

# Fundamentals of Data Analytics

SEN163A

Metadata

Jacopo De Stefani

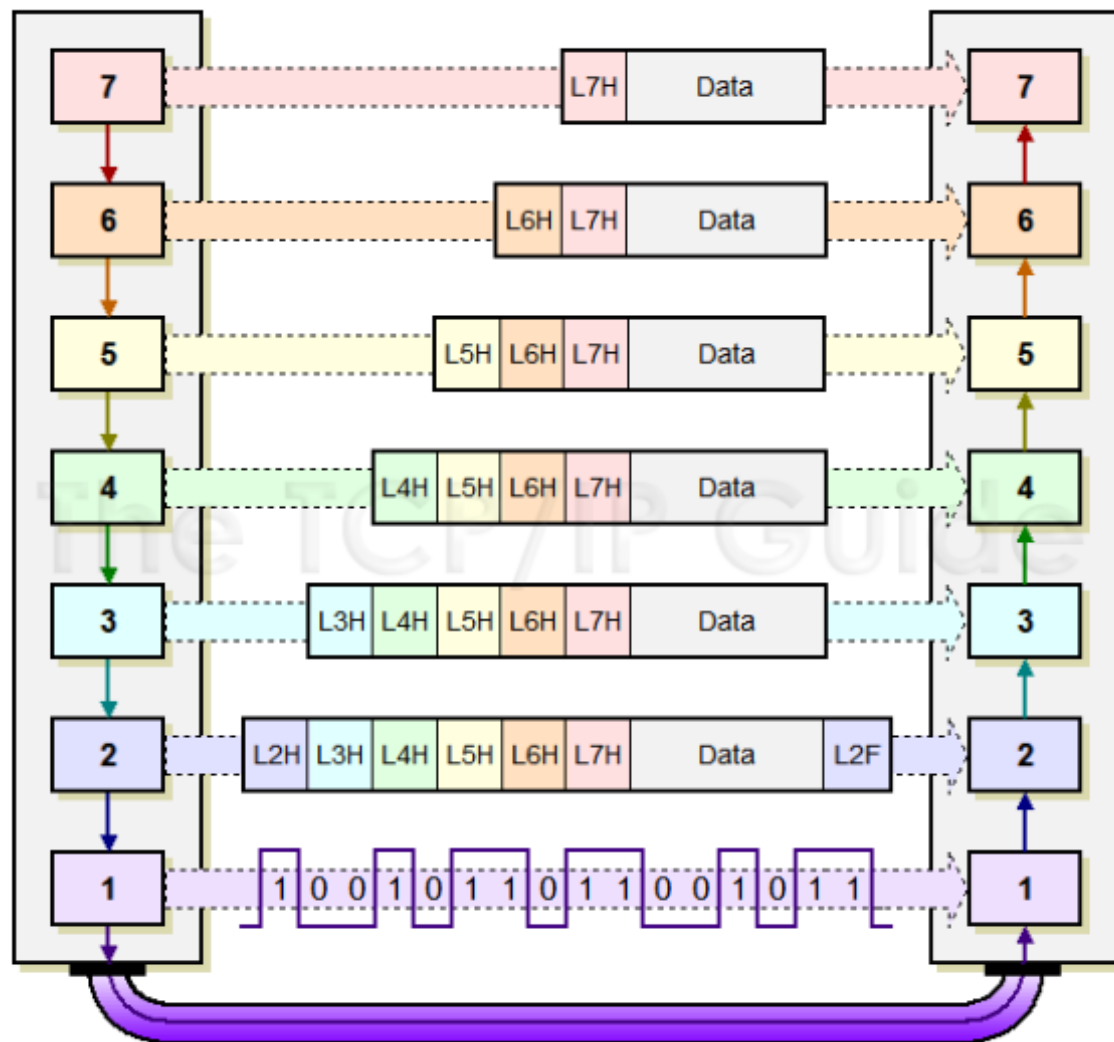
Based on the material by Tobias Fiebig

# Content notification (CN)

- Privacy violation
  - Surveillance
  - Pregnancy and abortion
  - Symbols of fascism
  - Genocide
  - Prosecution of LGBTQ+ people
- 
- If any of these topics emotionally affect you strongly, you are free to leave. We can then schedule an appointment where we can discuss the lecture's contents separated from these contents.

# Not all data is alike

- In network measurement:
  - Metadata
  - Payload



# Payload in a TCP packet

The image shows a Wireshark packet capture window titled "\*Ethernet". The filter bar at the top displays "tcp.port == 25". The packet list pane shows several packets, with packet 639 selected. The packet details pane shows the following structure:

- > Frame 639: 96 bytes on wire (768 bits), 96 bytes captured (768 bits) on interface \Device\NPF\_{BE4B534C-D25B-4B49-B73F-60345E167DC3}, Ethernet II, Src: Cisco\_67:53:c0 (28:6f:7f:67:53:c0), Dst: LCFCHeFe\_4a:91:e3 (50:7b:9d:4a:91:e3)
- > Internet Protocol Version 4, Src: 94.130.126.186, Dst: 145.94.42.199
- > Transmission Control Protocol, Src Port: 25, Dst Port: 51595, Seq: 1, Ack: 1, Len: 42
- > Simple Mail Transfer Protocol
  - > Response: 220 mail.aperture-labs.org ESMTP Postfix\r\n

The packet bytes pane shows the raw data of the selected packet, with the SMTP response payload highlighted in blue:

```
0000 50 7b 9d 4a 91 e3 28 6f 7f 67 53 c0 08 00 45 00 P{.J..(o.gS...E.
0010 00 52 8d fc 40 00 37 06 1c 48 5e 82 7e ba 91 5e .R..@.7. .H^..~..^
0020 2a c7 00 19 c9 8b b6 43 d2 3e 70 45 ec 37 50 18 *......C .>pE.7P.
0030 01 11 8f 55 00 00 32 32 30 20 6d 61 69 6c 2e 61 ...U..22 0 mail.a
0040 70 65 72 74 75 72 65 2d 6c 61 62 73 2e 6f 72 67 perture- labs.org
0050 20 45 53 4d 54 50 20 50 6f 73 74 66 69 78 0d 0a ESMTP P ostfix].
```

The status bar at the bottom indicates: "Response (smtp.response), 42 bytes" and "Packets: 3382 · Displayed: 33 (1.0%)".

# Metadata is protocol dependent

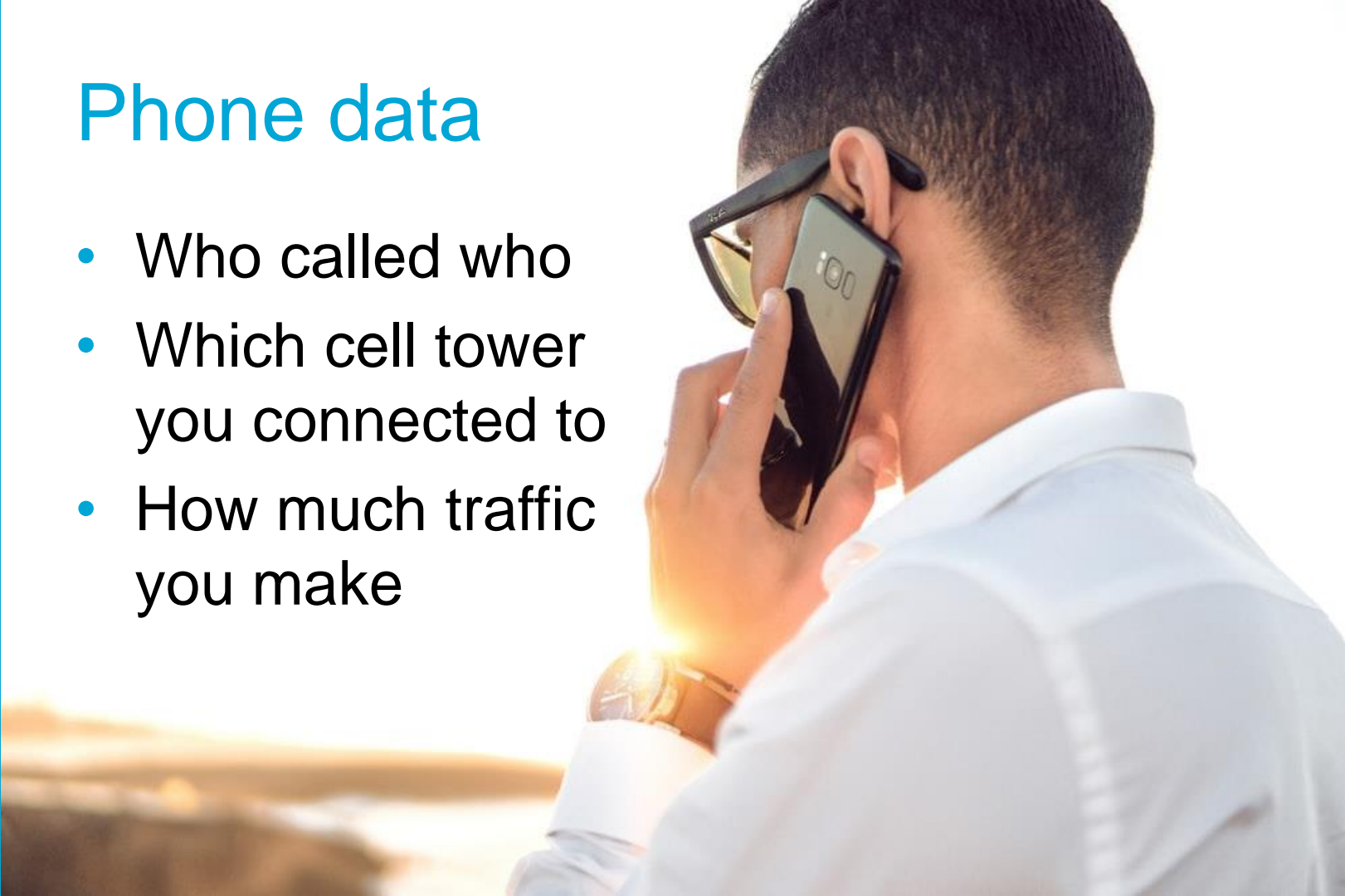
- For IP packets, TCP is already payload
- For TCP packets the protocol data is payload
- For an email, sender and recipient are metadata

