CSC2552 Project: The Digital Deluge

Thomas Hollis

University of Toronto

thollis@cs.toronto.edu

April 3, 2019

Overview

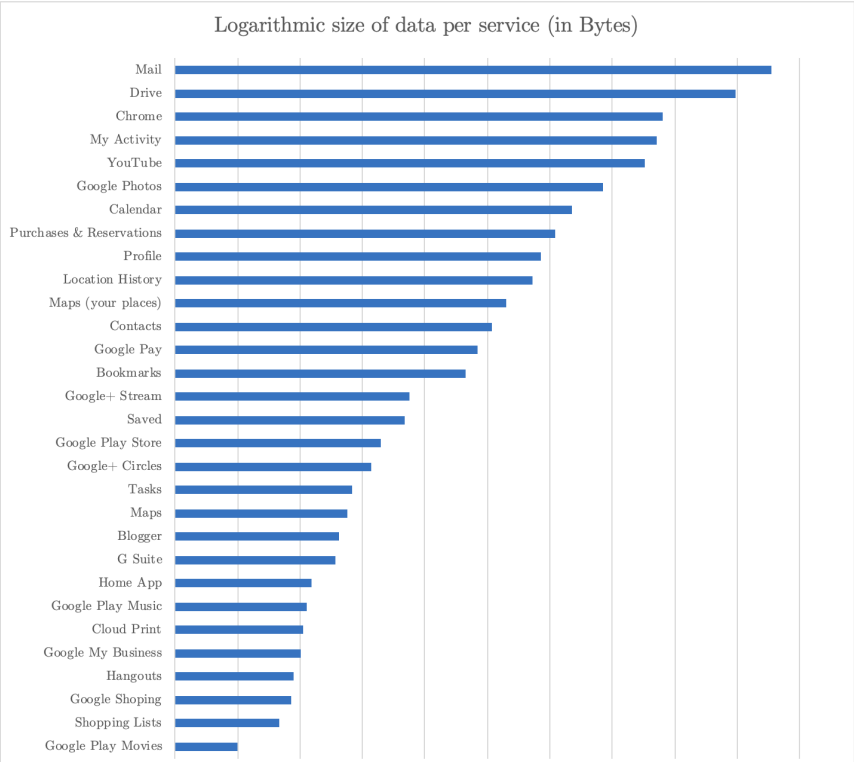
1.1 Research Question

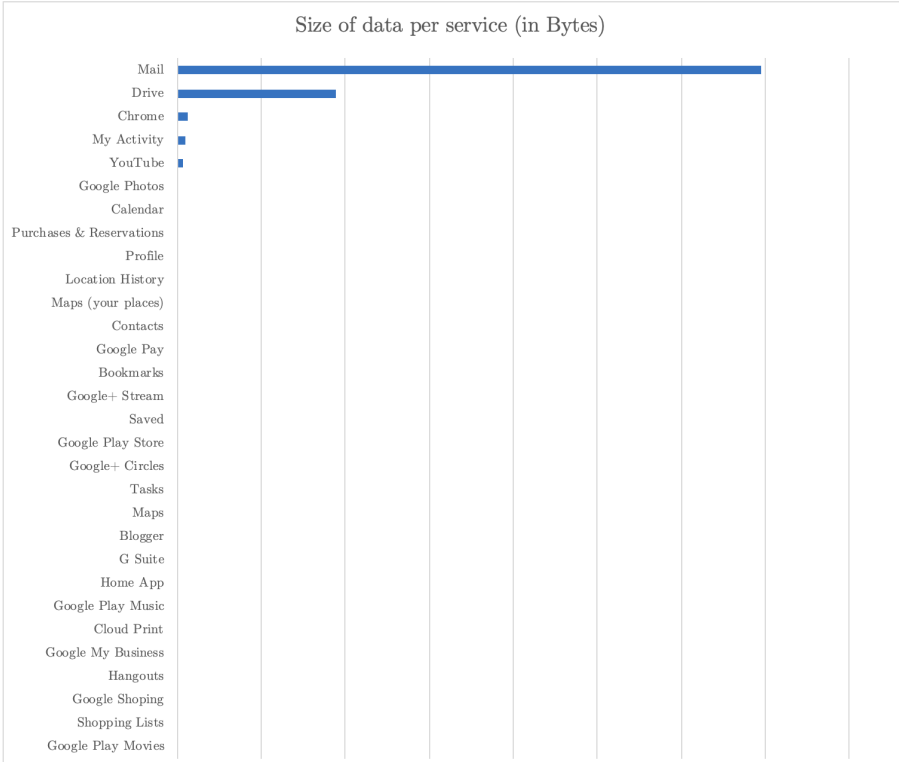
- "*How much can GDPR queries reveal about the use of consumer data in tech companies in 2019?*"

