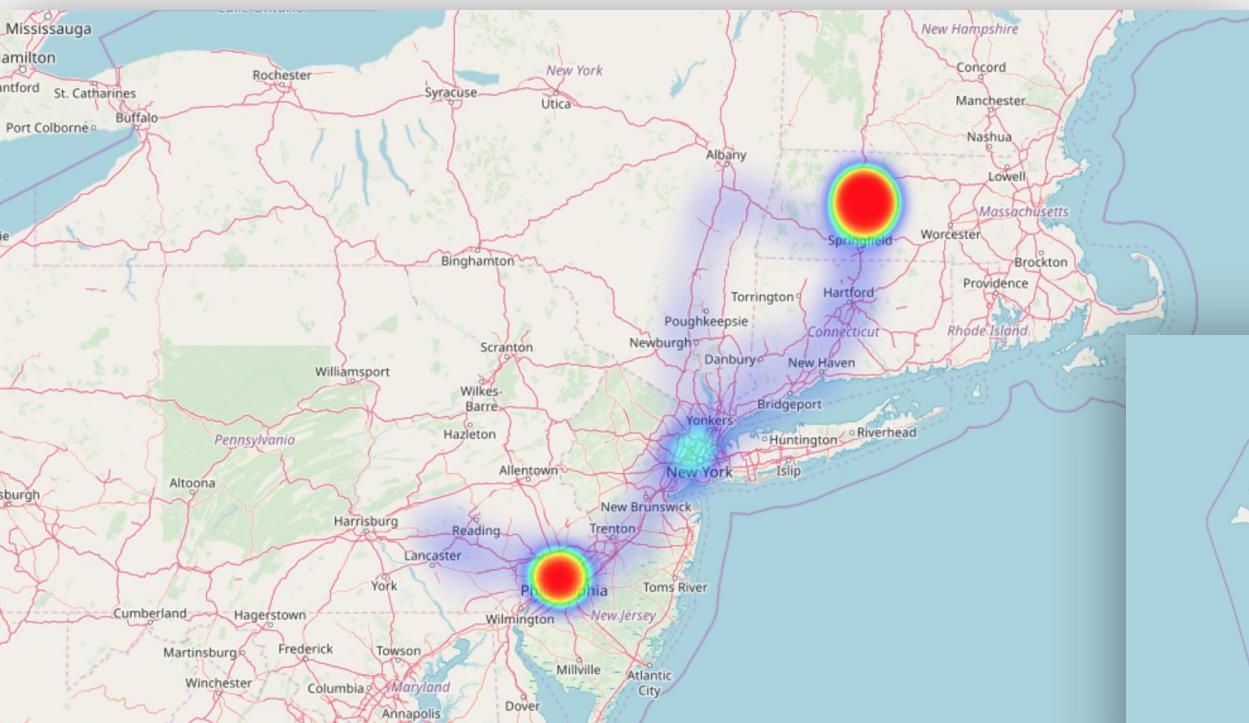
4. Participation (10%)

Digital behavioral data?

Take a look at your phone or laptop, what sort of digital behavioral data does your device have on you?



My digital footprint = my actual footprint



Digital footprints

At any second, a massive amount of user-generated content is being produced and recorded. They constitute our digital footprints that reveal how we think and live.

Three Vs

Volume, velocity, and variety



8,342 Tweets sent in 1 second



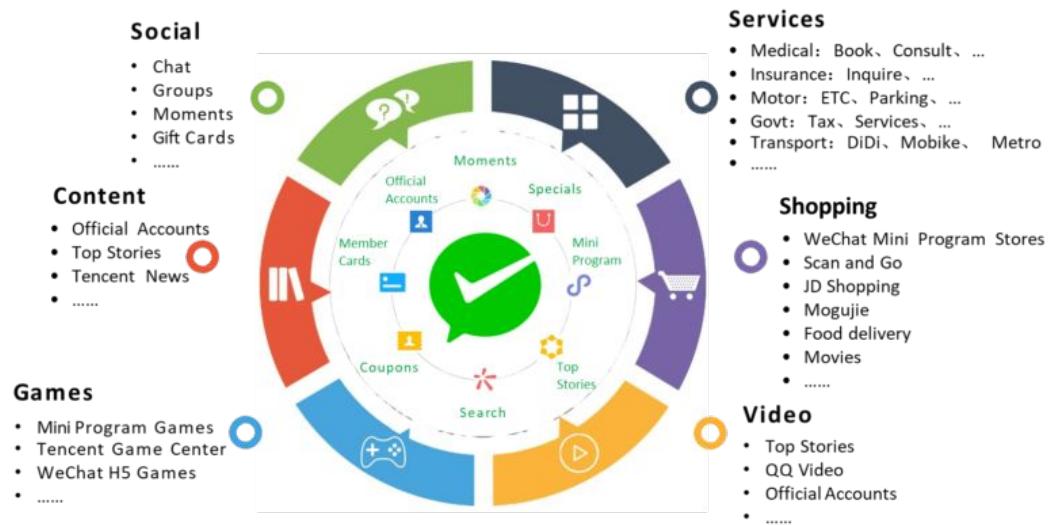
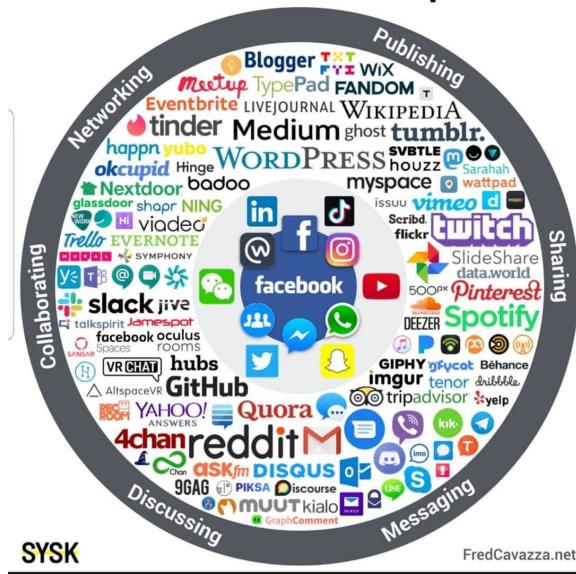
503,057 Tweets since opening this page
0:01:00 seconds ago