# Metadata of an email

Return-Path: <T.Fiebig@tudelft.nl>  
Received: from mail.aperture-labs.org  
by mail.aperture-labs.org with LMTP  
id yP03LPkVVW49awAAj3/rZg  
(envelope-from <T.Fiebig@tudelft.nl>)  
for <tobias@aperture-labs.org>; Tue, 25 Feb 2020 12:41:29 +0000  
Delivered-To: tobias@aperture-labs.org  
Return-Path: <T.Fiebig@tudelft.nl>  
Received: from mail.aperture-labs.org (localhost [127.0.0.1])  
by mail.aperture-labs.org (Postfix) with ESMTP id 6F293C3BFEC  
for <tobias@fiebig.nl>; Tue, 25 Feb 2020 12:41:29 +0000 (UTC)  
...  
From: Tobias Fiebig <T.Fiebig@tudelft.nl>  
To: "tobias@fiebig.nl" <tobias@fiebig.nl>  
Subject: test  
Thread-Topic: test  
Thread-Index: AQHV69js7PIpFFhVvkaPDFjzRuE7Uw==  
Date: Tue, 25 Feb 2020 12:41:38 +0000  
Message-ID: <712cab1b-f71a-41fb-925e-0aadbb311160@email.android.com>  
Accept-Language: en-US, nl-NL

# Phone data

- Who called who
- Which cell tower you connected to
- How much traffic you make





# Metadata

- Times when you write/post something
- What articles you read
- Which websites you visit



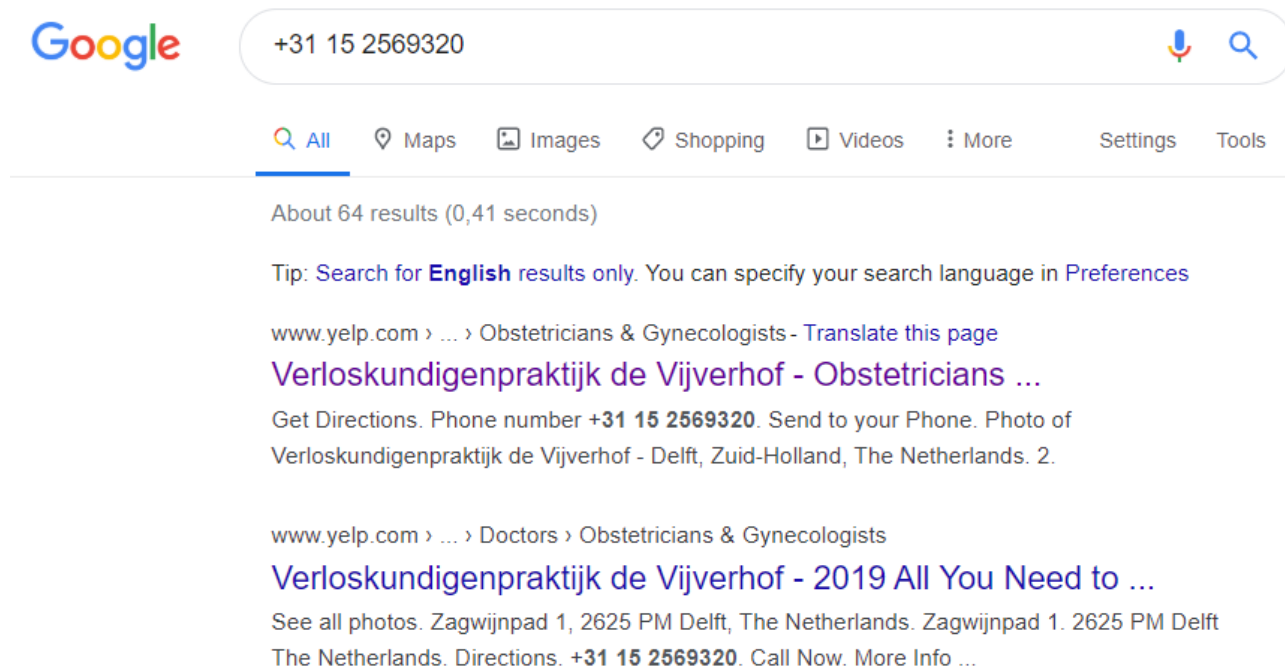
# Let's do a story in metadata...

- We have a user in our institution...
- The following things happen...

# Event 1: A received call

- +31 15 27 85700 receives a call from +31 15 2569320
- The former is our employee... ok...
- The second one... Let's do some OSINT (Open Source Intelligence; Fancy word for googling... ;-))

# Event 1: A received call



The screenshot shows a Google search interface. The search bar contains the phone number "+31 15 2569320". Below the search bar, the "All" tab is selected. The search results show "About 64 results (0,41 seconds)". A tip suggests searching for English results. The first result is from www.yelp.com, titled "Verloskundigenpraktijk de Vijverhof - Obstetricians ...". It includes a link to "Translate this page", a "Get Directions" button, the phone number "+31 15 2569320", and a photo of the practice. The second result is also from www.yelp.com, titled "Verloskundigenpraktijk de Vijverhof - 2019 All You Need to ...". It includes a link to "See all photos", the address "Zagwijnpad 1, 2625 PM Delft, The Netherlands", and a link to "Directions".

Google

+31 15 2569320

Q All Maps Images Shopping Videos More Settings Tools

About 64 results (0,41 seconds)

Tip: Search for **English** results only. You can specify your search language in [Preferences](#)

www.yelp.com › ... › Obstetricians & Gynecologists - [Translate this page](#)

**Verloskundigenpraktijk de Vijverhof - Obstetricians ...**

Get Directions. Phone number **+31 15 2569320**. Send to your Phone. Photo of Verloskundigenpraktijk de Vijverhof - Delft, Zuid-Holland, The Netherlands. 2.

www.yelp.com › ... › Doctors › Obstetricians & Gynecologists

**Verloskundigenpraktijk de Vijverhof - 2019 All You Need to ...**

See all photos. Zagwijnpad 1, 2625 PM Delft, The Netherlands. Zagwijnpad 1. 2625 PM Delft The Netherlands. Directions. **+31 15 2569320**. Call Now. More Info ...

# Ok... interesting...

- People often are called by their doctors at work
- People have the tendency of working when their doctors are working...

## Event 2: An uncommon call

- +31 15 27 85700 dials out to +31 616 80 98 99
- The call last for 96 minutes
- This is obviously a mobile phone number
- This number has never been called from that phone before...
- Hm...

## Event 2: An uncommon call

- We ultimately find the number as the emergency contact for that person... their sister...

# Event 3: A story in encrypted TCP Packets

\*Ethernet

File Edit View Go Capture Analyze Statistics Telephony Wireless Tools Help

tcp.port == 443 and ip.addr == 95.170.72.132

No.	Time	Source	Destination	Protocol	Length	Info
1560	31.990452	95.170.72.132	145.94.42.199	TCP	60	443 → 61273 [ACK] Seq=1 Ack=518 Win=30720 Len=0
1561	31.991610	95.170.72.132	145.94.42.199	TLSv1.2	1514	Server Hello
1562	31.991611	95.170.72.132	145.94.42.199	TLSv1.2	1514	Certificate [TCP segment of a reassembled PDU]
1563	31.991688	145.94.42.199	95.170.72.132	TCP	54	61272 → 443 [ACK] Seq=518 Ack=2921 Win=131328 Len=0
1564	31.992671	95.170.72.132	145.94.42.199	TLSv1.2	198	Server Key Exchange, Server Hello Done
1565	31.995721	95.170.72.132	145.94.42.199	TLSv1.2	1514	Server Hello
1566	31.995726	95.170.72.132	145.94.42.199	TLSv1.2	1514	Certificate [TCP segment of a reassembled PDU]
1567	31.995729	95.170.72.132	145.94.42.199	TLSv1.2	198	Server Key Exchange, Server Hello Done
1568	31.995946	145.94.42.199	95.170.72.132	TCP	54	61273 → 443 [ACK] Seq=518 Ack=3065 Win=131328 Len=0
1570	32.008131	145.94.42.199	95.170.72.132	TLSv1.2	180	Client Key Exchange, Change Cipher Spec, Encrypted Hand
1572	32.011525	145.94.42.199	95.170.72.132	TLSv1.2	180	Client Key Exchange, Change Cipher Spec, Encrypted Hand
1573	32.011581	95.170.72.132	145.94.42.199	TLSv1.2	105	Change Cipher Spec, Encrypted Handshake Message
1574	32.014113	95.170.72.132	145.94.42.199	TLSv1.2	105	Change Cipher Spec, Encrypted Handshake Message
1575	32.057501	145.94.42.199	95.170.72.132	TCP	54	61272 → 443 [ACK] Seq=644 Ack=3116 Win=131072 Len=0

> Frame 1567: 198 bytes on wire (1584 bits), 198 bytes captured (1584 bits) on interface \Device\NPF {BE4B534C-D25B-4B49-B73F-60345E1...}

0000 50 7b 9d 4a 91 e3 28 6f 7f 67 53 c0 08 00 45 00 P{·J··(o·gS··E·  
0010 00 b8 ff c6 40 00 37 06 df 25 5f aa 48 84 91 5e ····@·7··%·H··^  
0020 2a c7 01 bb ef 59 ec 61 29 3a be 89 1c 65 50 18 \*····Y·a·):···eP·  
0030 00 3c f3 dc 00 00 d8 c6 22 5a 1a 6b 8f b6 1a 72 ·<·....."Z·k····

Frame (198 bytes)    Reassembled TCP (338 bytes)

Source or Destination Hardware Address (eth.addr), 6 bytes    || Packets: 5136 · Displayed: 867 (16.9%)    || Profile: Default



## Event 3: A story in encrypted TCP Packets

```
tfiebig@tardis ~ % host 145.94.42.199
```

```
199.42.94.145.in-addr.arpa domain name pointer x-  
145-94-42-199.wired.tudelft.nl.
```

```
tfiebig@tardis ~ % host 95.170.72.132
```

```
132.72.170.95.in-addr.arpa domain name pointer  
webhosting-cluster.transip.nl.
```

## Event 4: Just another call...

- +31 15 27 85700 dials out to +31 20 693 21 51
- The call lasts for 10 minutes...

