Fundamentals of Data Analytics

SEN163A

Metadata

Jacopo De Stefani

Based on the material by Tobias Fiebig

TU Delft

Content notification (CN)

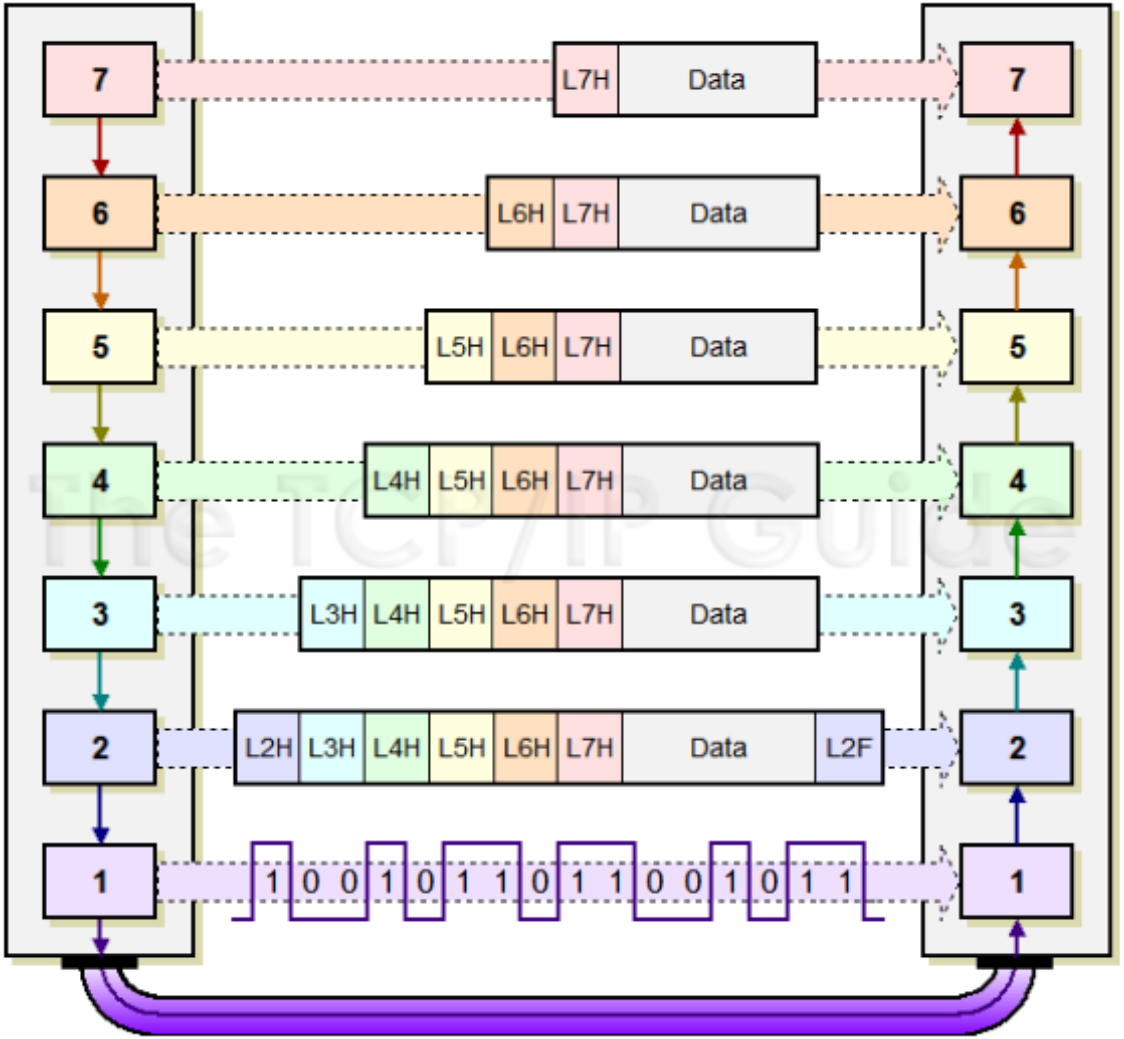
- Privacy violation
- Surveillance
- Pregnancy and abortion
- Symbols of fascism
- Genocide
- Prosecution of LGBTQ+ people

- 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.

TU Delft

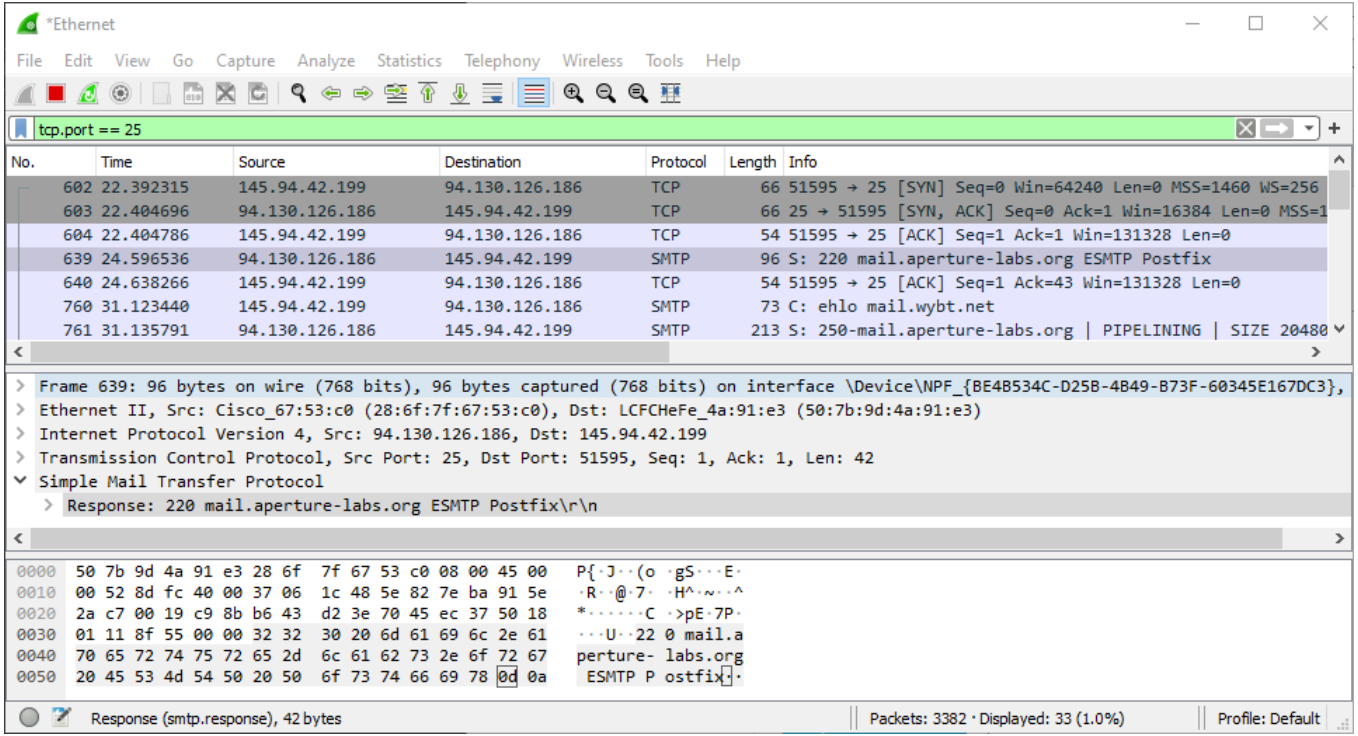
Not all data is alike

- In network measurement:
  - Metadata
  - Payload



TU Delft

Payload in a TCP packet



Metadata is protocol dependent

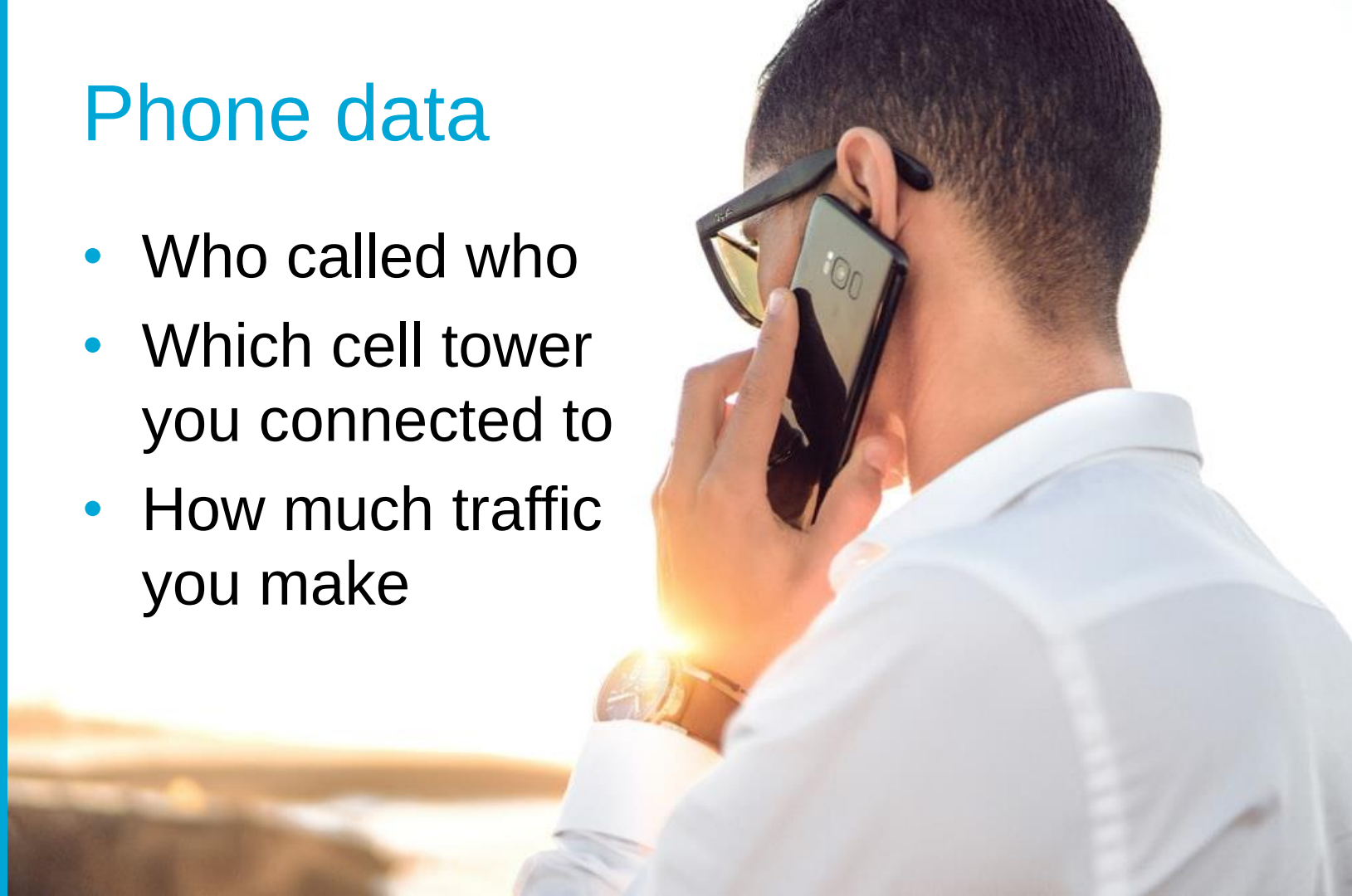
- For IP packets, TCP is already payload
- For TCP packets the protocol data is payload
- For an email, sender and recipient are metadata