1.2 Methodology

- *Armada strategy* inspired by BitByBit
- First: general overview (shown in proposal presentation)
- Second: detailed investigations (focus on this presentation)

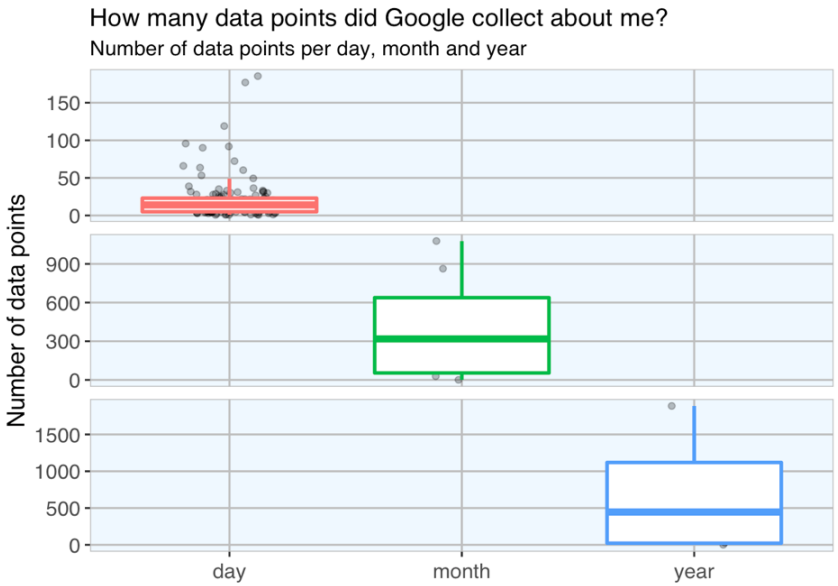
2.1 Results: Google Results

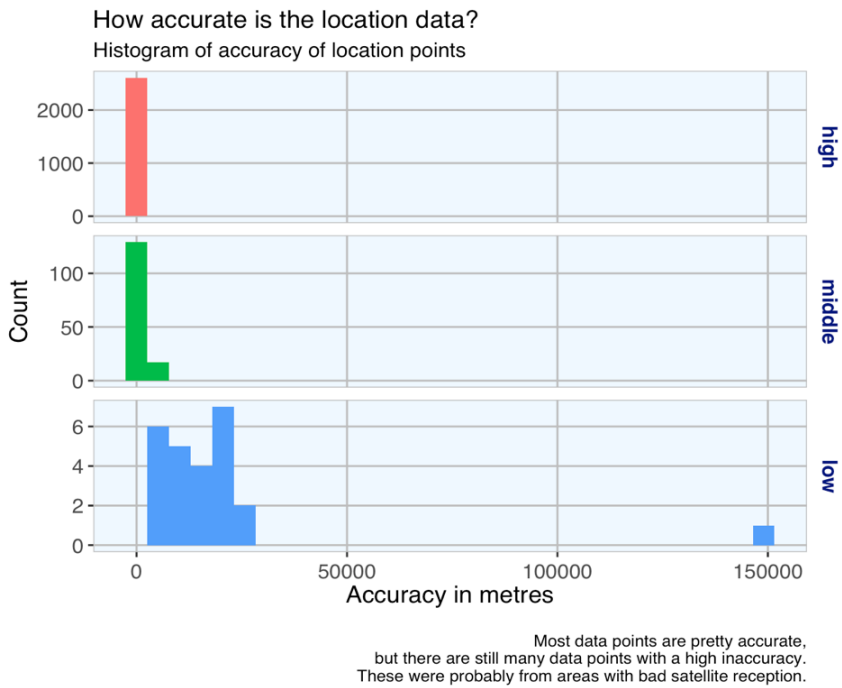




- Strength from using breadcrumb data (Maps vs. Drive 2GiB)
- Most worrying details do not take up the most volume (Maps/Search)
- Example of a troubling log: all YouTube searches since account creation