## Event 5: The employee calls in sick...

- The next day the employee calls in sick for a day
- A week later they call in sick for a week

# Question

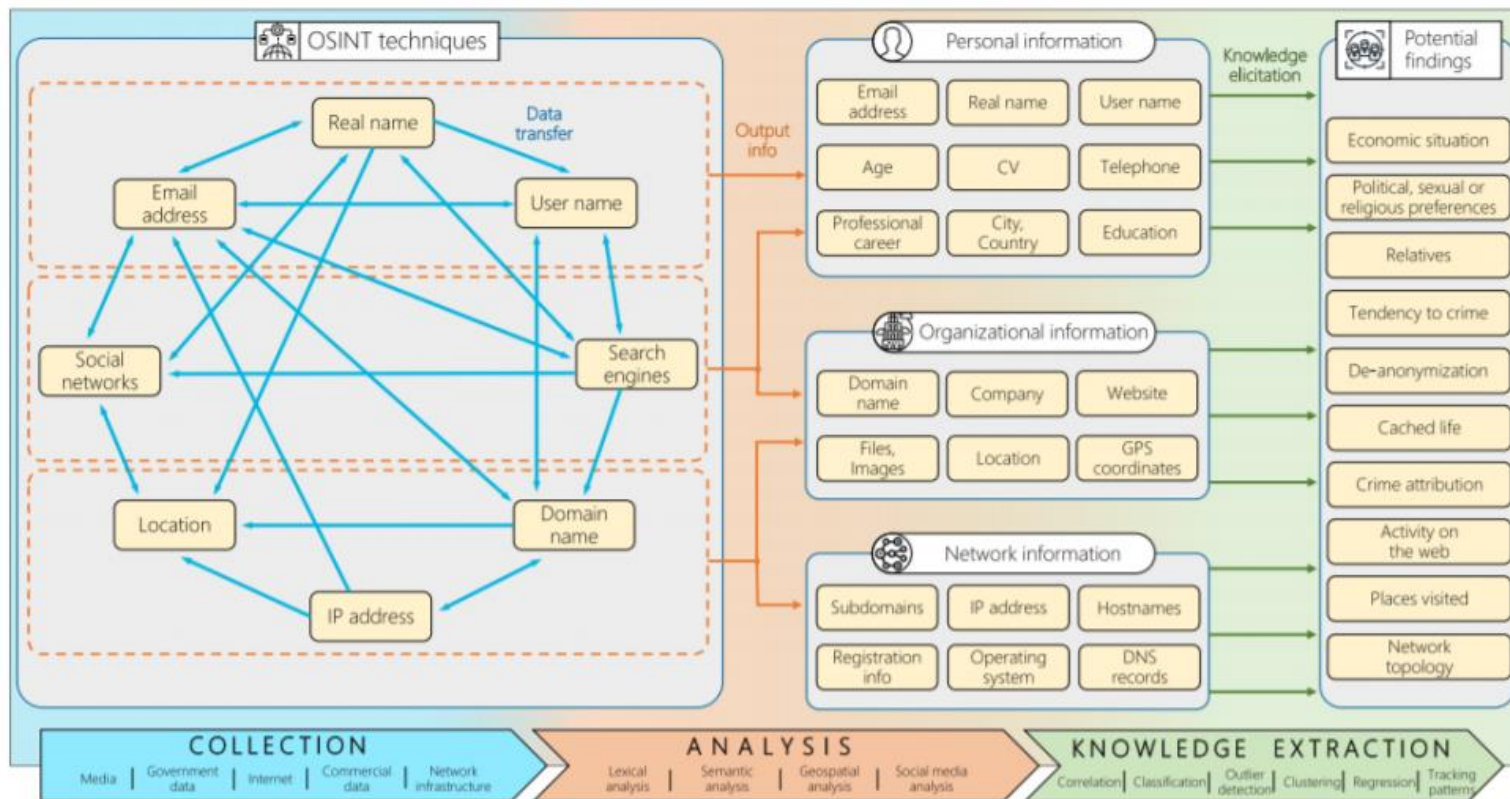
What happened?

# A story in words...

- The employee received a call from their OBGYN informing her of her pregnancy
- The employee calls her sister, a person she trusts, and discusses this
- She decides to have an abortion, and visits <https://abortuskliniek-amsterdam.nl> which is reachable on 95.170.72.132
- As you can not make an appointment via mail, she gives them a call and makes an appointment for the next day
- After a mandatory wait time she receives medical care

# Moral of the story... as a data scientist

- The data you collect and work with can tell you a lot about the humans creating that data
- As an engineer and data scientist, you carry the responsibility for that data, and only wilding it in a responsible manner



**FIGURE 2. Principal OSINT workflows and derived intelligence.**

# Moral of the story... as a end-user

- Trade-off between ease-of-use and data privacy/security
- Some tips and tricks for increasing online privacy: <https://nordvpn.com/blog/how-to-be-anonymous-online/>
- Disclaimer: I am not affiliated with NordVPN, DYOR and find the best provider.





# A story of not so meta-data

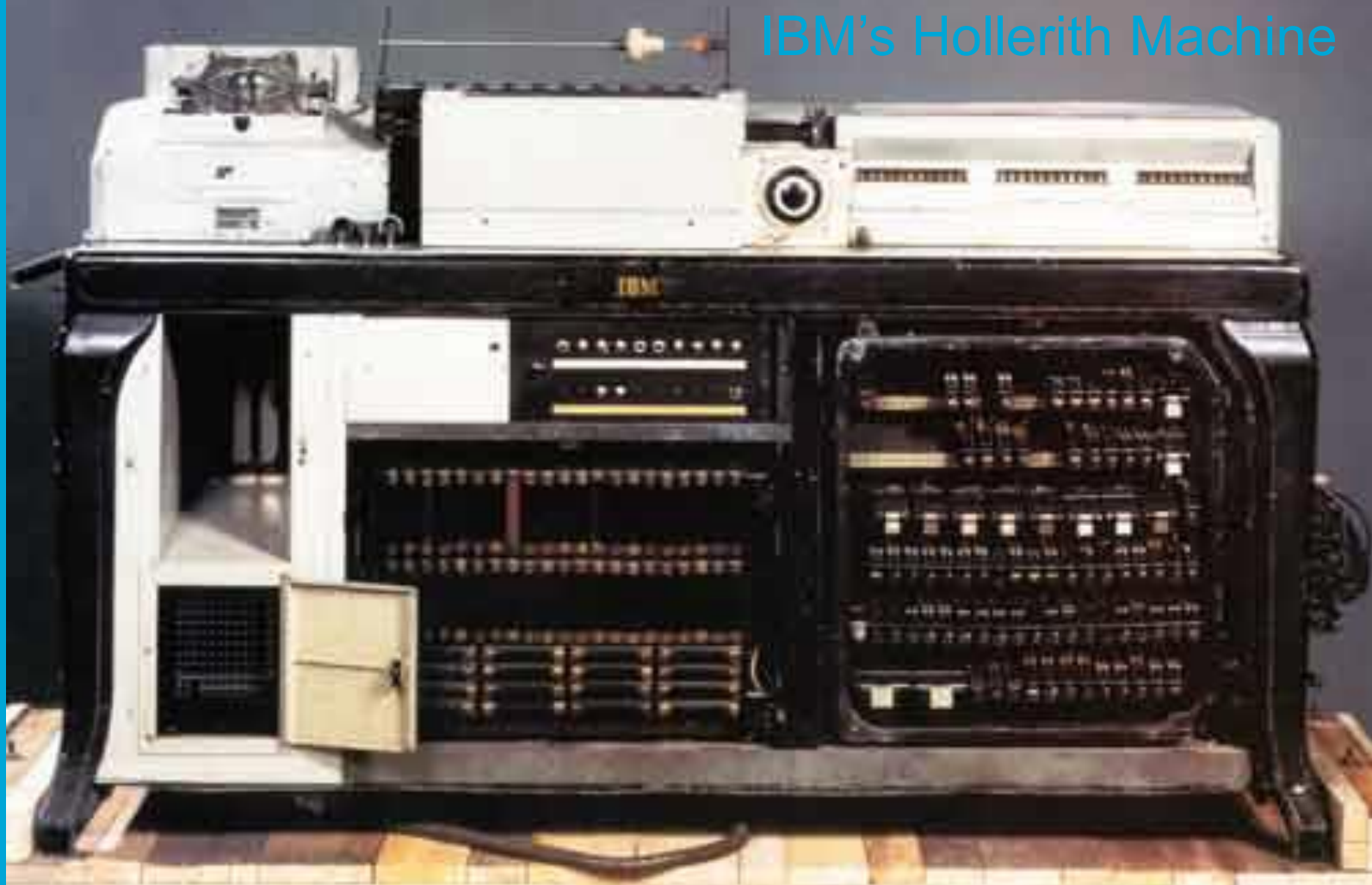
# Digital Government 101

- Imagine you make a small database of all people in you country... for easier administration
- Like every good government project:
  - Run by IBM
  - Insanely efficient
  - Big data for the 20<sup>th</sup> century!

# The Dutch: Always early adopters

- We have a highly digital government
  - Nearly everything—from taxes to registering in a flat—can be done online
  - Applying for social benefits? Online!
  - We have DigiD (a central authentication framework) to log into, e.g., our healthcare provider

## IBM's Hollerith Machine







# Public Administration with IBM

- Rolled out in the Netherlands
- Used to ease municipal administration, taxation, census etc.
- Recorded names, addresses, and—among other things—religion

# No privacy issue...

- ... the handling authority is well intentioned...
- ... it actually has to have that data...
- What could possibly go wrong?



# Invasion...





And then the new cards looked like this

# The Holocaust was supported by the first 'IT' systems (Punchcard Machines)

- Basically ran an IT infrastructure
  - (even before the Netherlands were occupied; Just happened to use the same supplier)
- One of the biggest crimes in human history would not have been as easy without engineers!



This is why the Dutch resistance...





...mostly burned archives.

# Background information

- Read these:
  - <http://db.yadvashem.org/righteous/family.html?language=en&itemId=4043044>
  - [https://nl.wikipedia.org/wiki/Aanslag\\_op\\_het\\_Amsterdams\\_bevolkingsregister\\_1943](https://nl.wikipedia.org/wiki/Aanslag_op_het_Amsterdams_bevolkingsregister_1943)
  - [https://www.verzetsmuseum.org/museum/nl/tweede-wereldoorlog/begrippenlijst/achtergrond,aanslag/amsterdamse\\_bevolkingsregister](https://www.verzetsmuseum.org/museum/nl/tweede-wereldoorlog/begrippenlijst/achtergrond,aanslag/amsterdamse_bevolkingsregister)
- If you have family members still remembering those times, talk to them, learn from them. Understand history, keep the memory alive.



Break

# But we don't collect that anymore!

- Dutch authorities may no longer collect religion. 😊
- So, things are good...
- ...
- ... right?

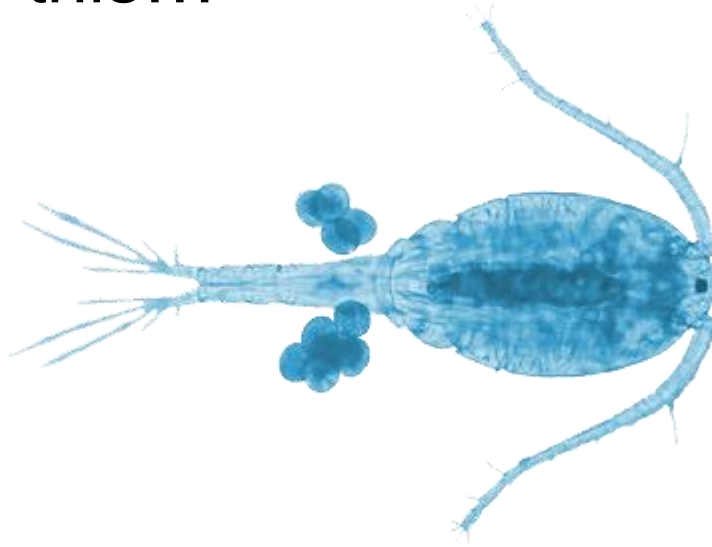
# But we don't collect that anymore!