A screenshot of the first tweet ever posted on Twitter. The tweet reads "just setting up my twttr" and was posted at 12:50 PM on March 21st, 2006, from a web browser. The tweet is attributed to "jack" (Jack Dorsey) with a small profile picture.

The first ever tweet

We will primarily rely on public Twitter data, but Twitter constitutes only a small fraction of the giant global social media landscape

Social Media Landscape 2019



~~Three~~ Vs Four Vs

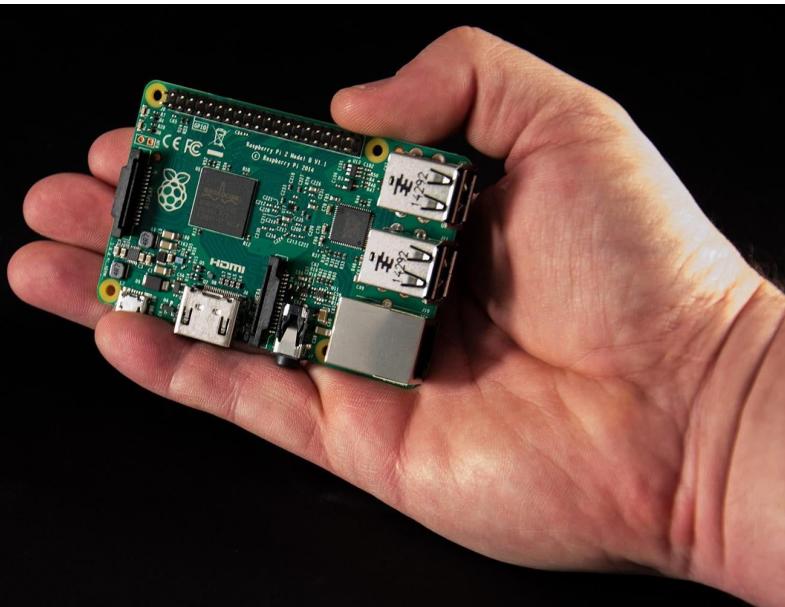
Volume, velocity, and variety **+ value**

Business value: data vendors, business analytics

Political value: Cambridge Analytica, facial recognition for social control

Social value: predictive policing, data-driven humanitarian aids, smart city

More powerful and affordable computing resources



Micro-computers such as
Raspberry Pi <\$100



Cloud computing

Digital footprints

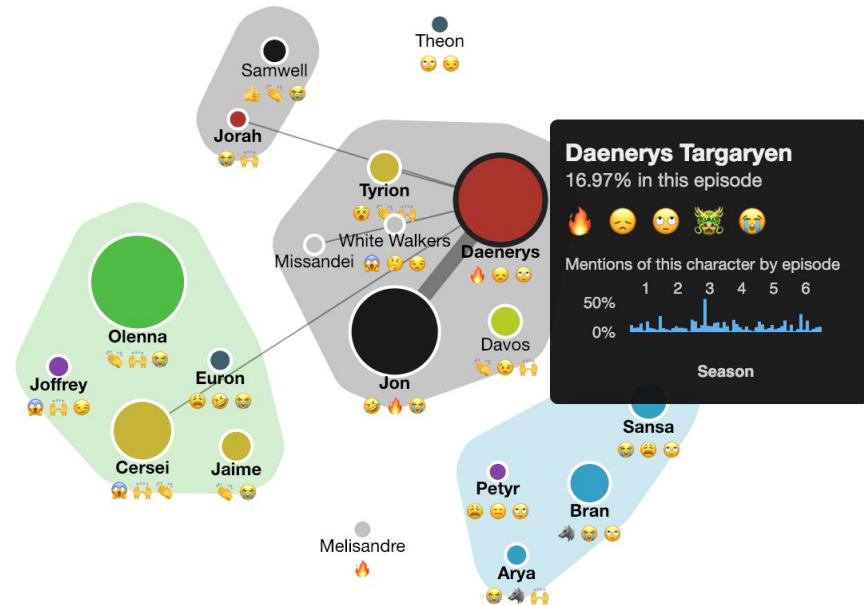
Three examples

- [How every #GameOfThrones episode has been discussed on Twitter](#)
- [How Latinos in the U.S. connect with Latin America on Twitter](#)
- [3 Million Russian Troll Tweets](#)
-