Metadata of an email

```
Return-Path: <T.Fiebig@tudelft.nl>
Received: from mail.aperture-labs.org
        by mail.aperture-labs.org with LMTP
        id yPO3LPkVVV49awAAj3/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



TU Delft

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…

TU Delft

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… ;-))

TU Delft

Event 1: A received call

Google

+31 15 2569320

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 ...

TU Delft

Ok… interesting…

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

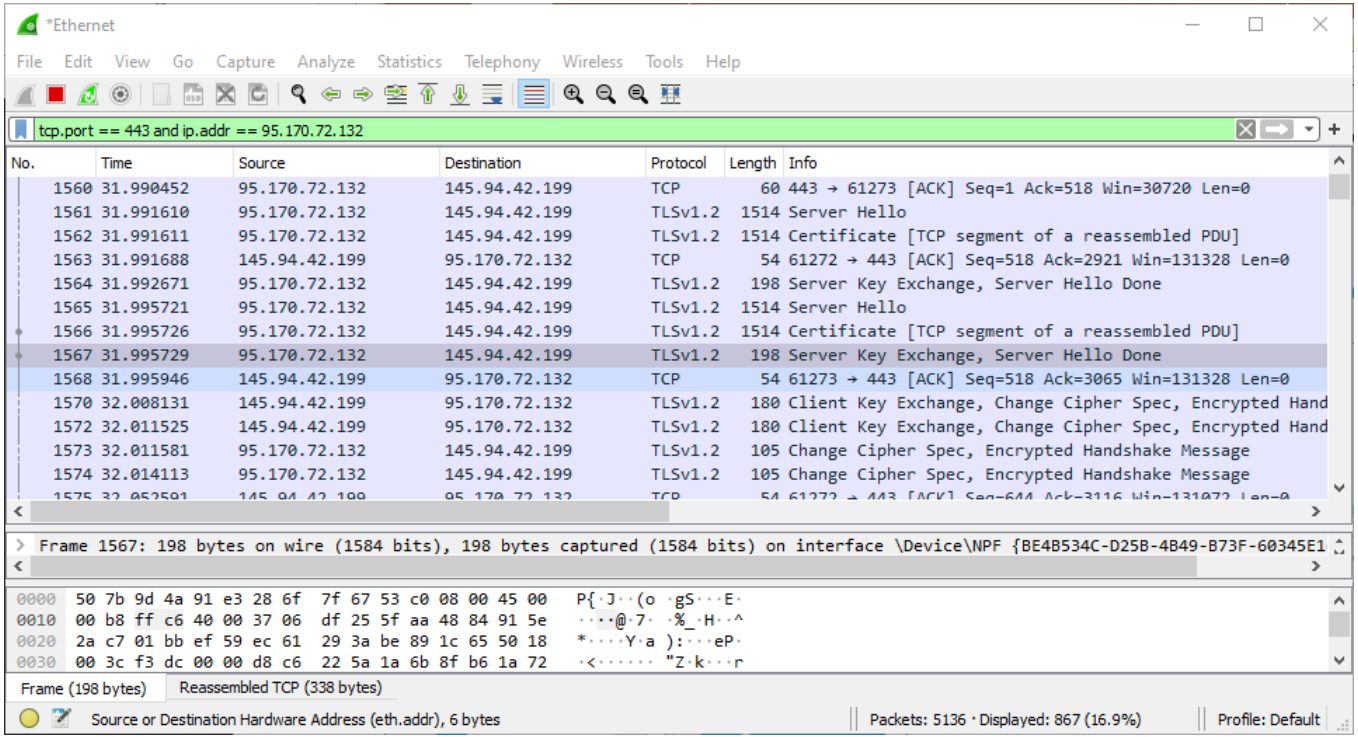
Event 2: An uncommon call

- +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



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.
```

TU Delft

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

TU Delft

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

TU Delft

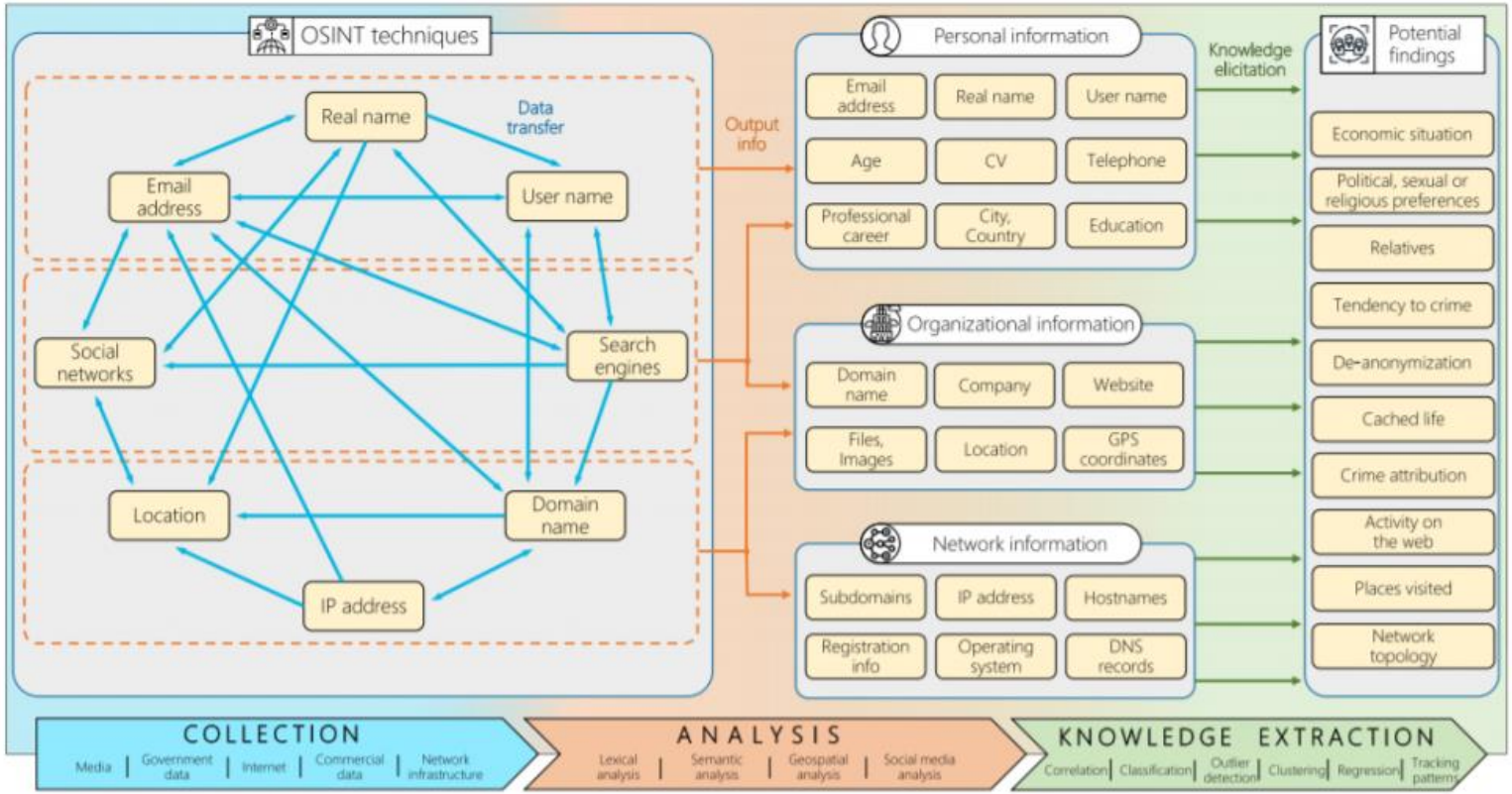


FIGURE 2. Principal OSINT workflows and derived intelligence.

TU Delft

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.

TU Delft

A story of not so meta-data

TU Delft

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 20th century!