- Dutch authorities may no longer collect religion. 😊
- So, things are good...
- ...
- ... right?
- METADATA!



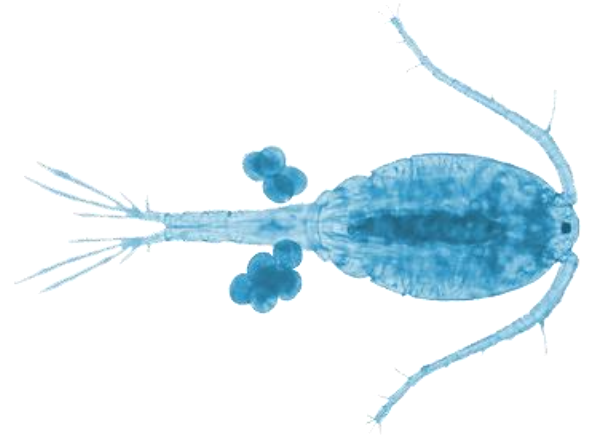
# Little excursion...

- What is this...



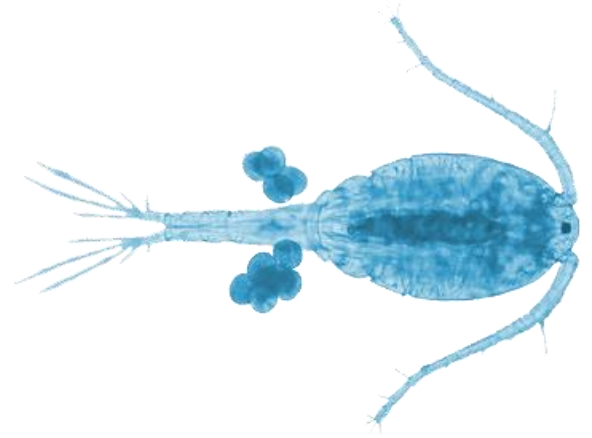
# This is a copepod

- Small crustacean living in New York tap water
- Not health critical
- Not...



# This is a copepod

- Small crustacean living in New York tap water
- Not health critical
- Not... kosher.



Amazon Home

Shop by Room

Scout | Style Explorer

Shop by Style

Home Décor

Furniture

Kitchen & Dining

Bed & Bath

Garden & Outdoor

Home Improvement

## SCOUT | Find your just right

Explore now ▸

◀ Back to search results for "kosher faucet"



### E-Z Filter Glatt **Kosher** EZ Water Filter, E-Z-Filter hooks on Faucet

by GiftGadge



1 customer review

Available from these sellers.

- Completely Removes Bugs
- Safe for Hot and Cold Water
- Easy Installation Right Onto Faucet
- Intended for Residential Use

New (1) from \$12.00 + \$5.49 shipping

Share    

 Deliver to Netherlands

See All Buying Options

Add to List

Have one to sell?

Sell on Amazon



#### Prepare your lawn and garden for fall

As cooler fall temperatures approach, take time to remove leaves, loosen soil, add fertilizer, spread seed, and improve the appearance of your yard [Learn](#)

# What data do you have...

- ...that you need?
- ...that you accidentally collect?
- ...that you want?
- ...that you can not prevent yourself from collecting?

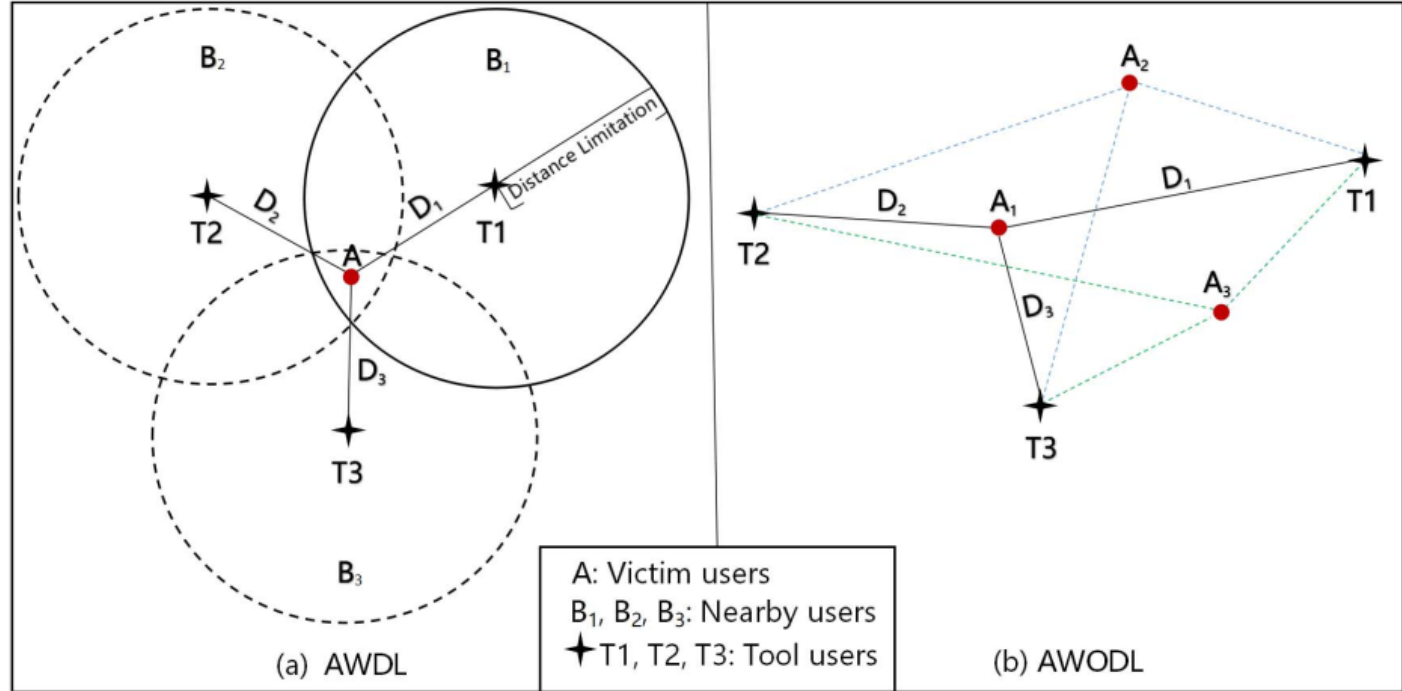
# Who knows Tinder?



Who knows... Grindr?



# Location triangulation



Zhao, Fanghua, et al. "You Are Where You App: An Assessment on Location Privacy of Social Applications." 2018 IEEE 29th International Symposium on Software Reliability Engineering (ISSRE). IEEE, 2018.



# So why is this bad... ?

- Grindr used to claim that it had users in over 140 countries...
- Grindr actually suggests it can be trusted in their TOS

# So why is this bad... ?

- Grindr used to claim that it had users in over 140 countries...
- Grindr actually suggests it can be trusted in their TOS
- Being gay is not legal everywhere...

# So why is this bad... ?

- Grindr used to claim that it had users in over 140 countries...
- Grindr actually suggests it can be trusted in their TOS
- Being gay is not legal everywhere...

# So why is this bad... ?

- Grindr used to claim that it had users in over 140 countries...
- Grindr actually suggests it can be trusted in their TOS
- Being gay is not legal everywhere...
- Actually used to hunt gay men:  
<https://www.independent.co.uk/news/world/africa/egypts-police-using-social-media-and-apps-like-grindr-to-trap-gay-people-9738515.html>

# Let's look at another datapoint...

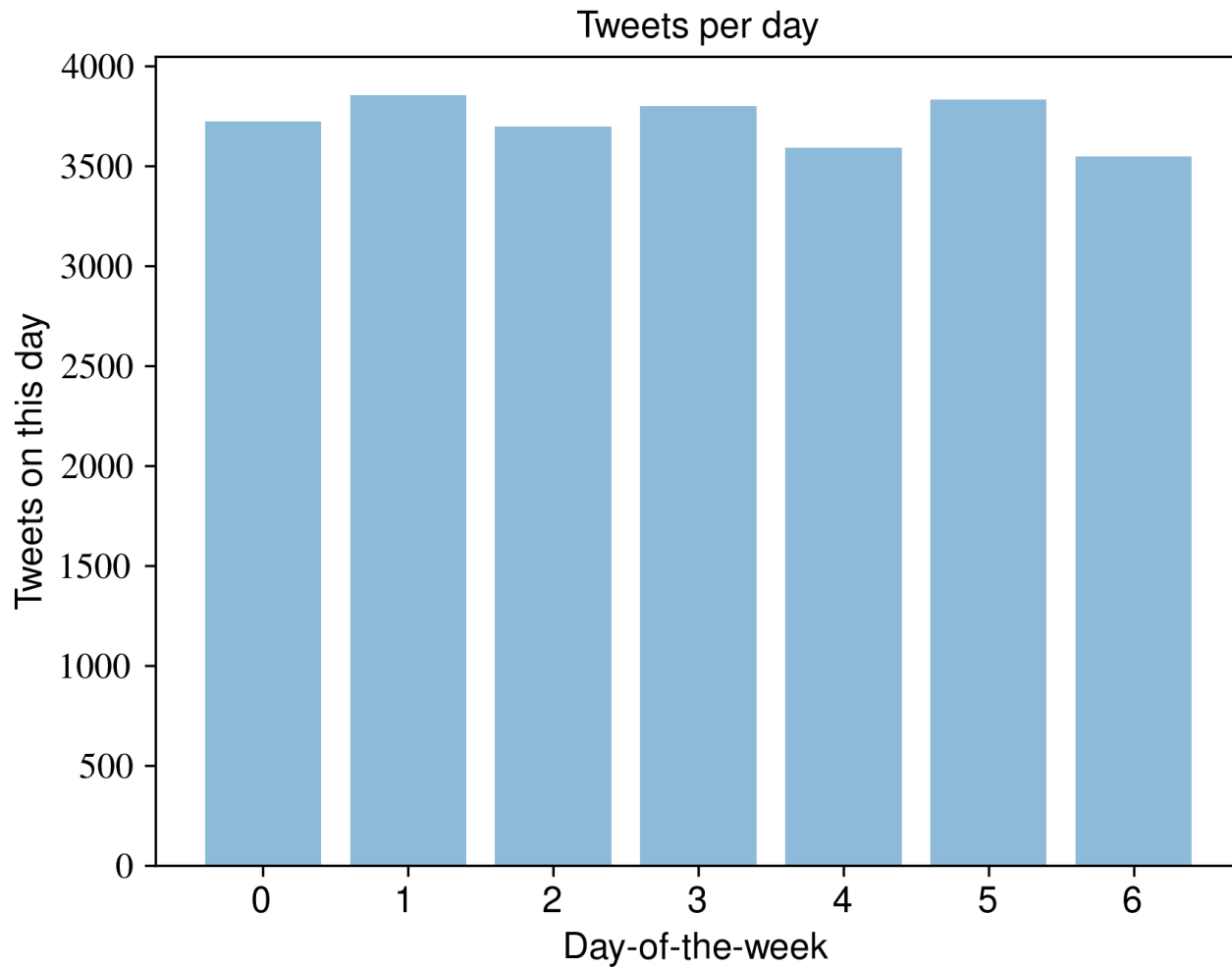
- Wed Mar 4 14:17:46 CET 2020

# Question

What can we learn from this?