Dorne Dothraki Ironborn King's Landing Lannister Neutral Night's Watch North Others Reach Riverlands
Stormlands Targaryen Vale Wildlings

#GameOfThrones

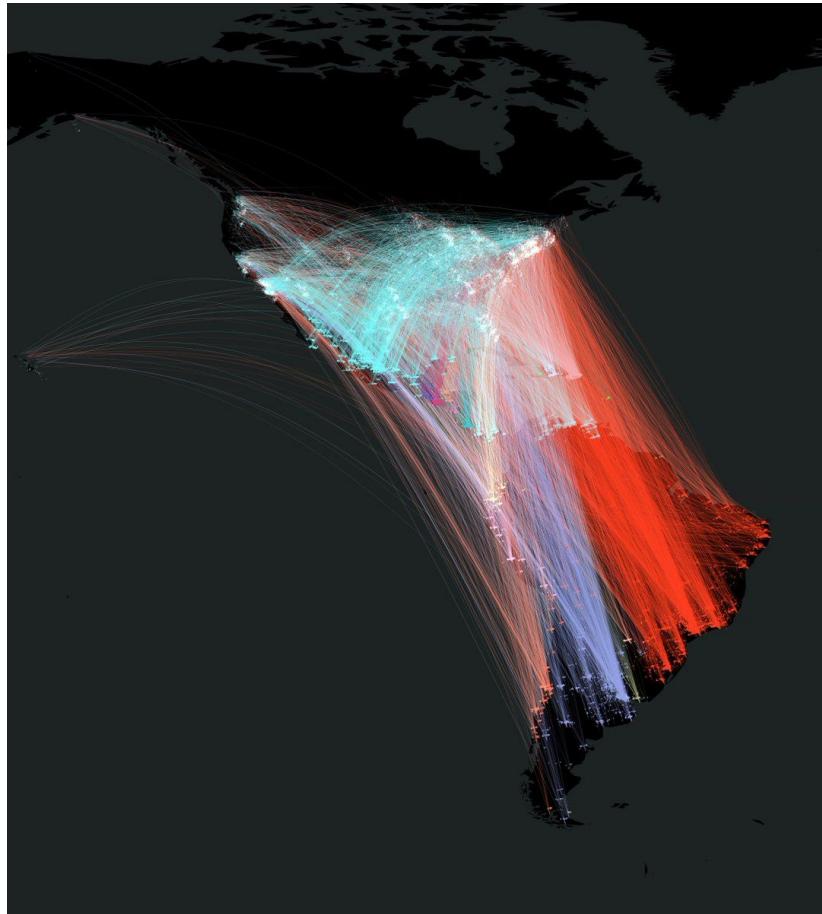
Q: How do characters relate to each other?



<https://interactive.twitter.com/game-of-thrones/#?episode=63>

Latinos on Twitter

Q: How do Twitter users relate to each other?