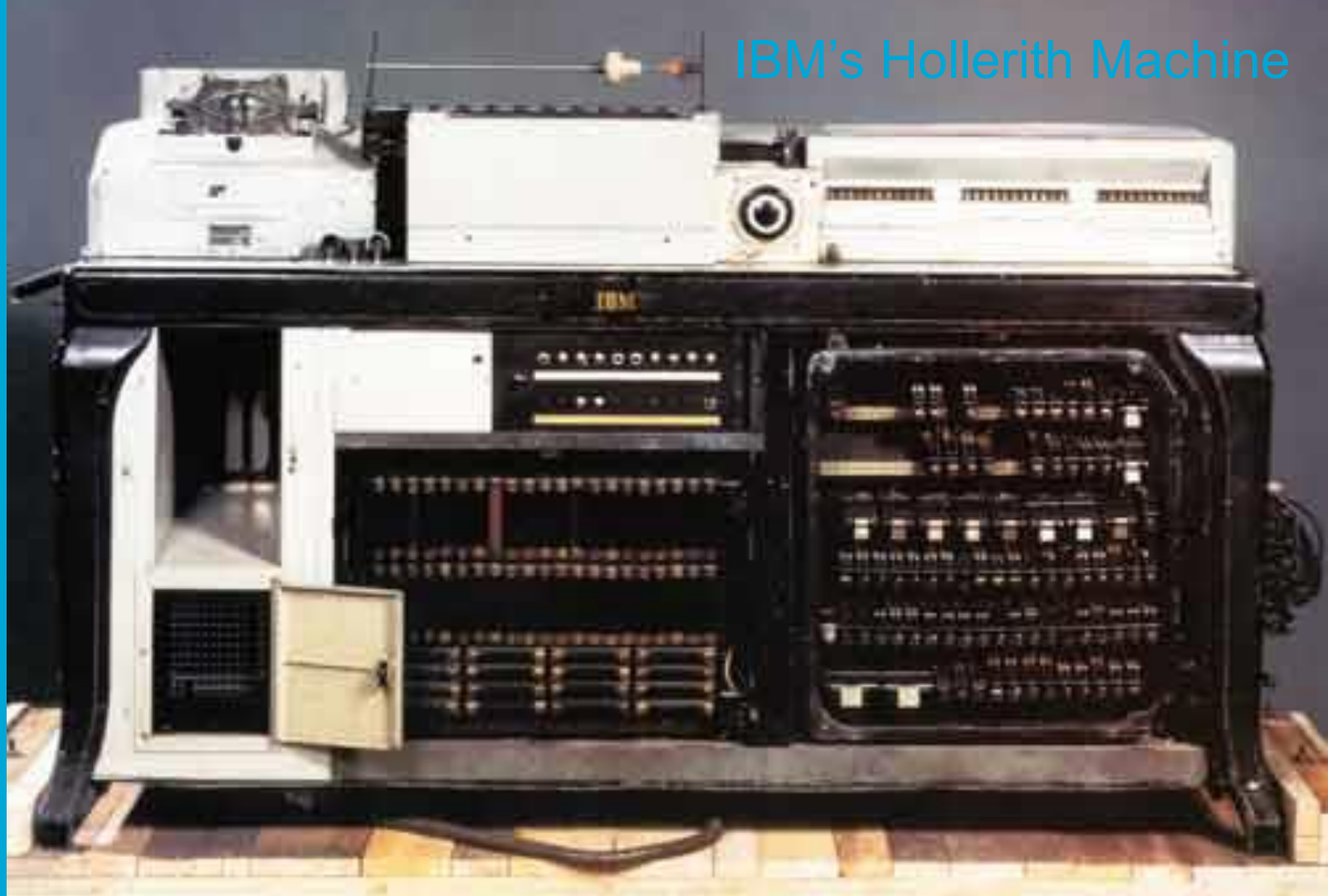
TU Delft

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

TU Delft

IBM's Hollerith Machine



TU Delft

TU Delft

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?

TU Delft

Invasion…

TU Delft

And then the new cards looked like this

TU Delft

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!

TU Delft

This is why the Dutch resistance…

TU Delft


…mostly burned archives.
TU Delft

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://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.

TU Delft


Break

But we don’t collect that anymore!

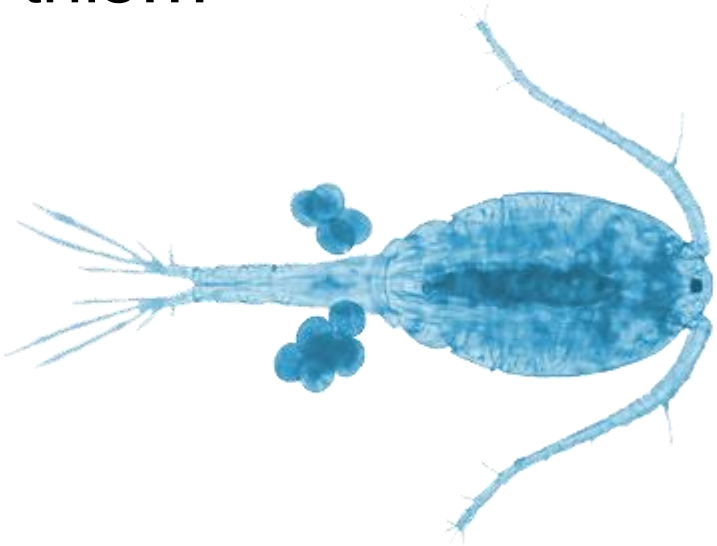
- Dutch authorities may no longer collect religion. ☺
- So, things are good...
- …
- … right?

But we don’t collect that anymore!

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

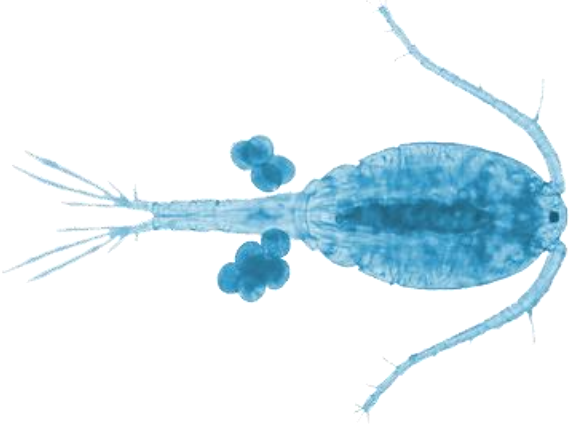
Little excursion…

- What is this…



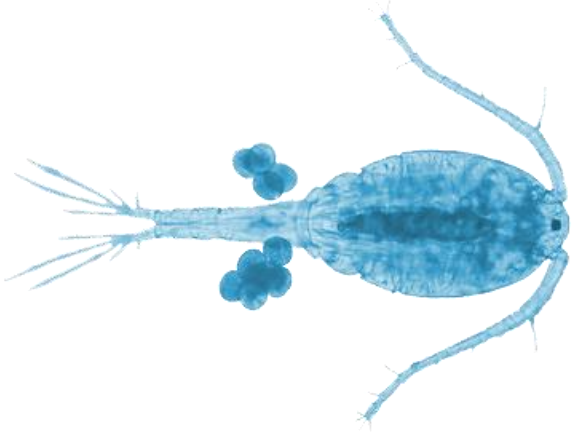
This is a copepod

- 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.



‹ 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

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

Share

Deliver to Netherlands

See All Buying Options

Add to List

Have one to sell? Sell on Amazon

TUDelft

What data do you have…

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

TU Delft

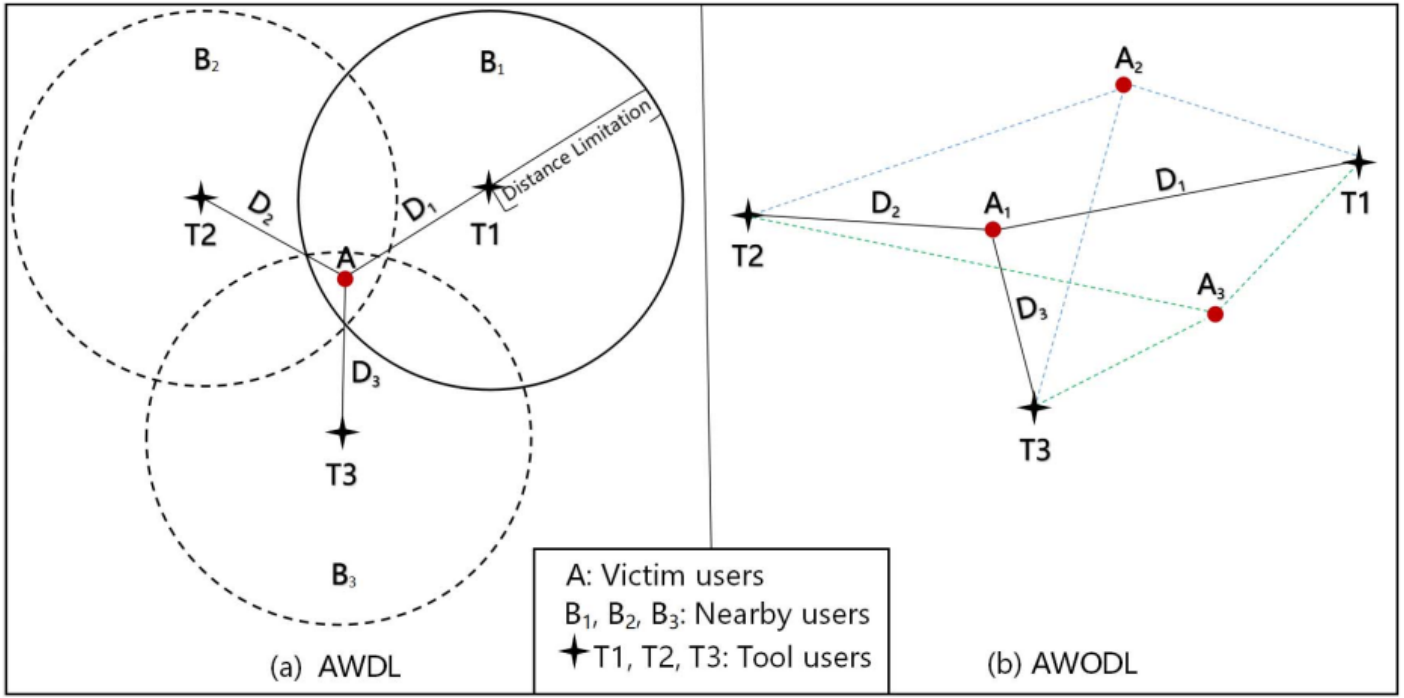
Who knows Tinder?

TU Delft

Who knows… Grindr?