# Let's try...

- What we have: 26,045 tweets from a single person over 7 years
- Let's see what we can... see

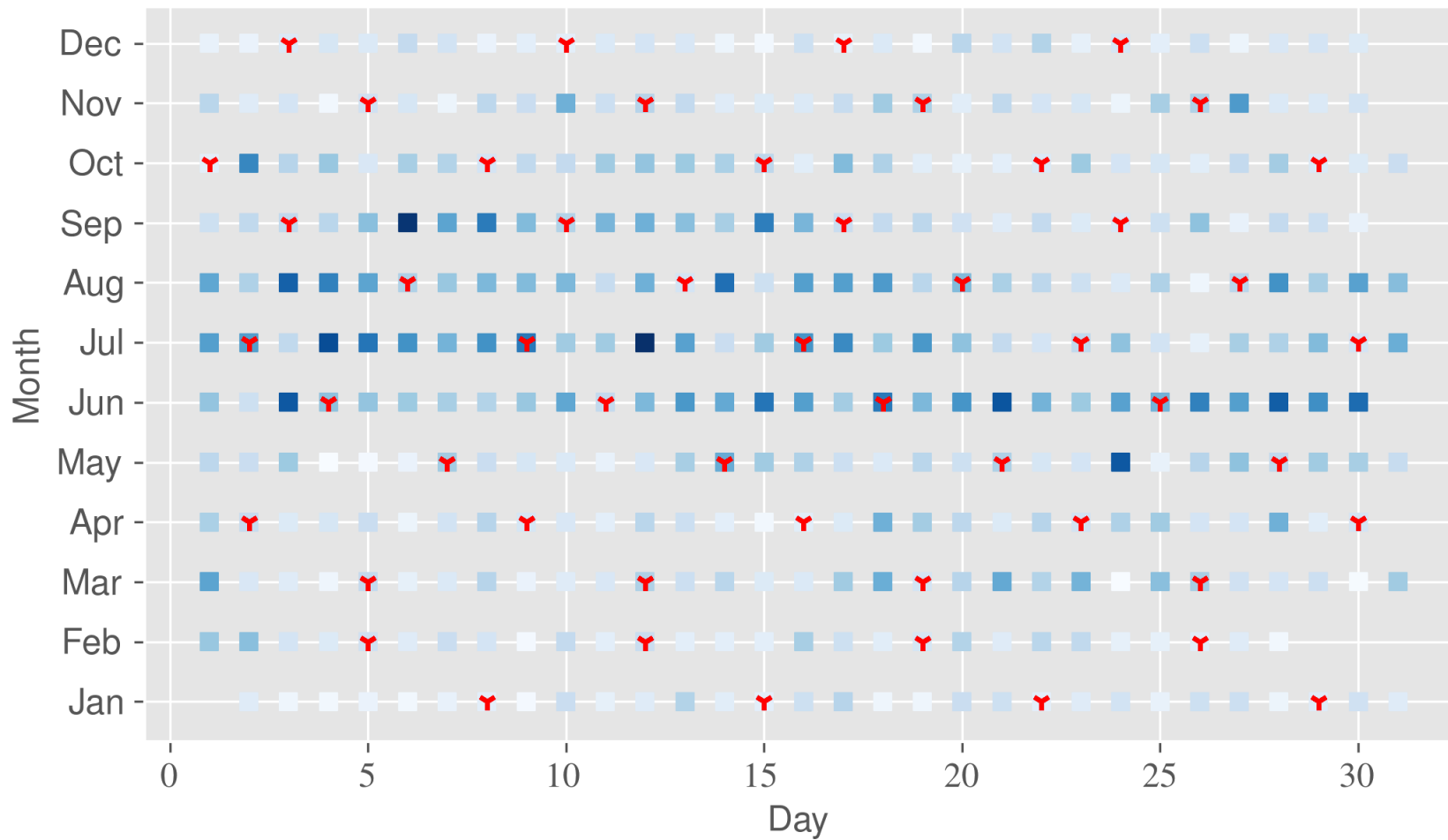




A 10x32 grid of colored squares representing a heatmap. The x-axis is labeled from 0 to 30, and the y-axis is labeled from 0 to 9. The colors range from light blue to dark blue, with some squares being white. The pattern shows varying intensities across the grid, with some rows having more dark blue squares than others.