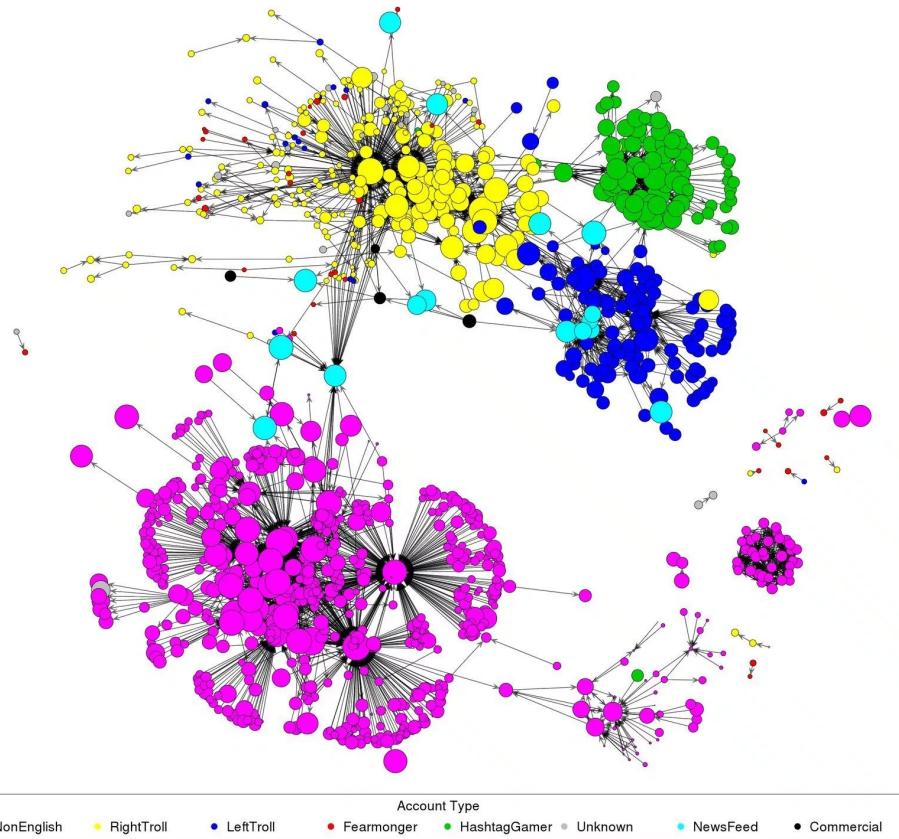


<https://twitter.com/TwitterAlas/status/712058816041345024>

Russian Trolls

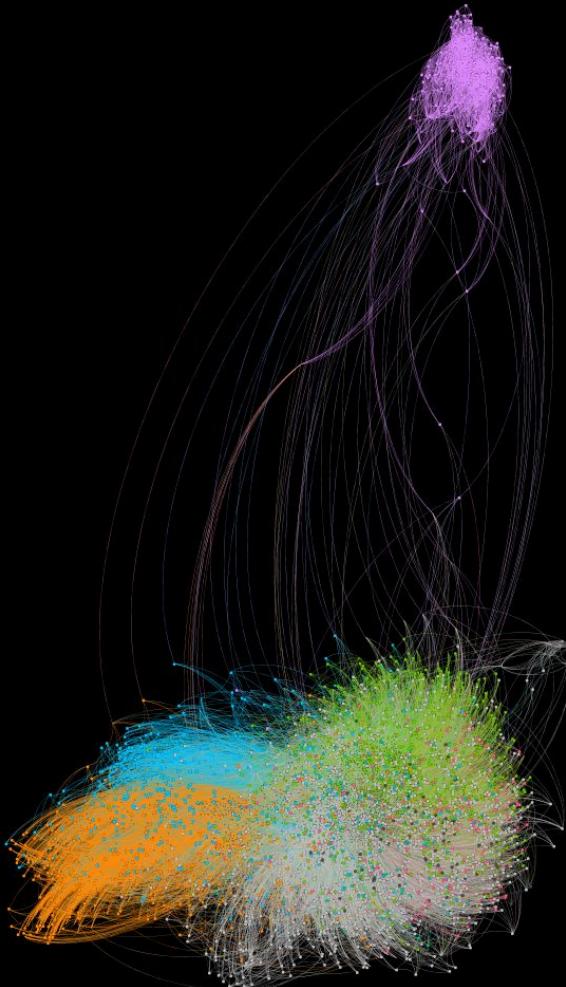
Q: How do Twitter users relate to each other?

Q: What sort of patterns can you identify from this graph?



An example from my own project

Q: What sort of questions
would you ask about this
graph?



Two types of data

Content data: what we can see

The what

Donald J. Trump Following

There are now 77 major or significant Walls built around the world, with 45 countries planning or building Walls. Over 800 miles of Walls have been built in Europe since only 2015. They have all been recognized as close to 100% successful. Stop the crime at our Southern Border!

7:33 AM - 16 Jan 2019

31,026 Retweets 120,471 Likes

32K 31K 120K

Tweet your reply

Eugene Gu, MD @eugenegu · 9h
Replying to @realDonaldTrump

Walls are symbols of hate and division. That's why Ronald Reagan, hero of like all Republicans, told Mr. Gorbachev to tear down that wall. You know, the one in Berlin.

822 543 5.1K

What's in digital behavioral data

Metadata: The data of data

The where, when, who, how, and etc.

Donald J. Trump @realDonaldTrump

Following

There are now 77 major or significant Walls built around the world, with 45 countries planning or building Walls. Over 800 miles of Walls have been built in Europe since only 2015. They have all been recognized as close to 100% successful. Stop the crime at our Southern Border!

7:33 AM - 16 Jan 2019

31,026 Retweets 120,471 Likes

32K 31K 120K

Eugene Gu, MD @eugenegu · 9h

Replying to @realDonaldTrump

Walls are symbols of hate and division. That's why Ronald Reagan, hero of like all Republicans, told Mr. Gorbachev to tear down that wall. You know, the one in Berlin.