TU Delft

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…

TU Delft

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…

TU Delft

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

TU Delft

Let’s look at another metadatapoint…

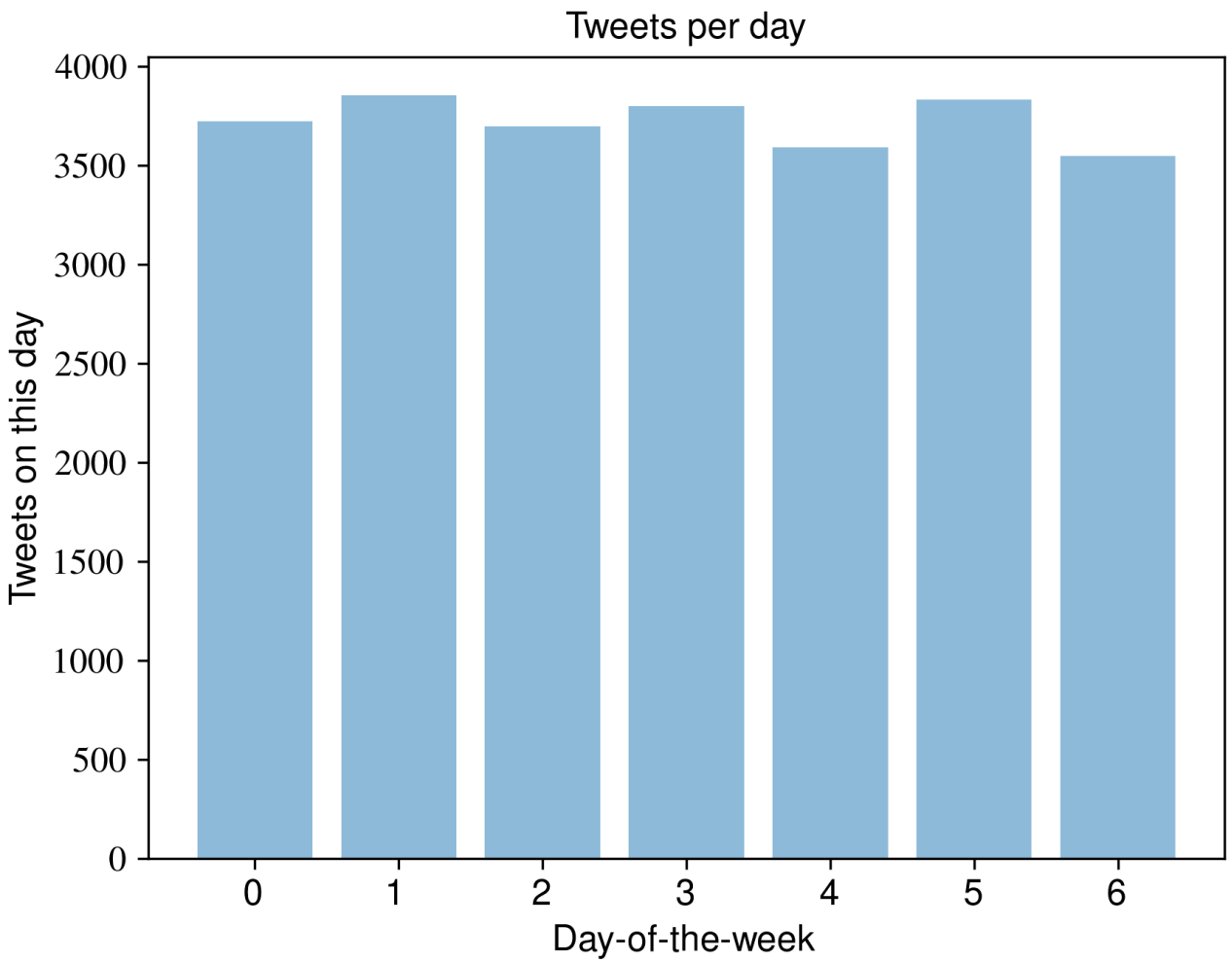
- Wed Mar 4 14:17:46 CET 2020

Question

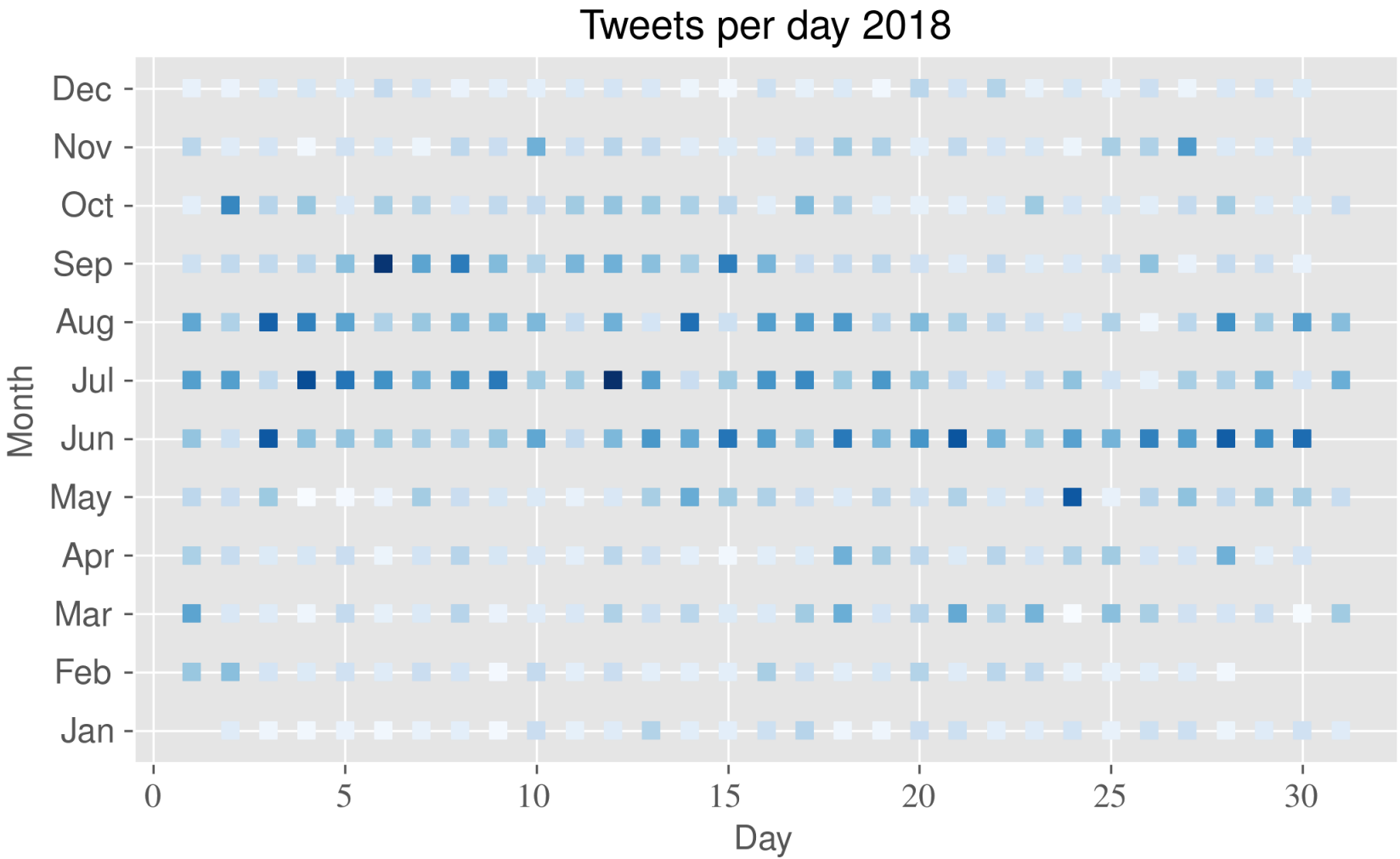
What can we learn from this?