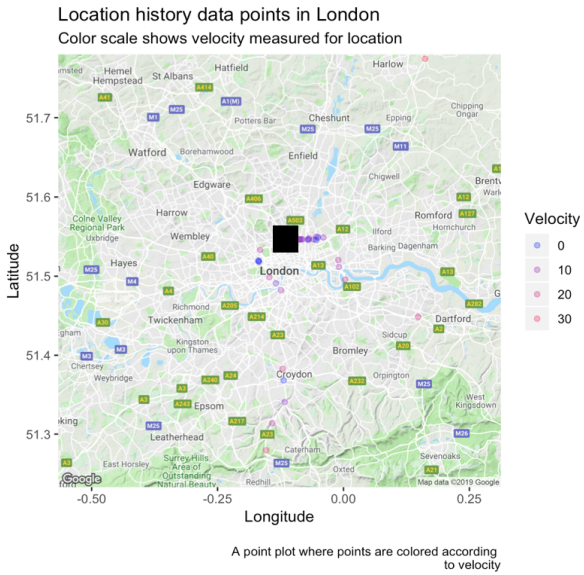
2.1 Results: Google Results

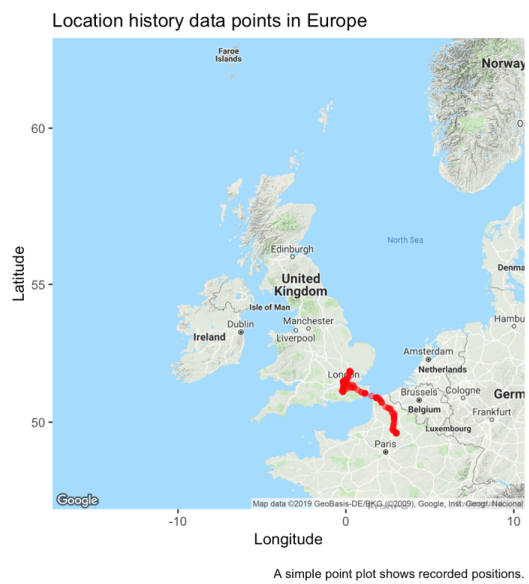




- Default setting was to have Google Timeline on
- Accuracy of measurements mostly high (nearly all - no user-control)
- Above is just using a few months of data (duration of experiment)