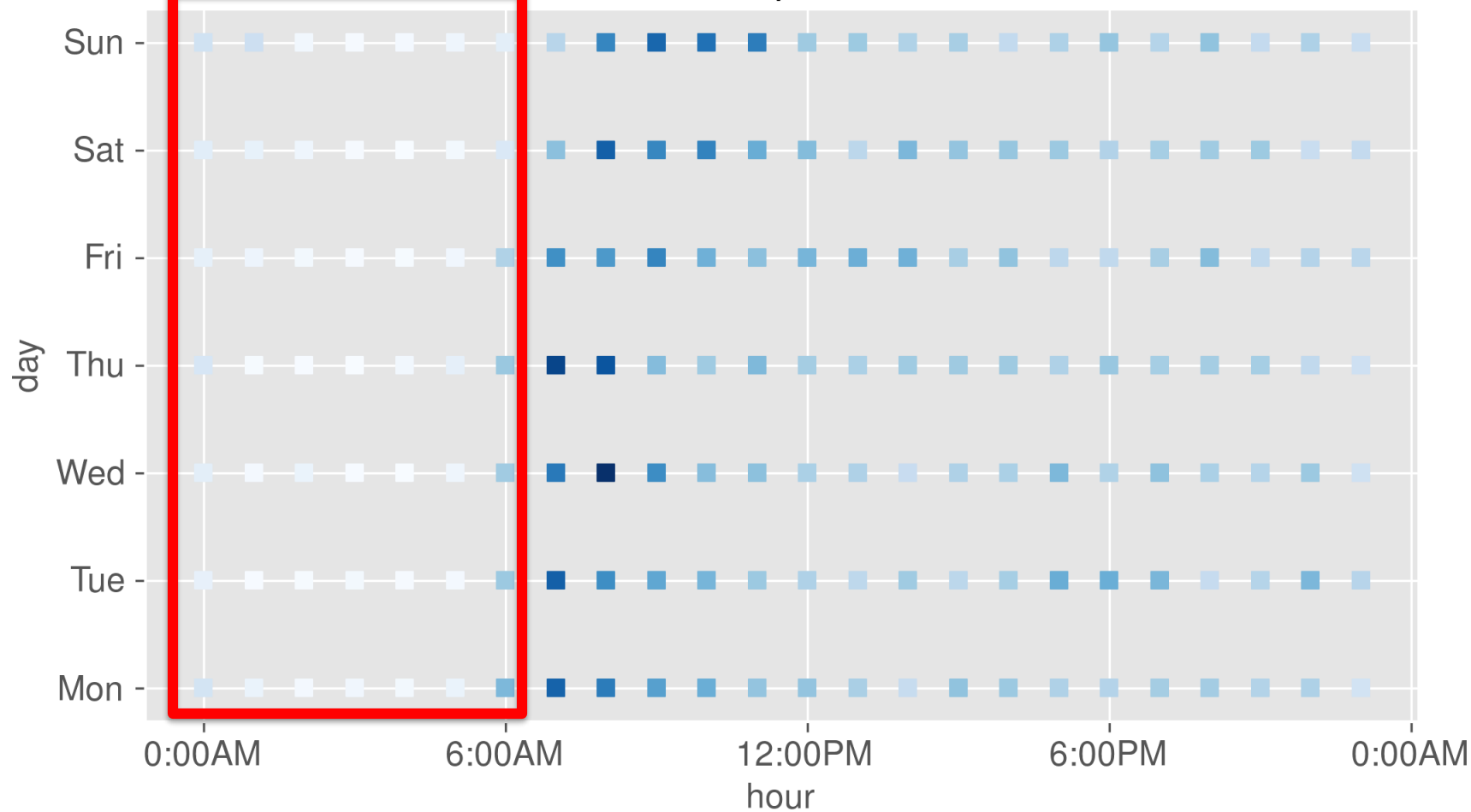
Day

# Tweets per day 2018

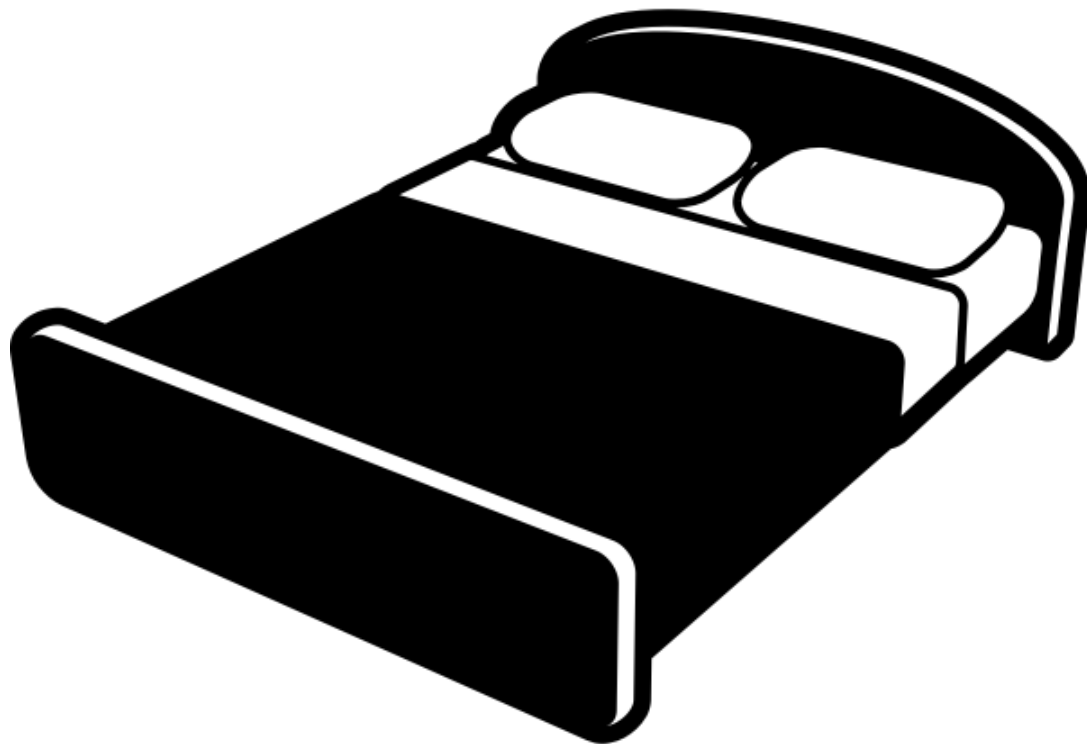


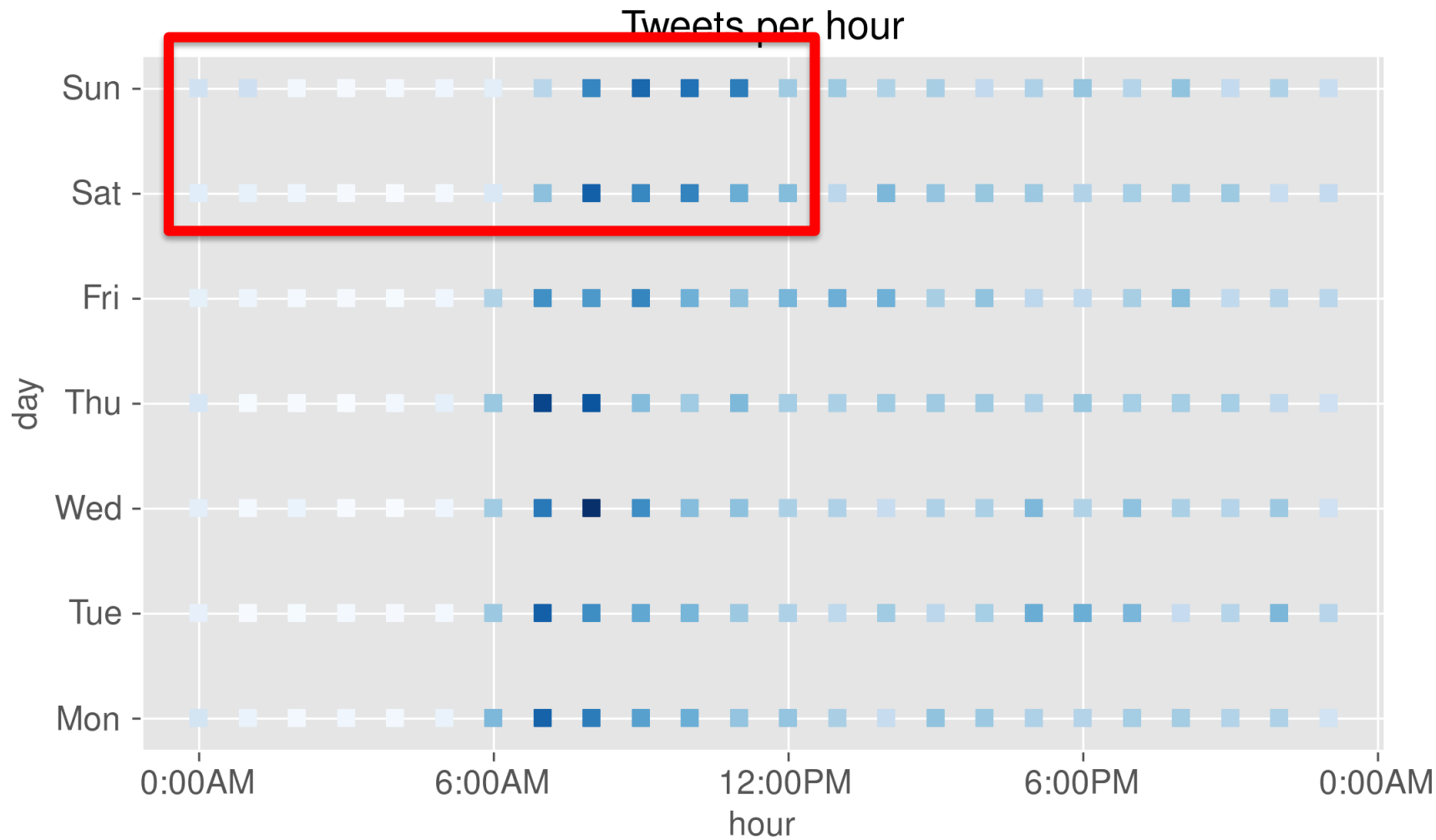
0:00AM

hour



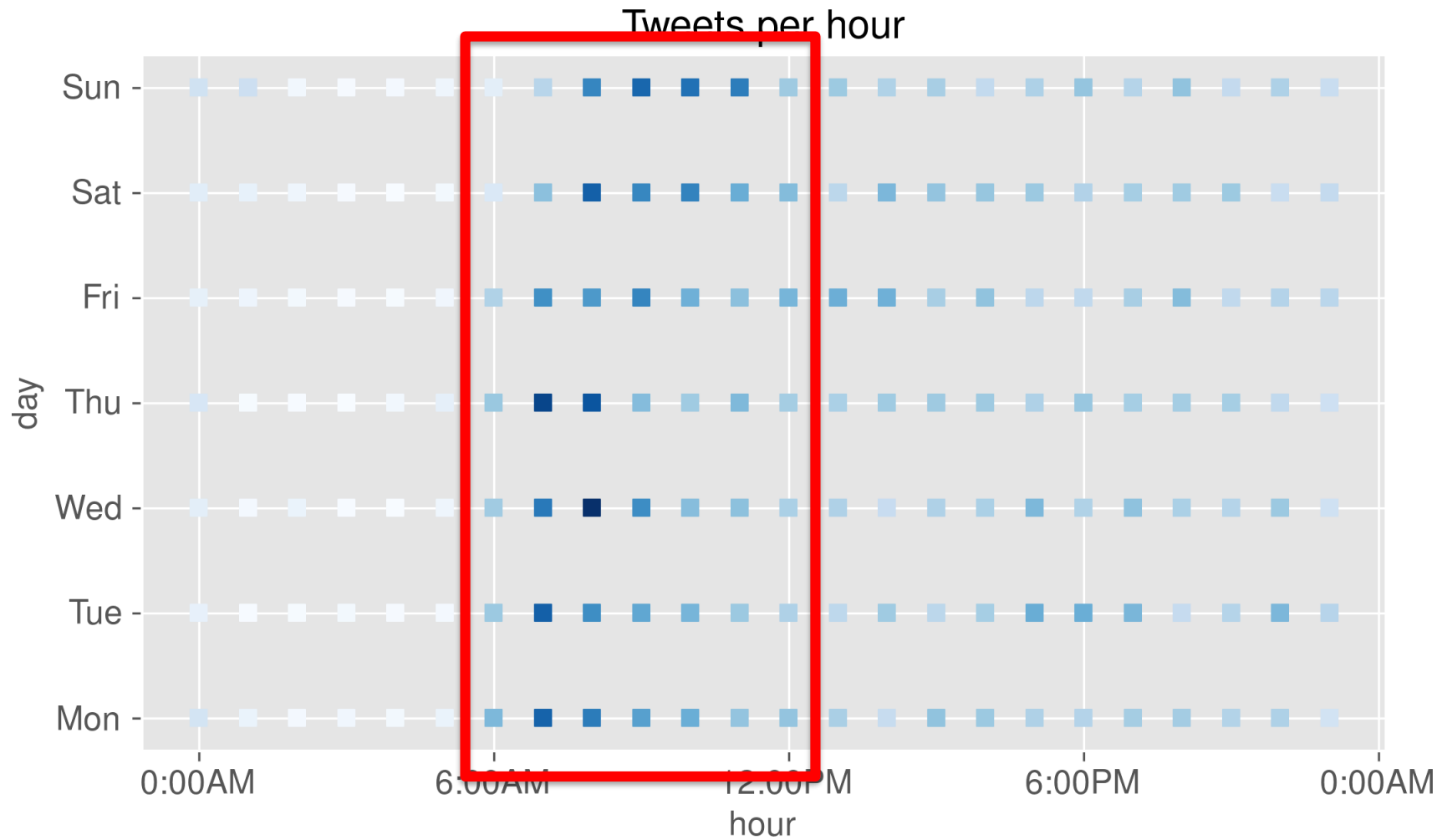
# Sleepingtime...