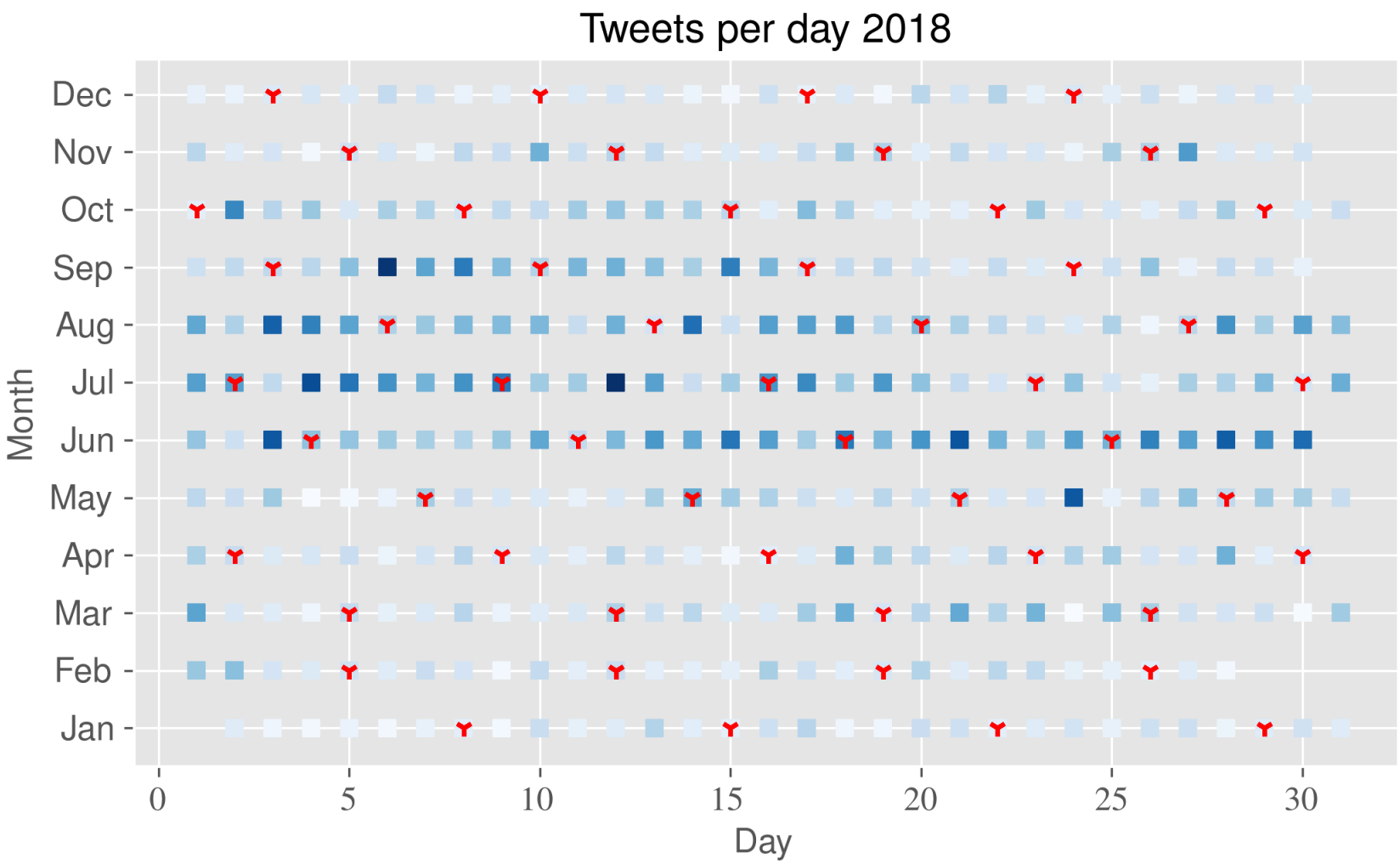
Let’s try…

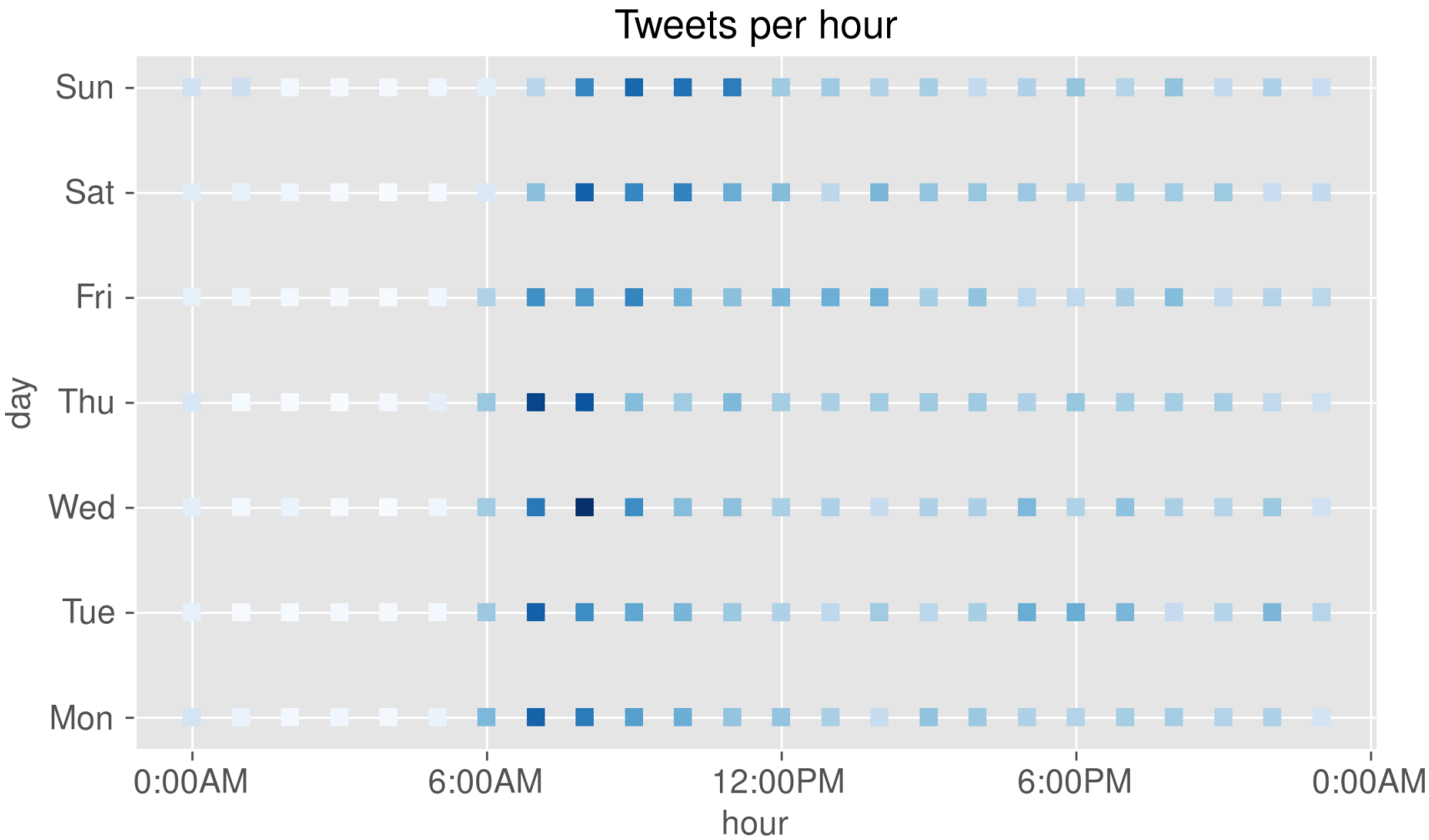
- What we have: 26,045 tweets from a single person over 7 years
- Let’s see what we can… see

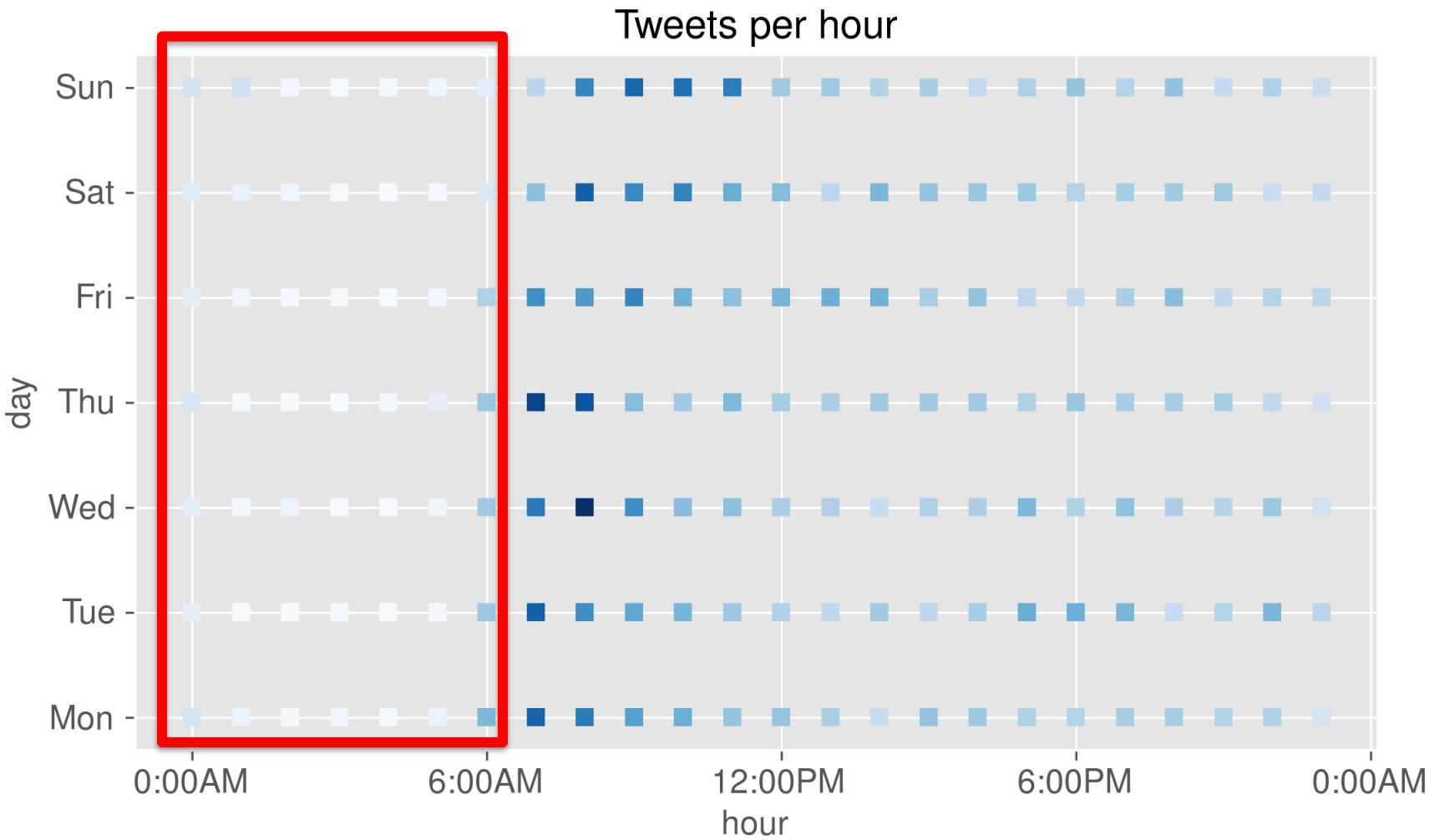

Tweets per day
Tweets on this day
4000
3500
3000
2500
2000
1500
1000
500
0
0
1
2
3
4
5
6
Day-of-the-week