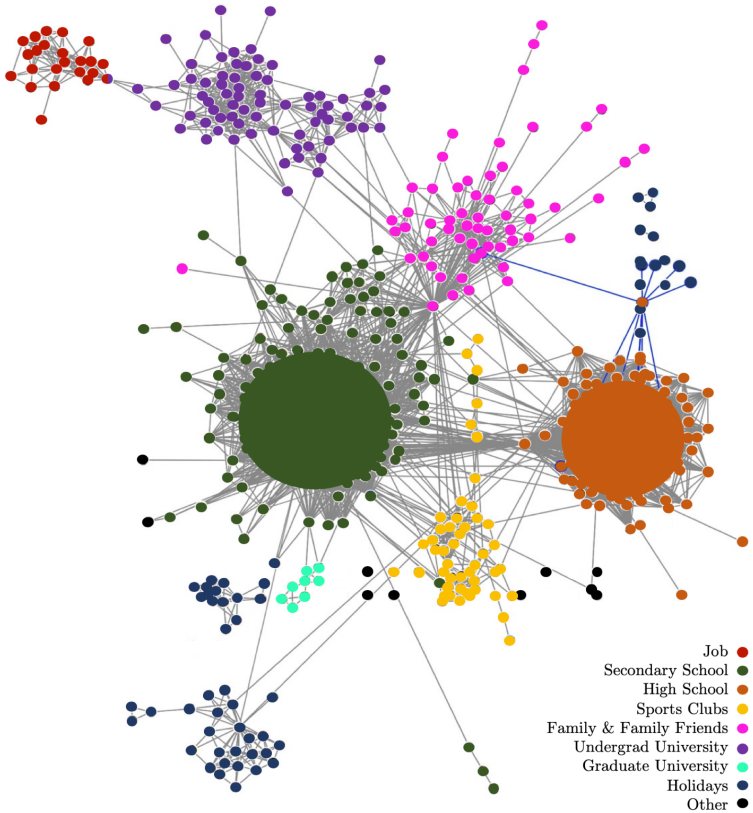
2.1 Results: Google Results

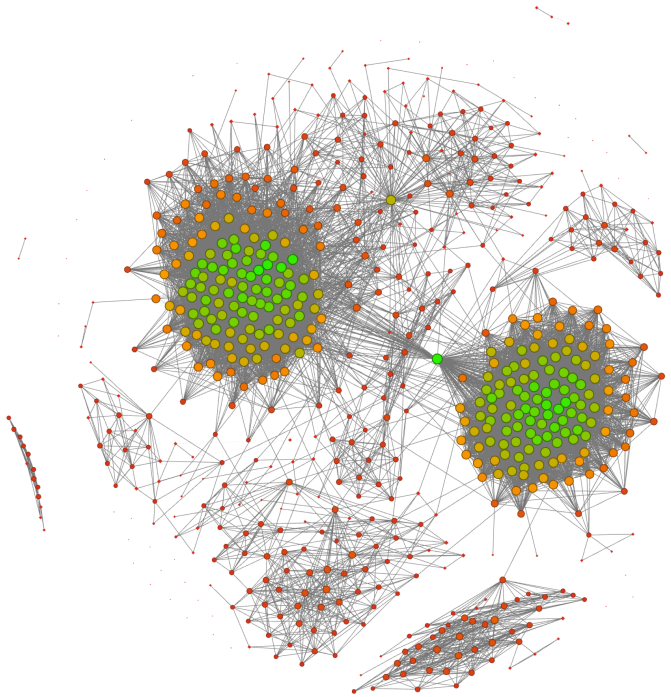




- Can use the location tracking to plot velocity on a map
- Accuracy is usually higher near home, data can track address
- Continental travelling (even under the channel) is also tracked