# Sleeping in on weekends...



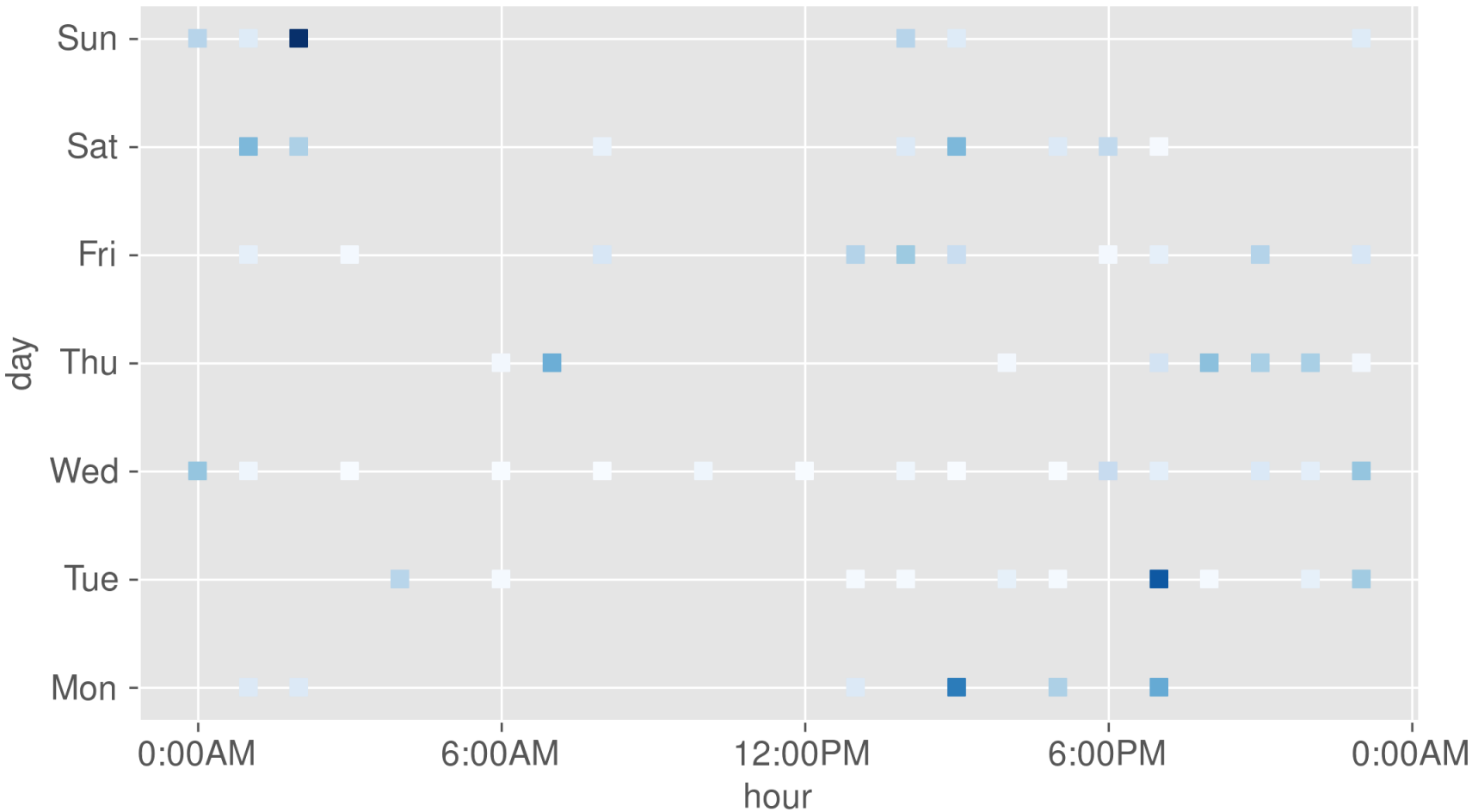




Well... coffee and newspapers...  
I am not a boomer...



## Tweets per hour 14-05-2019 to 26-05-2019

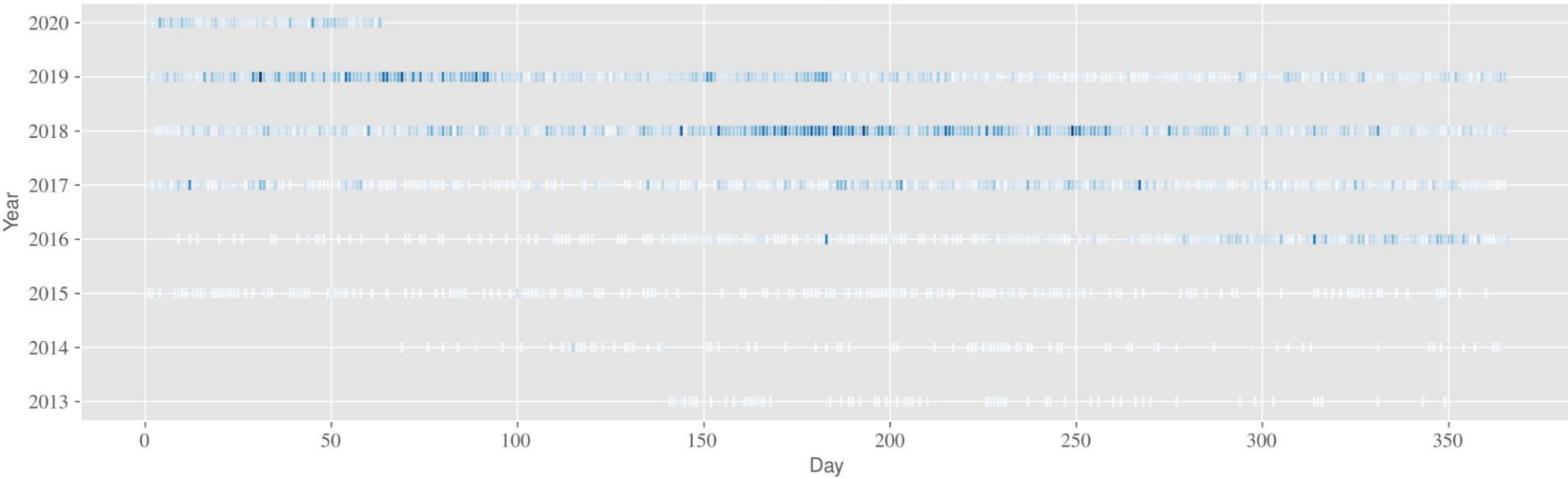


# These weeks look different...

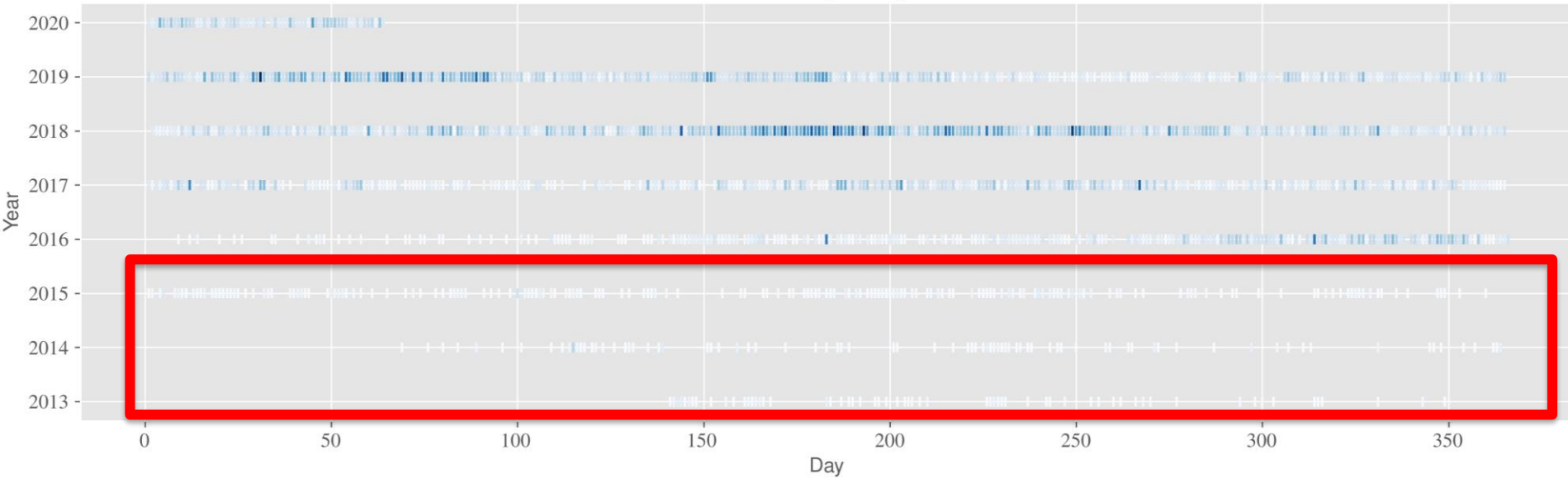
# These weeks look different...

- Went to the US for two conferences that week
- No time for tweeting
- Highly different timezone