TU Delft

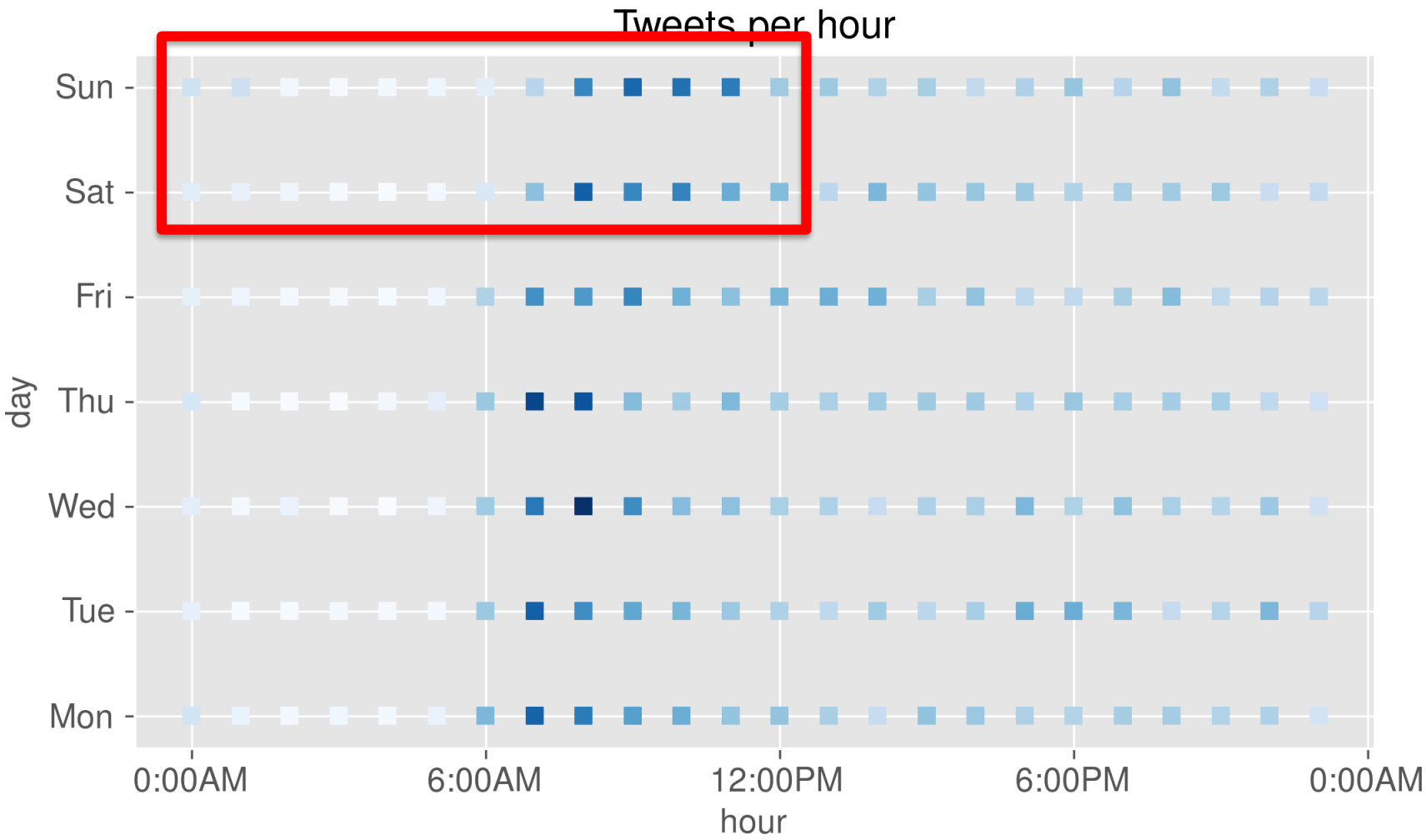

Tweets per day 2018
Dec
Nov
Oct
Sep
Aug
Jul
Jun
May
Apr
Mar
Feb
Jan
Month
0
5
10
15
20
25
30
Day


Tweets per day 2018
Dec
Nov
Oct
Sep
Aug
Jul
Jun
May
Apr
Mar
Feb
Jan
Month
0
5
10
15
20
25
30
Day

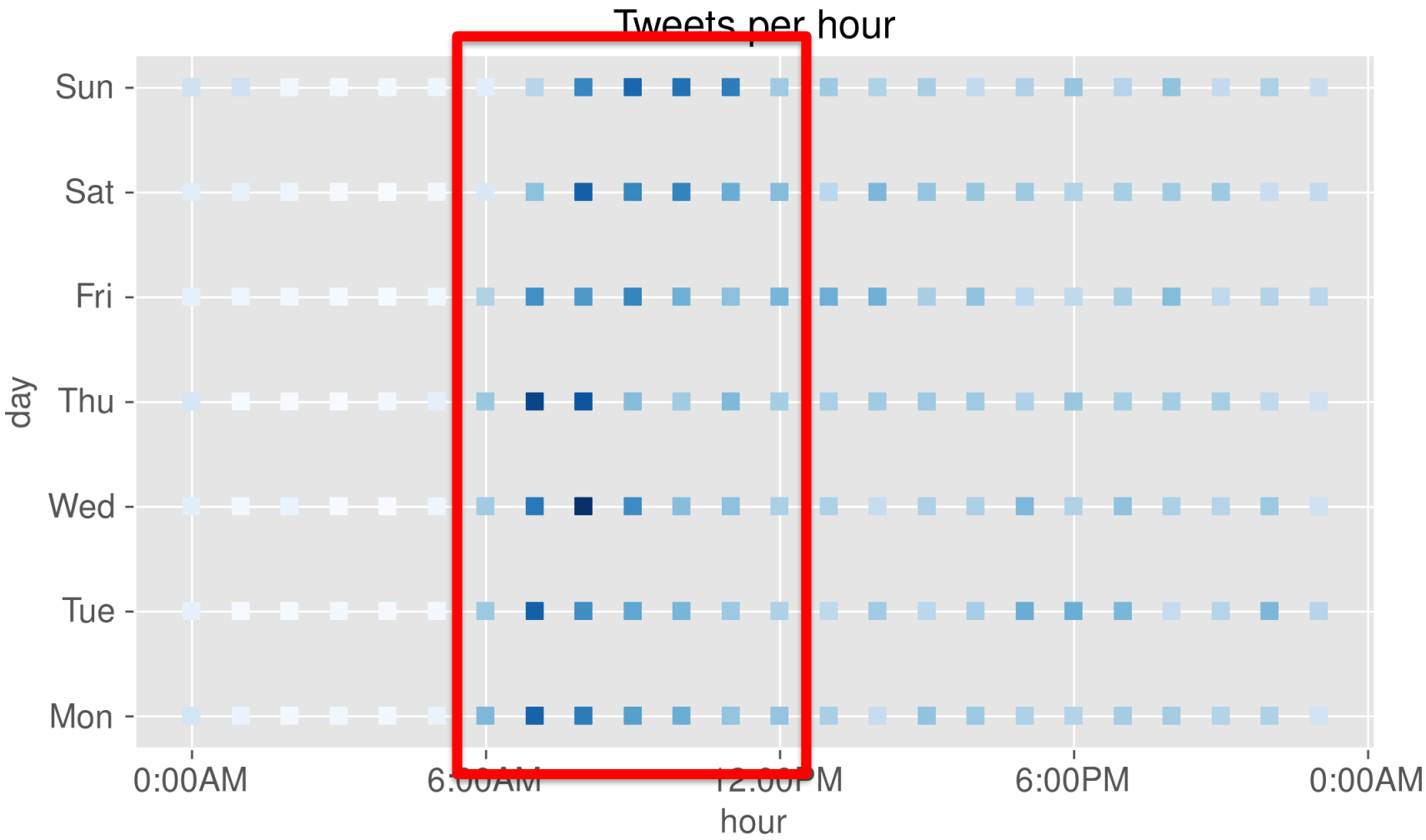

Tweets per hour
Sun
Sat
Fri
Thu
Wed
Tue
Mon
day
0:00AM
6:00AM
12:00PM
6:00PM
0:00AM
hour


Tweets per hour
Sun
Sat
Fri
Thu
Wed
Tue
Mon
day
0:00AM
6:00AM
12:00PM
6:00PM
0:00AM
hour

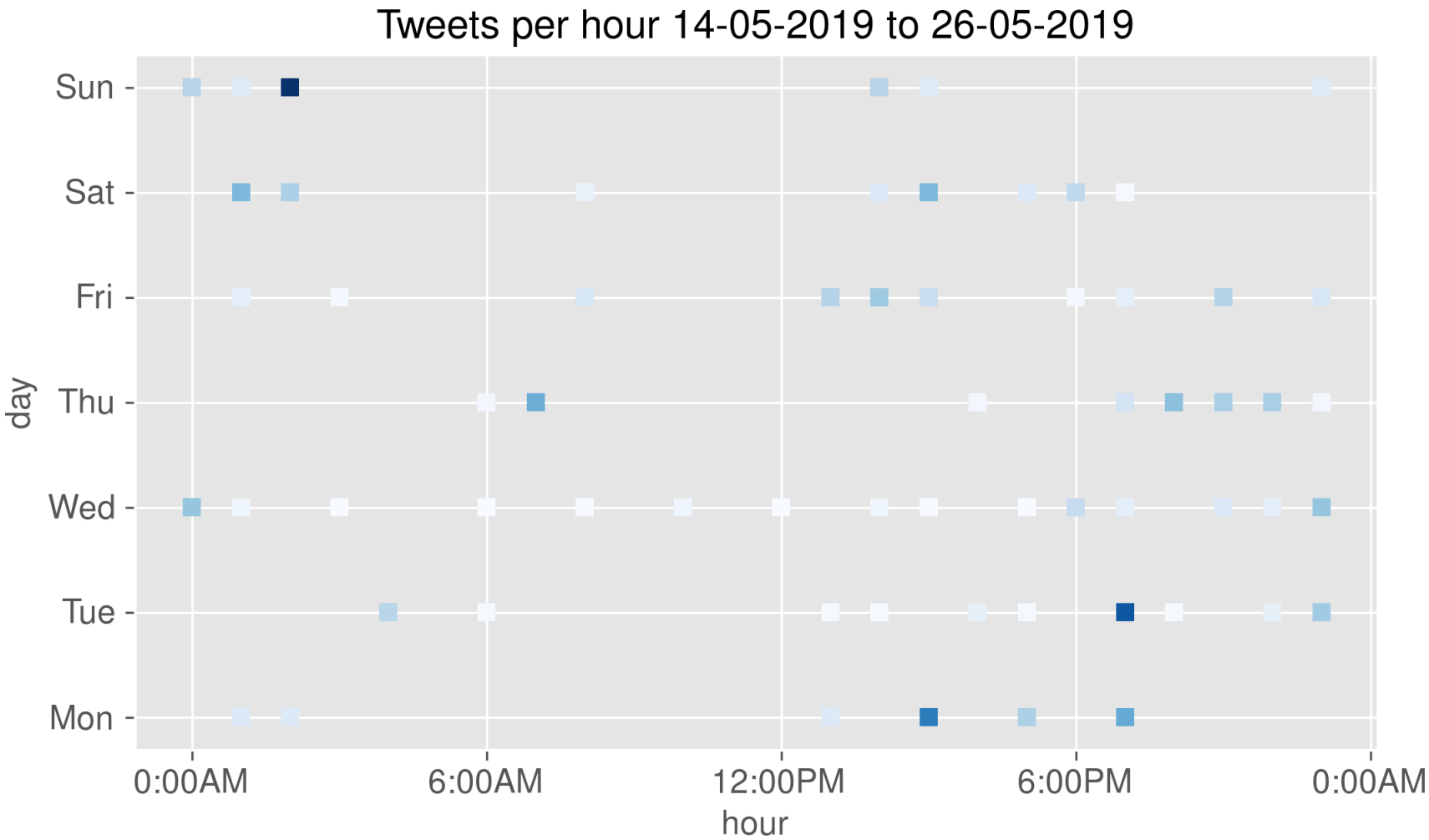
Sleepingtime…


Tweets per hour
Sun
Sat
Fri
Thu
Wed
Tue
Mon
day
0:00AM
6:00AM
12:00PM
6:00PM
0:00AM
hour