822 543 5.1K

Why metadata matters

NSA triples metadata collection numbers, sucking up over 500 million call records in 2017

Devin Coldewey @techcrunch / 9 months ago

 Comment

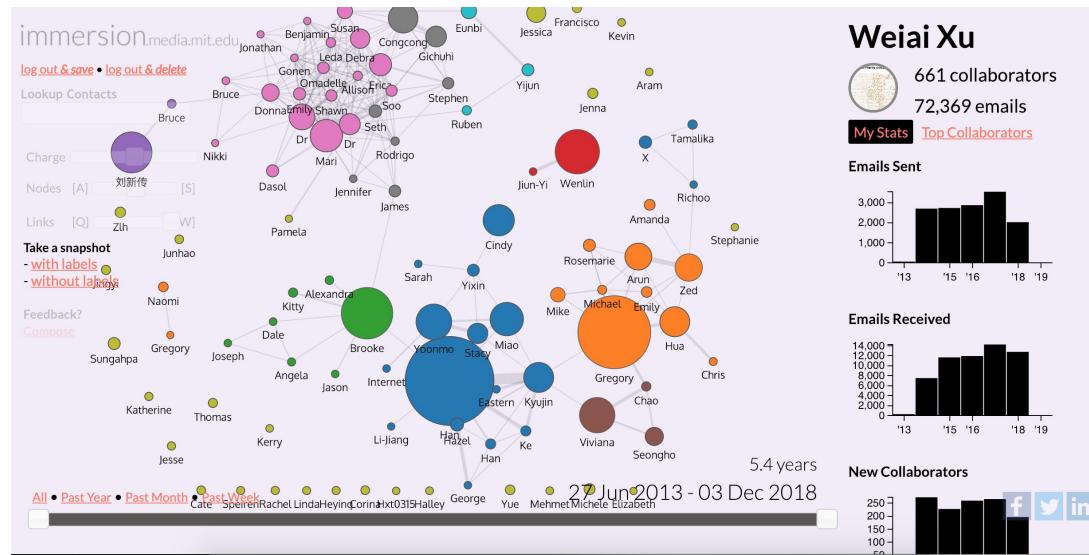


NSA is *not* tapping your phone calls. It is *not* collecting content data. But it does collect metadata, that is, which numbers were called and when, the duration of the call and so on.

Metadata can reveal a lot about our lives!

Metadata from your digital life
Try Immersion (immersion.media.mit.edu)

Q: What metadata do you think Immersion will collect from your Gmail account?



Try Immersion

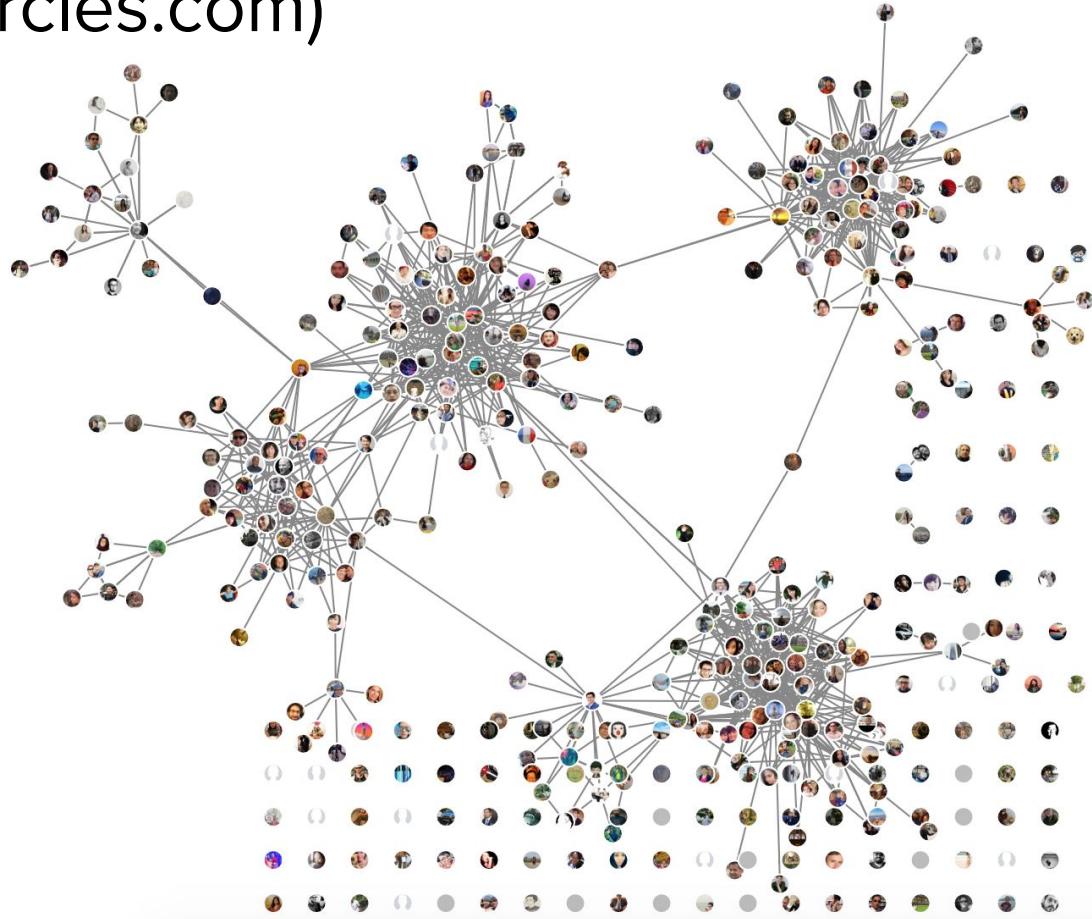
3. What information from my inbox does Immersion collect?