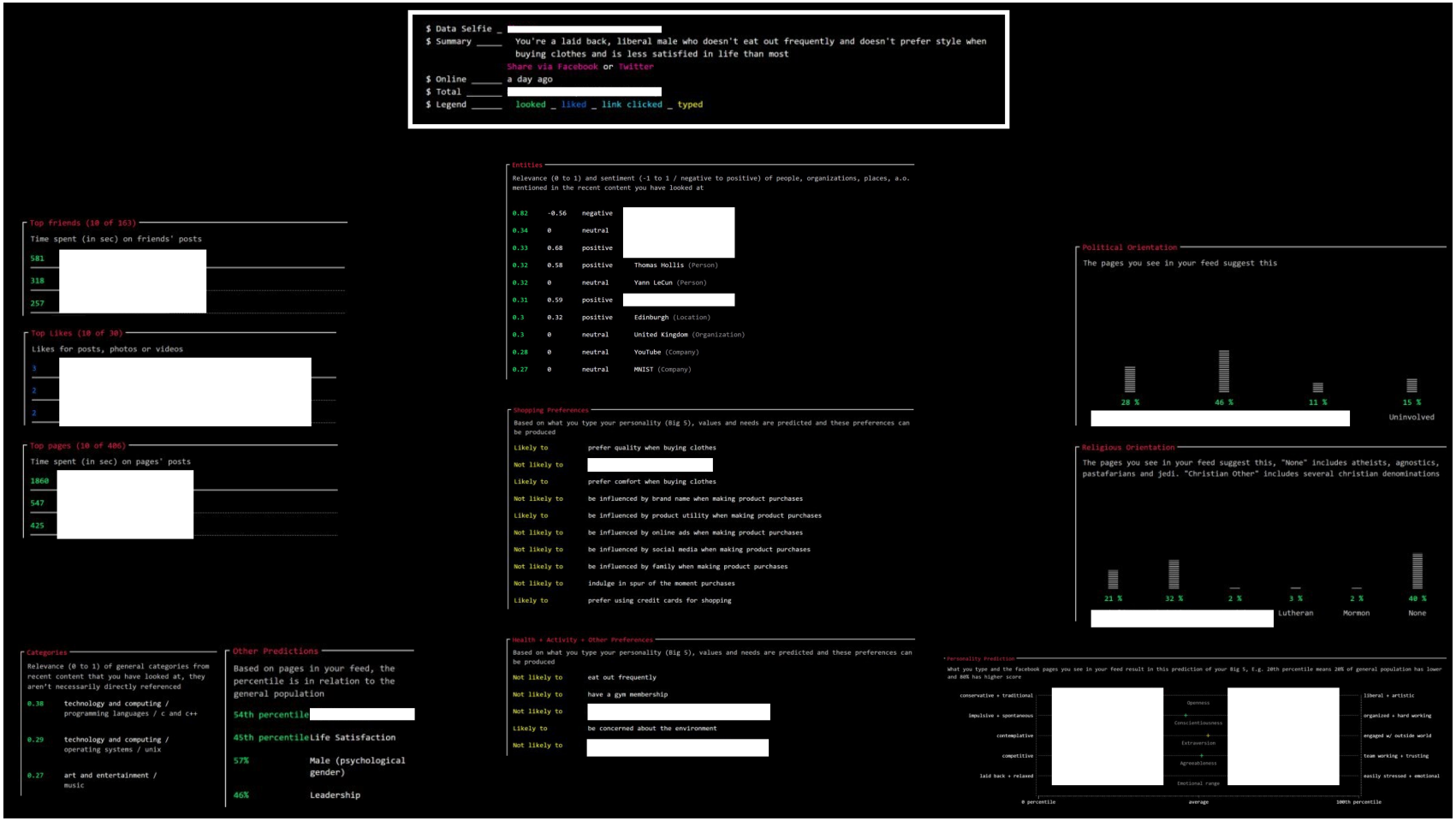
2.2 Results: Facebook Results





- Incredibly easy to find out social clusters by weighting mutual friends
- Can infer how you know someone and what the interaction is like
- Worrying that Facebook don't show us this even if they have the data