Sleeping in on weekends…


Tweets per hour
Sun
Sat
Fri
Thu
Wed
Tue
Mon
day
0:00AM
6:00AM
12:00PM
6:00PM
0:00AM
hour

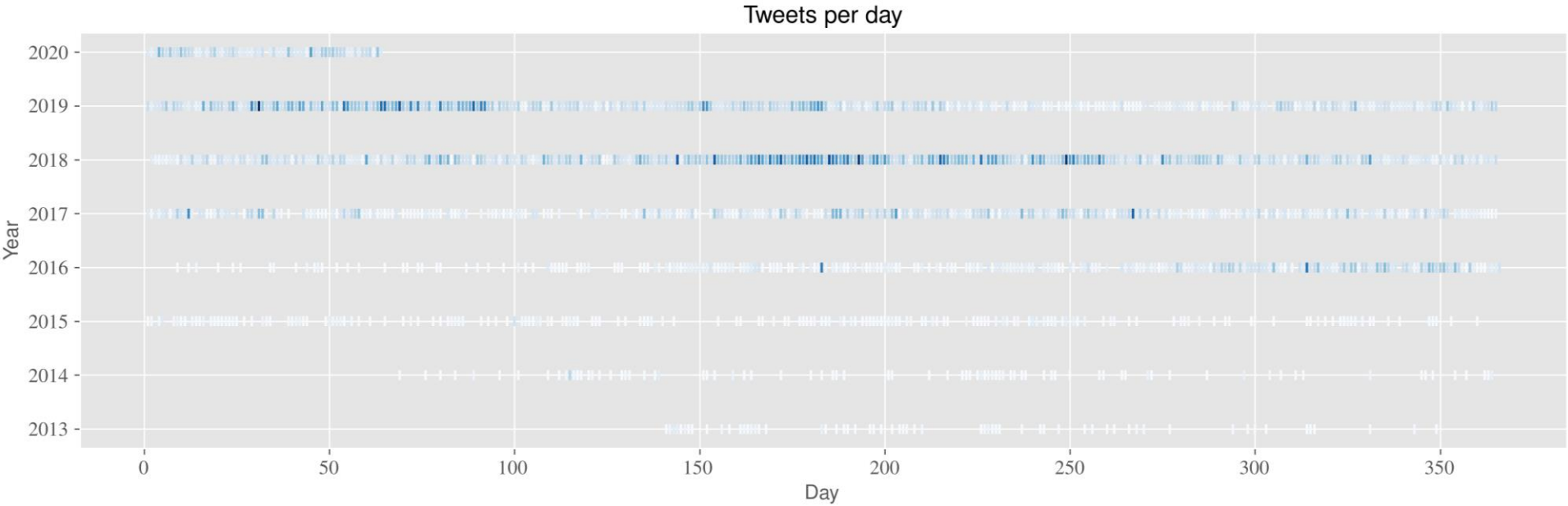
Well… coffee and newspapers…
I am not a boomer…


Tweets per hour 14-05-2019 to 26-05-2019
Sun
Sat
Fri
Thu
Wed
Tue
Mon
day
0:00AM
6:00AM
12:00PM
6:00PM
0:00AM
hour

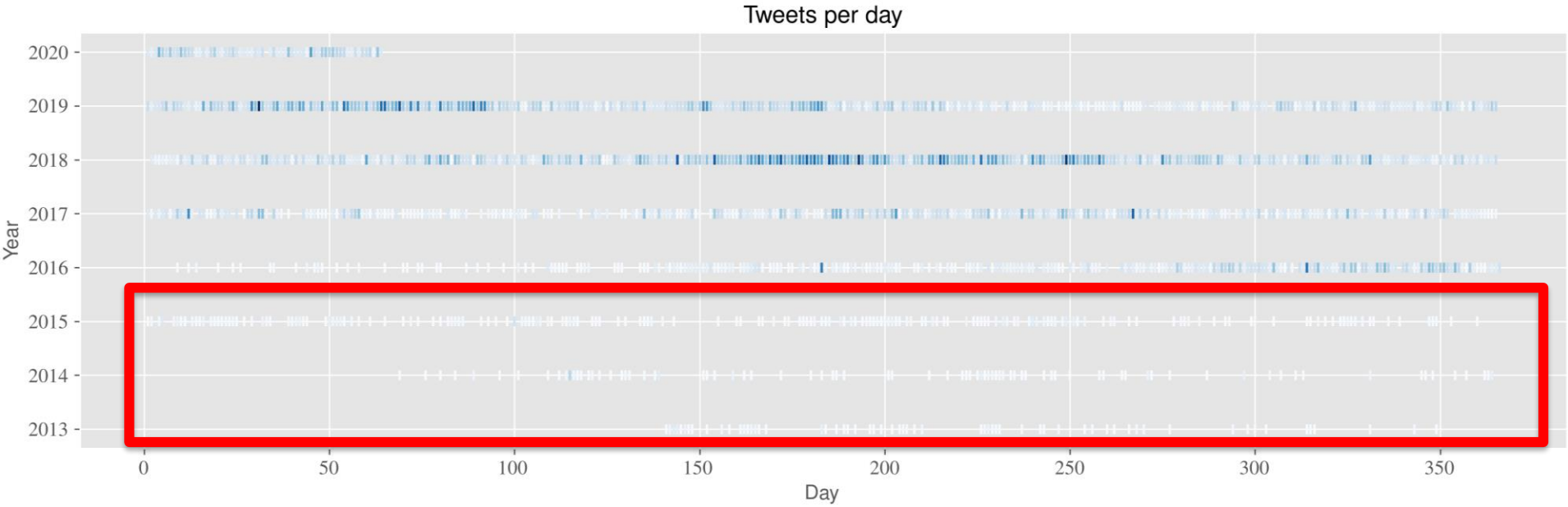
These weeks look different…

These weeks look different…

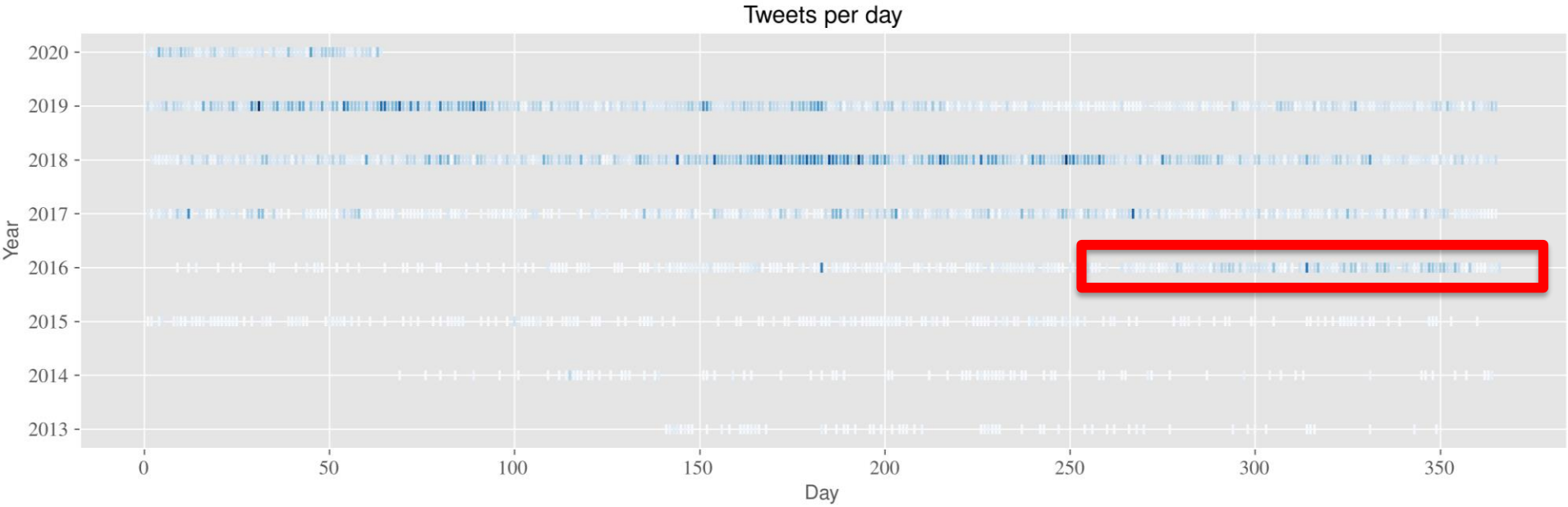
- Went to the US for two conferences that week
- No time for tweeting
- Highly different timezone


Tweets per day
Year
2020
2019
2018
2017
2016
2015
2014
2013
0
50
100
150
200
250
300
350
Day