To create your visualization Immersion collects only the metadata (*From, To, Cc and Timestamp*) of emails. Immersion does not access the subject or body of any of your emails. Due to the architecture of the underlying protocol, technically Immersion could access any part of your email, but we chose not to. Once again, this is true of all email clients. Privacy, and users' ownership of their data, are very important to us, so we've designed Immersion to prioritize the privacy of users and their ability to control their own data.

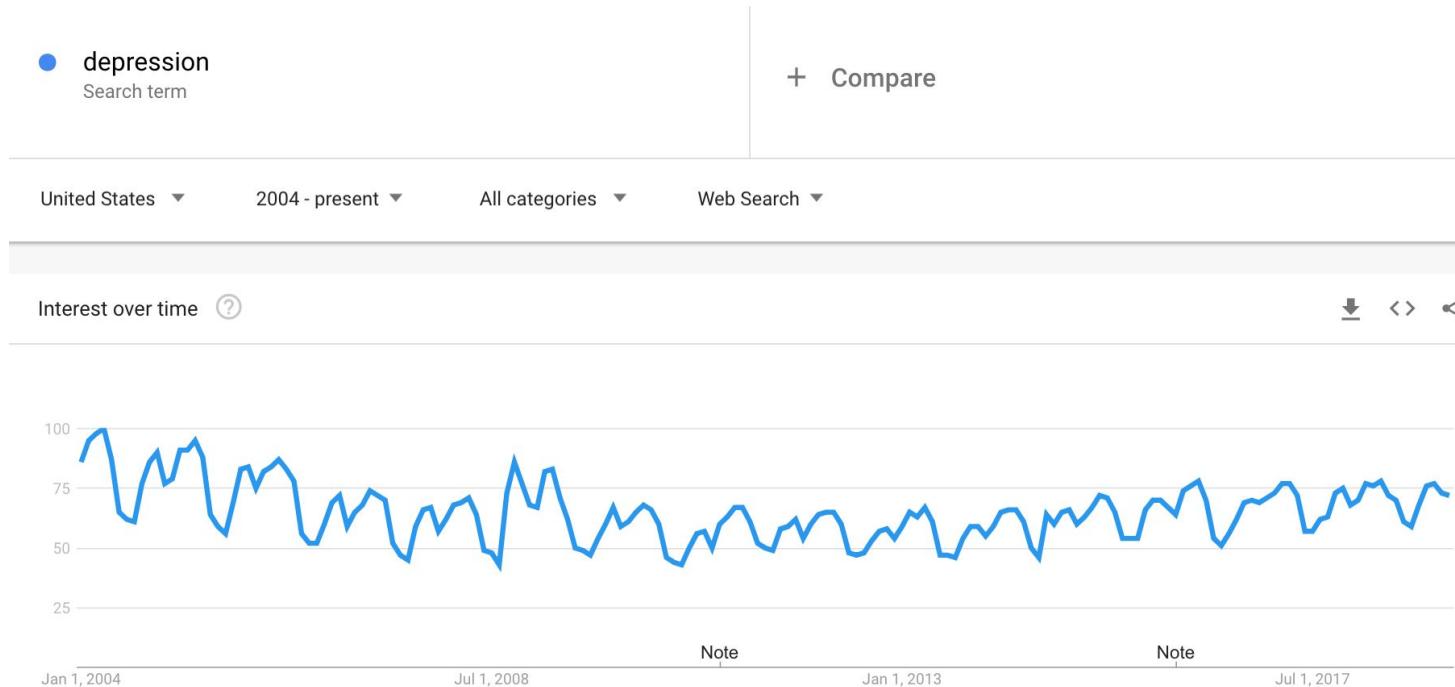
Q: Are you comfortable with Immersion's data collection?