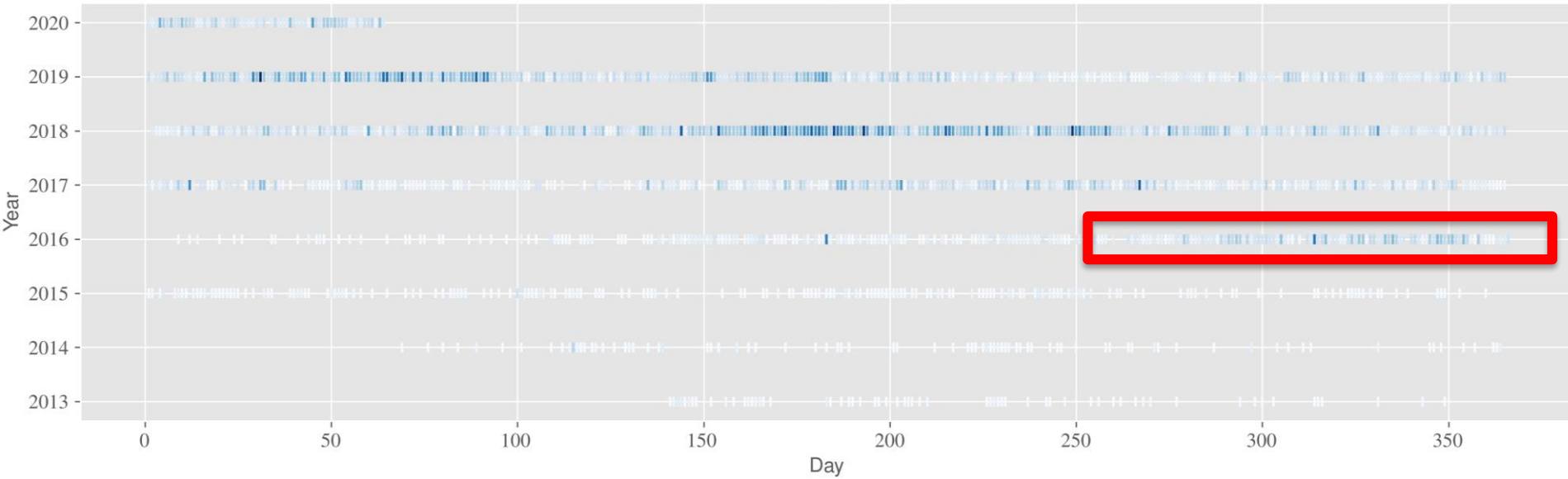
Tweets per day

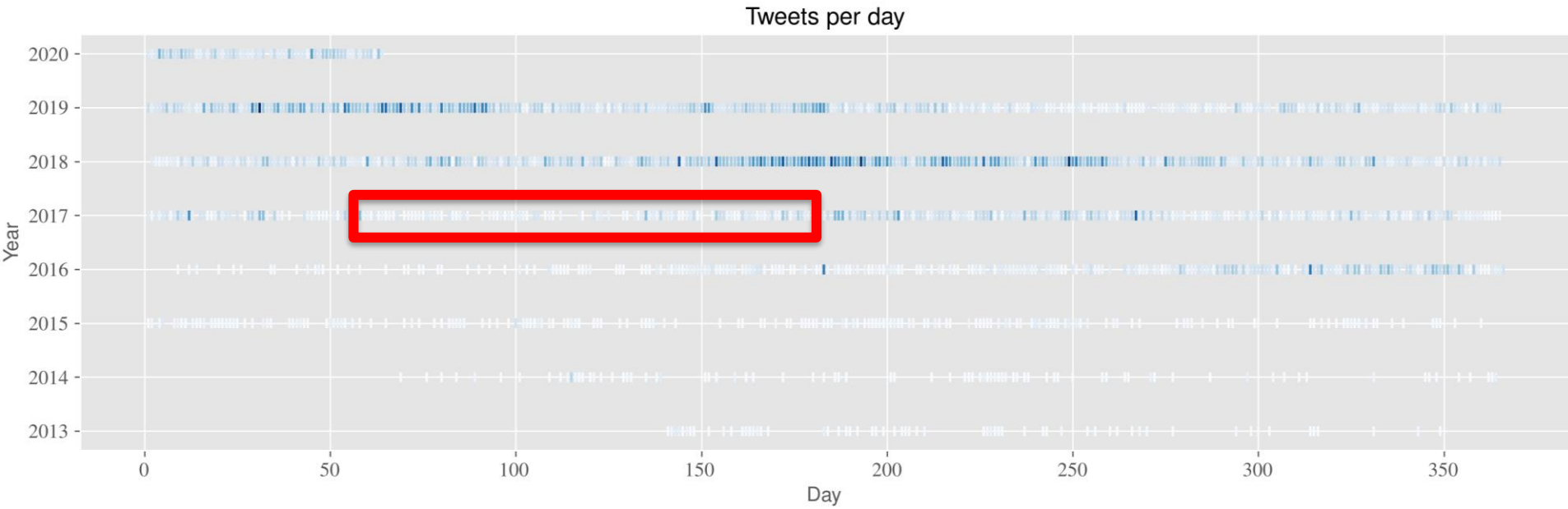


Tweets per day



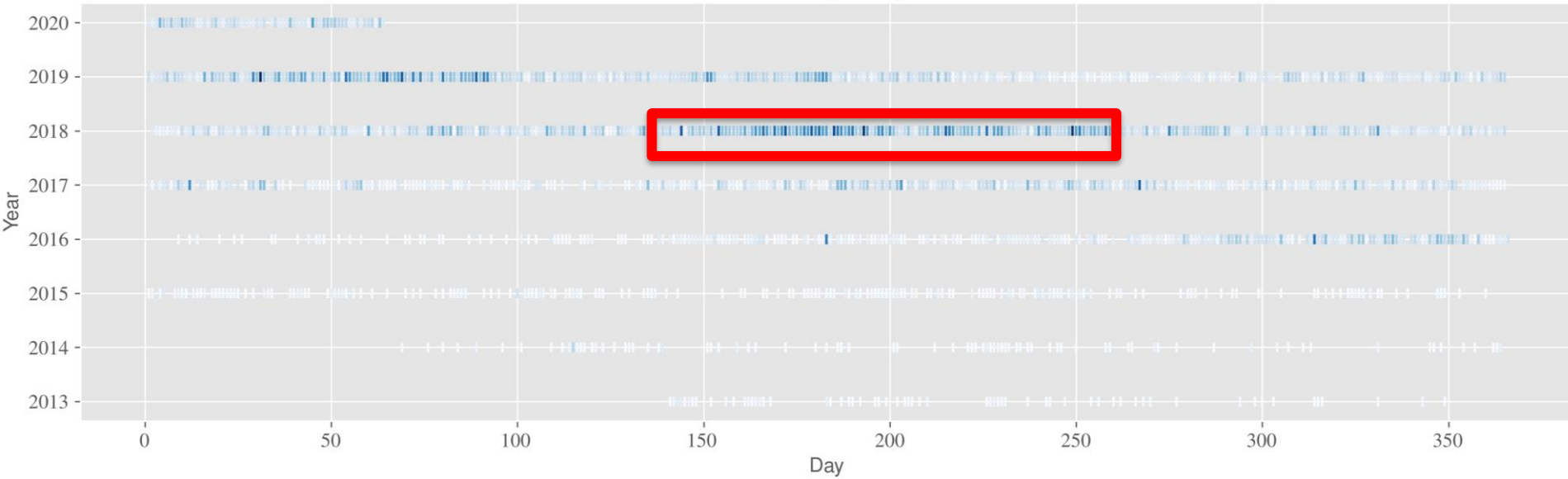
Tweets per day



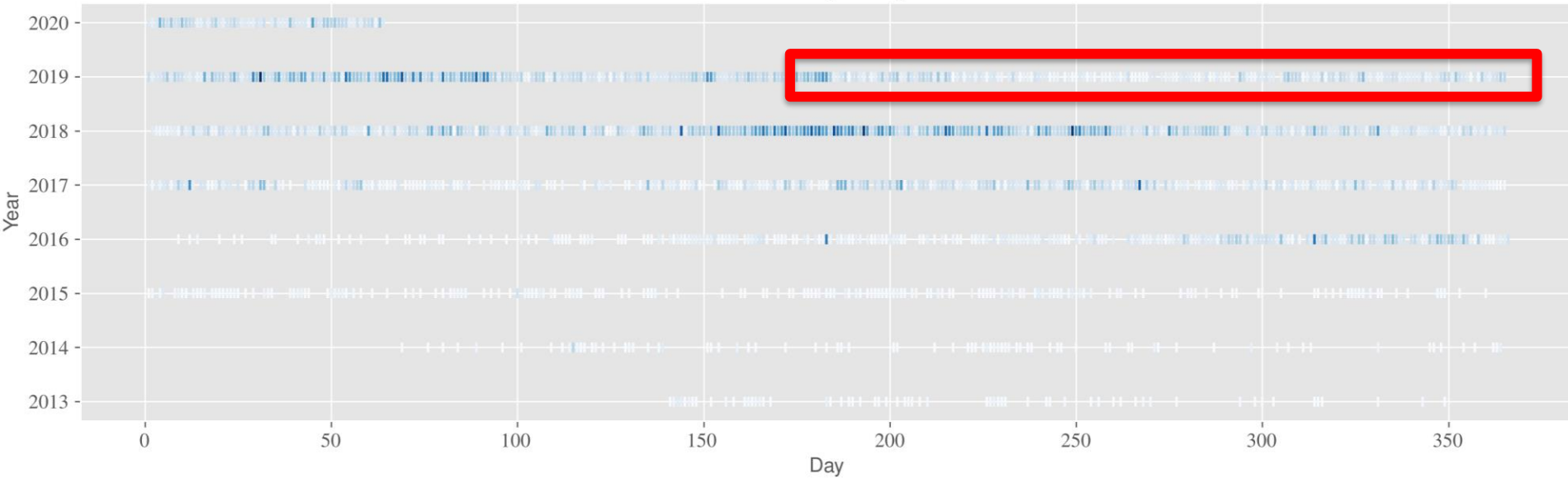




Tweets per day



Tweets per day



# Data may lie

- You interpret the data
- What you infer may have totally different causes
- Always cross-validate what you see with additional sources (if you can 😊)

# Getting Data

- Webscaping!
- You can roll your own toolchain (Using python's URL library urllib3)
- You can use existing ones (selenium, BeautifulSoup)
- Example: <https://www.edureka.co/blog/web-scraping-with-python/>

## Assignment 3 – Webscaping

# From the online example

- Import necessary libraries

```
from selenium import webdriver  
from BeautifulSoup import BeautifulSoup  
import pandas as pd
```

# From the online example

- Configure the 'driver' for selenium (a local browser)
- Might look different on your system!  
Search the web on how to do this.

```
driver = webdriver.Chrome("/usr/lib/chromium-  
browser/chromedriver")
```

# From the online example

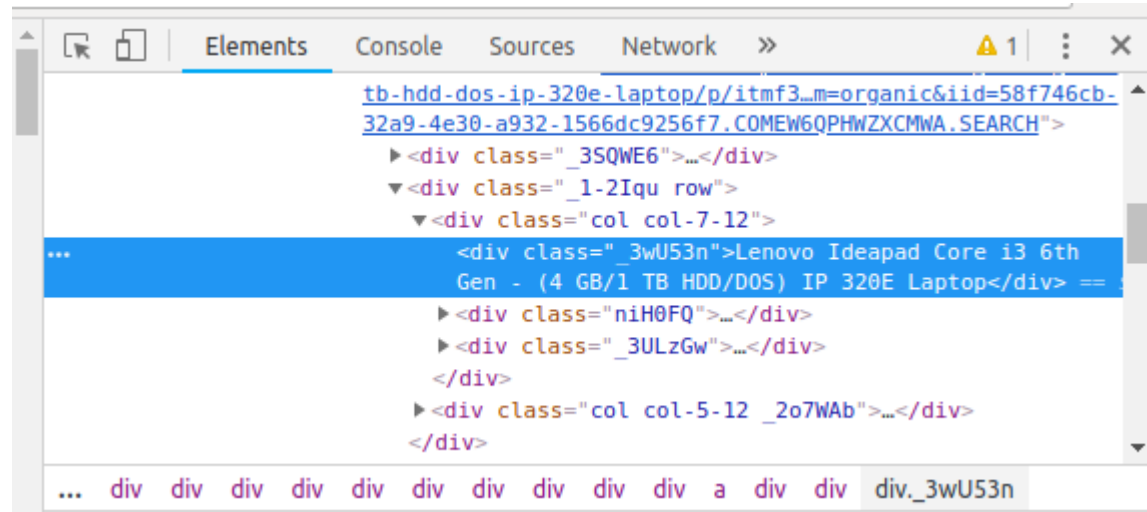
- Set up some variables to store results

```
products=[]  
prices=[]  
ratings=[]  
driver.get("https://www.flipkart.com/laptops/~buyback-  
guarantee-on-laptops-/pr?sid=6bo%2Cb5g&uniq")
```



# From the online example

- Identify the right <div>s
- Klick 'inspect tab' (or hit F12)
- Identify the right <div> names:



# From the online example

- Parse the content

```
content = driver.page_source
soup = BeautifulSoup(content)
for a in soup.findAll('a', href=True,
    attrs={'class': '_31qSD5'}):
    name=a.find('div', attrs={'class': '_3wU53n'})
    price=a.find('div', attrs={'class': '_1vC40E _2rQ-NK'})
    rating=a.find('div', attrs={'class': 'hGSR34 _2beYZw'})
    products.append(name.text)
    prices.append(price.text)
    ratings.append(rating.text)
```

# From the online example

- Store the data (or continue working with it)

```
df = pd.DataFrame({'Product Name':products, 'Price':prices,  
                  'Rating':ratings})  
df.to_csv('products.csv', index=False, encoding='utf-8')
```

# With urllib3

- See: <https://urllib3.readthedocs.io/en/latest/>
- `r.data` has the remote resource as text, ready for parsing

```
>>> import urllib3
>>> http = urllib3.PoolManager()
>>> r = http.request('GET', 'http://httpbin.org/robots.txt')
>>> r.status
200
>>> r.data
'User-agent: *\nDisallow: /deny\n'
```