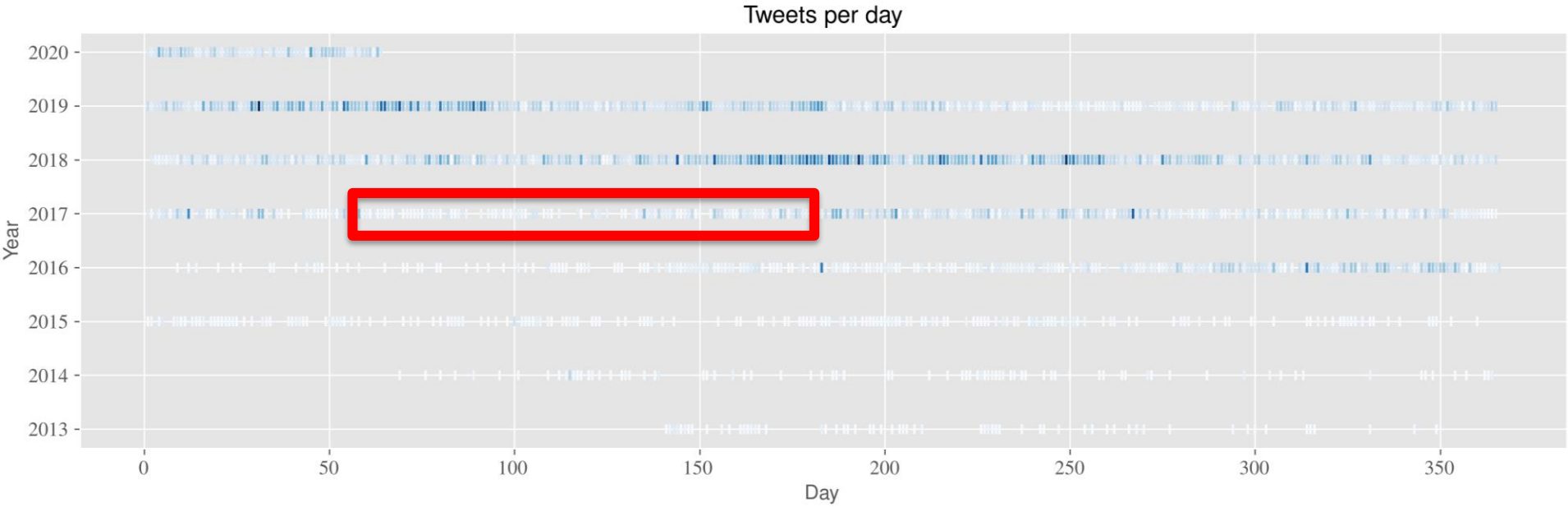
TU Delft

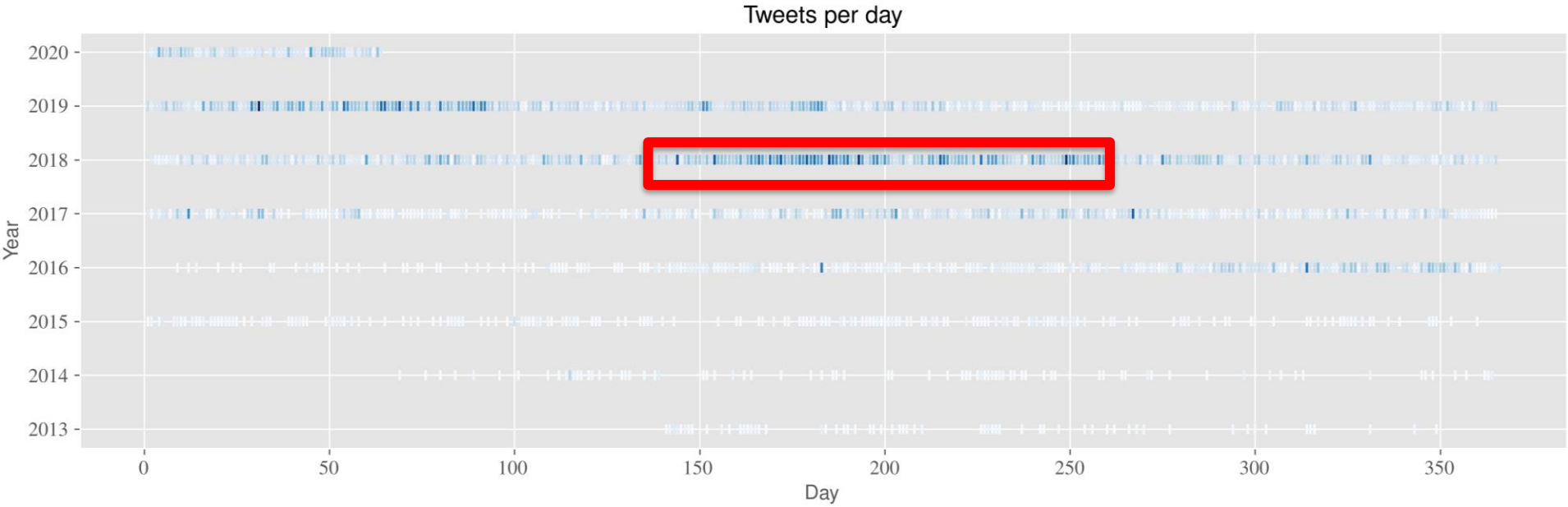

Tweets per day
Year
2020
2019
2018
2017
2016
2015
2014
2013
0
50
100
150
200
250
300
350
Day

TU Delft

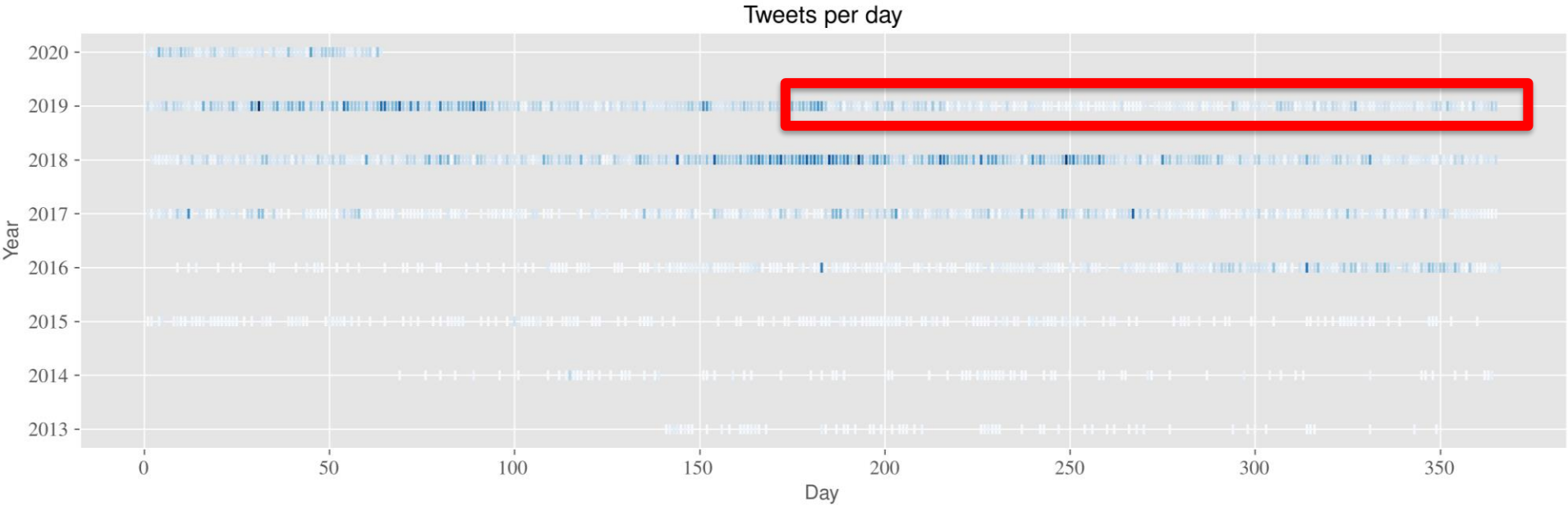

Tweets per day
Year
2020
2019
2018
2017
2016
2015
2014
2013
0
50
100
150
200
250
300
350
Day

TU Delft


Tweets per day
Year
2020
2019
2018
2017
2016
2015
2014
2013
0
50
100
150
200
250
300
350
Day


Tweets per day
Year
2020
2019
2018
2017
2016
2015
2014
2013
0
50
100
150
200
250
300
350
Day

TU Delft


Tweets per day
Year
2020
2019
2018
2017
2016
2015
2014
2013
0
50
100
150
200
250
300
350
Day

TU Delft

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")
```

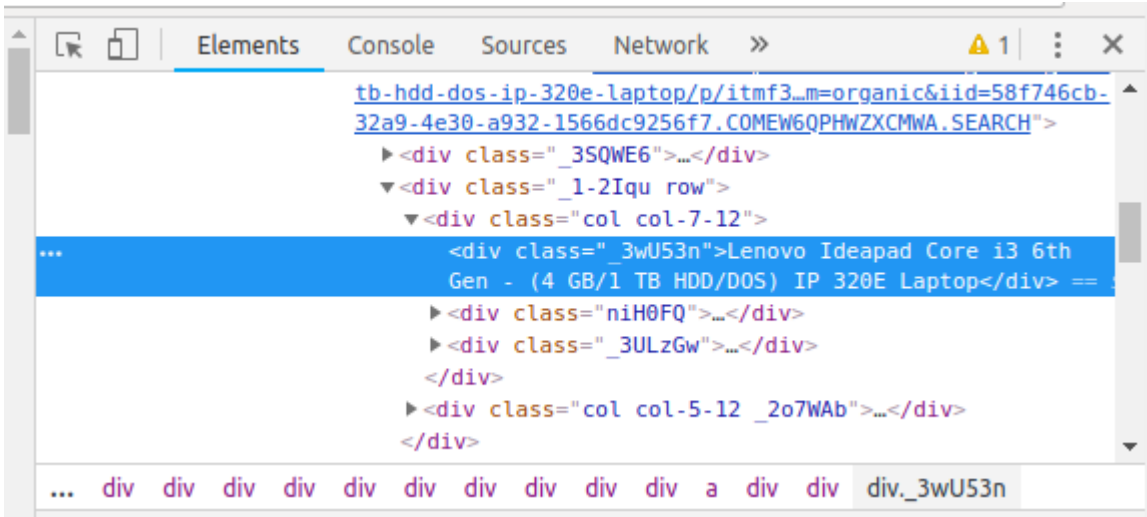
From the online example

- 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 s
- Klick ‘inspect tab’ (or hit F12)
- Identify the right names:



div div div div div div div div div div a div div div._3wU53n

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':'_1vC4OE _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’
```