Try Lost Circles (lostcircles.com)

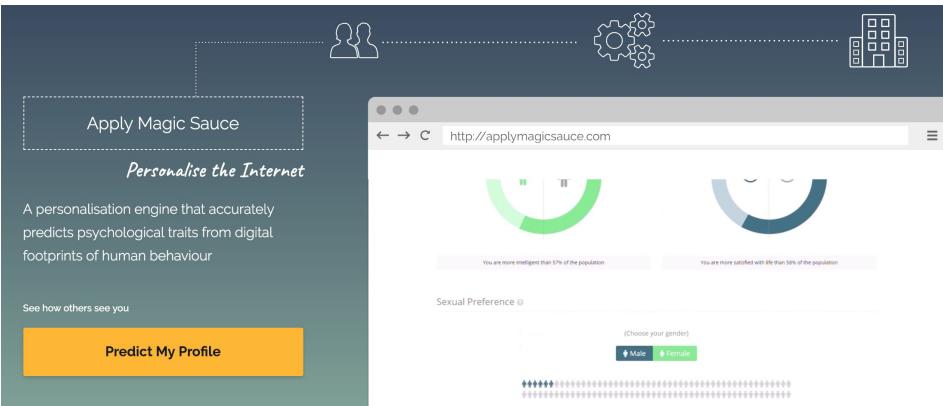
Q: This is my Facebook friendship network. What does the graph say about my life?



Try Google Trends (trends.google.com)



Try Magic Sauce (<https://applymagicsauce.com/>)



Private traits and attributes are predictable from digital records of human behavior

Michał Kosinski, David Stillwell, and Thore Graepel

PNAS April 9, 2013 110 (15) 5802-5805; <https://doi.org/10.1073/pnas.1218772110>

Edited by Kenneth Wachter, University of California, Berkeley, CA, and approved February 12, 2013 (received for review October 29, 2012)

Article

Figures & SI

Info & Metrics

PDF

Remember the Cambridge Analytic scandal? This app called *Magic Sauce*, developed by Cambridge University's Psychometrics Centre, is said to have laid the groundwork for Cambridge Analytic's use of Facebook data for political campaign.

Summarizing digital behavioral data

Based on your experience with the above visualization apps, what do you have to say about the three Vs about big data?

Volume, velocity, and variety + value

We should be concerned about

veracity, validity, etc

Toolkits for exploring digital behavioral data

According to thinkdigitalfirst.com



10 Free Twitter Analytics Tools For 2018

- Hootsuite. Hootsuite has been my favourite Social Media scheduling **tool** for a number of years. ...
- Buffer. Similar to Hootsuite in its ability to automate content, Buffer is a **great** option when looking at the **best** times to post your content on **Twitter**. ...
- Twitonomy. ...
- Socialert. ...
- Klear. ...
- Klout. ...
- ManageFlitter. ...
- Tweepi. ...

More items... • Jan 3, 2018

10 Free Twitter Analytics Tools For 2018 | Think Digital First
<https://www.thinkdigitalfirst.com/2018/01/03/10-free-twitter-analytics-tools-2018/>

There are plenty of web-based Twitter Analytics tools. The list keeps growing, underlying a market demand for data analytics.

See a list:

<https://www.thinkdigitalfirst.com/2018/01/03/10-free-twitter-analytics-tools-2018/>

Why we are NOT using them?

1. They are too basic;
2. They are limited to analyzing a small number of accounts; not designed for the big data era
3. They are NOT free;
4. They are foolproof tools, so using them won't help your analytical thinking;
5. The algorithm used by the web-based analytics tools (if any) is opaque. You don't know what sauce they use in cooking results.

There are more advanced data analytics softwares

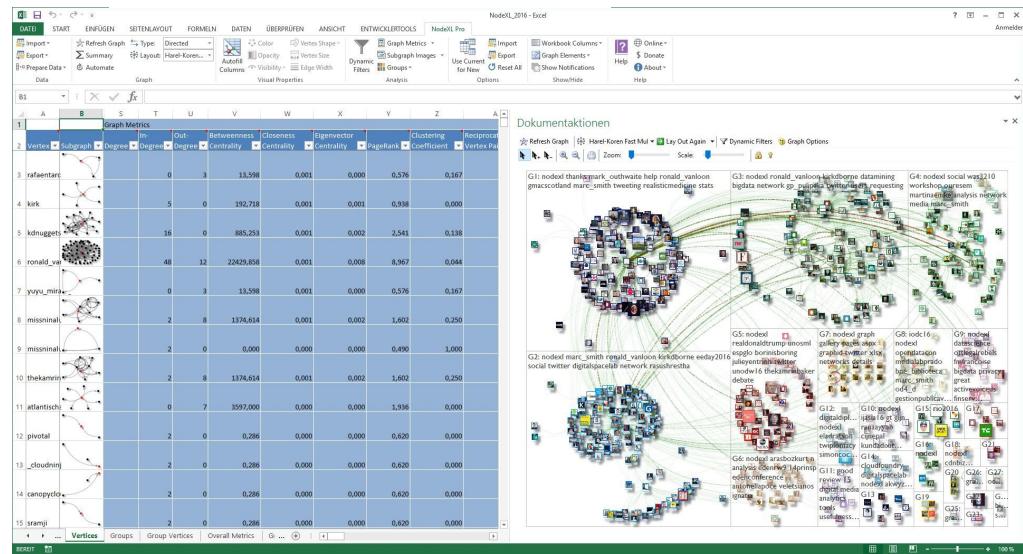
About Our Text Data Science Software

Collaborative text analytics and Gnip PowerTrack Twitter data

With dozens of powerful text analytics, data science, human coding, and machine-learning features, including instant access to the [Gnip PowerTrack 2.0 for Twitter](#) and the [free Twitter Search API](#), DiscoverText provides cloud-based software tools to quickly evaluate large amounts of text, survey, and Twitter data.