2.2 Results: Facebook Results

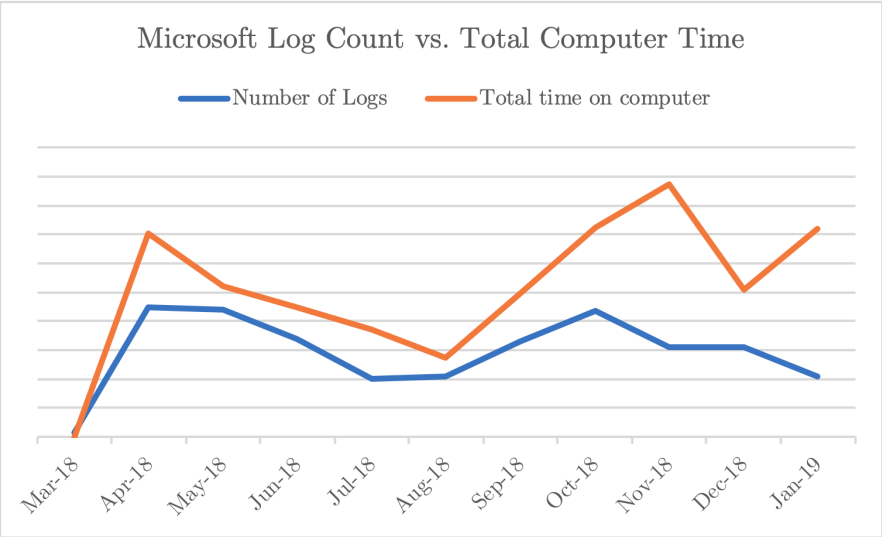


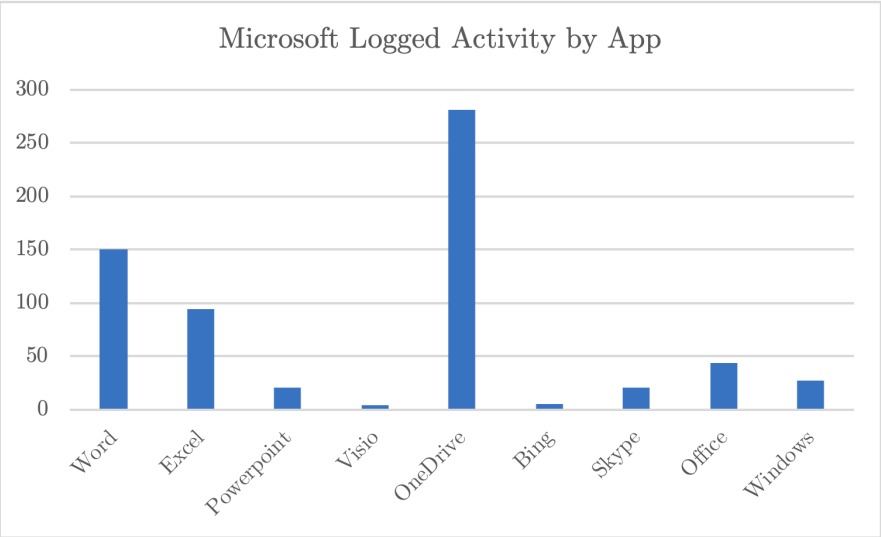
- Incredibly easy to infer accurate personality traits from very little data
- Possible even without interactions, just timing your attention

2.2 Results: Facebook Results

- Can also infer friend's attributes, using only your own data e.g. political leanings from likes (shared b.d.)

2.3 Results: Microsoft Results





- Most of the logging is using the OneDrive app (in my case)
- Moved to macOS in July and quit my OneDrive subscription $\rightarrow$ still tracked (not used Windows, own iPhone $\rightarrow$ MSOffice & OneDrive?)
- Forecasting total time on computer from logged data gets much less accurate after switching to Apple in July

2.4 Results: Other Results


```
> babble(ng = ng5, genlen = 100)
[1] "help hey thomas i was just writing my resume and i need a lil help from you
so it s regarding the template of the resume generally i write my resume in latex
a traditional software engineering one but i was thinking if i can do something d
ifferent this time what do you think i should pick some template or design my own
```

```
> babble(ng = ng3, genlen = 100)
[1] "for you guys to come visit me in toronto but wait till my exam timetable is
out because it might mean that you can come earlier than april th it better mean
```

- Ask me about a company at the end (or after class)!
- Most companies collect far more data than they need for their services
- Often able to infer very sensitive information about self or friends
- Often users are unaware of what is being recorded (Searches, IP...)

3.1 Conclusion: Opinion changes

- Survey changes reveal I became even more paranoid than I already was w.r.t. data
- Overestimated the volume of PII that was collected on me
- Overestimated the number of companies that store PII on me
- I made a good estimate of company compliance rates (around 50%)

3.2 Conclusion: General findings

- Google: Biggest concern is the greediness of the data collections (should be the opposite - err on the side of caution). "Collect first, worry later" approach (especially for low volume data) is an ethical concern not in line with "minimise harm" in Salganik's BitByBit
- Facebook: Overall I am more worried about the potential of the data than the data itself (changed my methodology accordingly)
- Microsoft: Fundamental issue of trust remains unresolved (cannot trust GDPR output is complete, cannot trust right to be forgotten...)
- There is a lot of information I found out that I could not include in this report $\rightarrow$ encourage everyone at some point in their life to look into their GDPR data (I open sourced all the tools I used or wrote + extra tools like "Noiszy" and "Random User Agent" to counter effects).

Can I open-source the full report?

Questions?