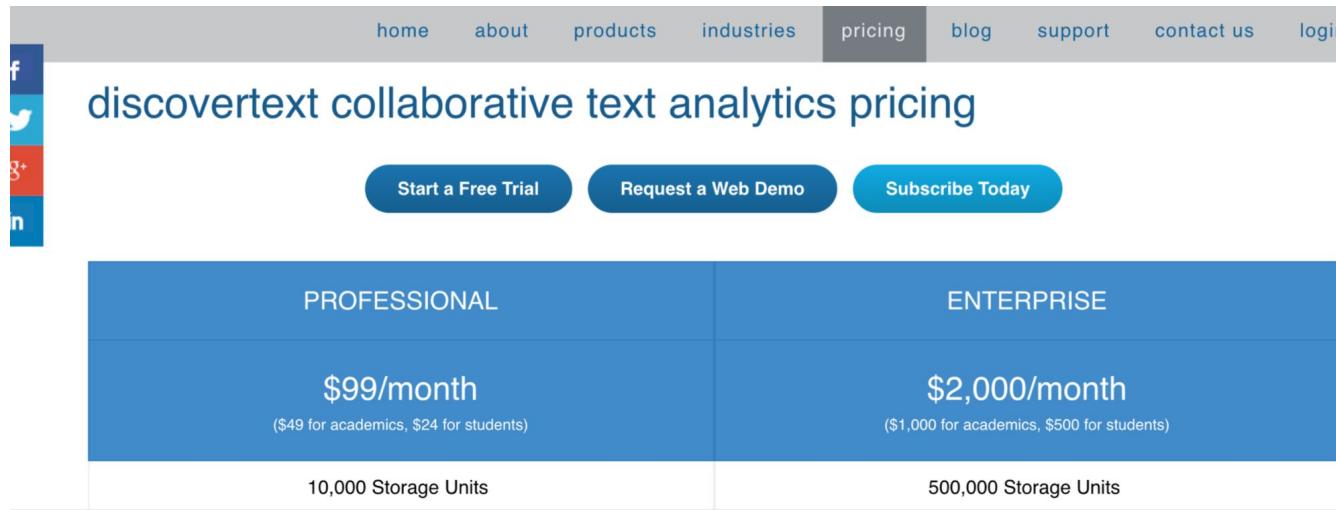
[Start a Free Trial](#)

netlytic



But, we won't use them either...

1. They provide advanced analytics;
2. But they require expensive software licenses or subscription
3. Most algorithms that drive the software are already open-source.



The screenshot shows the discovertext website's pricing section. At the top, there is a navigation bar with links: home, about, products, industries, pricing (which is highlighted in dark grey), blog, support, contact us, and login. To the left of the main content area, there are social media icons for Facebook, Twitter, Google+, and LinkedIn.

The main heading is "discovertext collaborative text analytics pricing". Below it are three buttons: "Start a Free Trial", "Request a Web Demo", and "Subscribe Today".

The pricing table is divided into two columns: "PROFESSIONAL" and "ENTERPRISE".

PROFESSIONAL	ENTERPRISE
\$99/month (\$49 for academics, \$24 for students)	\$2,000/month (\$1,000 for academics, \$500 for students)
10,000 Storage Units	500,000 Storage Units

Why using R

1. Free, open-source
2. Versatile and powerful; the latest algorithms are always implemented first in R and/or Python (another programming language used in data science)
3. It is an ecosystem and a community
4. The de facto language of data science, along with Python
5. R and Python are like PC and Mac. Most data scientists use both but prefer one over another.

Why open-source coding can be fun



Coding is a lot like assembling furnitures, except you don't have to purchase parts from IKEA. All parts are already made freely available through open-source platforms.



You can find a lot of open-source R codes on Github (<https://github.com>) and RPubs (<https://rpubs.com>). You can copy and adapt the codes to your own projects. For example, you can view the codes that produce the Russian trolls visualization [here](#).

Install R in your machine



Download and install R:

<https://www.r-project.org/>

Download and install RStudio:

<https://www.rstudio.com/>



Task for Week 1

▼ Week 1



Complete the ***Libraries/packages tutorial*** (https://curiositybits.shinyapps.io/R_social_data_analytics/#section-librariespackages) before the class on **9/5**.

Complete the ***Data frames*** (https://curiositybits.shinyapps.io/R_social_data_analytics/#section-data-frames) and ***Connecting to the Twitter API*** (https://curiositybits.shinyapps.io/R_social_data_analytics/#section-connecting-to-the-twitter-api) tutorials before the class on **9/10**

Don't forget to submit your completion reports (submission link is now live on Moodle)