What is the Internet doing to me? (witidtm 2021/2022 - TEU00311)

Lab Session #1

Stephen Farrell
stephen.farrell@cs.tcd.ie

https://github.com/sftcd/witidtm
https://down.dsg.cs.tcd.ie/witidtm

Public Service Announcement

- Don’t be here if: symptoms, waiting for test result etc.
- Face coverings (e.g. masks) are mandatory (except: verified medical reason)
- Maintain 1m distance in lab
- Clean your workstation
- Take a pic of your seat number/surroundings (potential contact tracing)
- We have to stop at :45, so we’ll start at :00 sharp
- Many more details at:

  https://www.tcd.ie/about/coronavirus/phased-reopening-plans/assets/student-induction.pdf
  https://www.tcd.ie/about/coronavirus/

F2F vs. Hybrid/Remote

- Definition:
  - “remote” == “we’re all online”
  - “hybrid/remote” == “many of us are in one room but some are still online”
- Reminders for me: 1) Check if anyone is hybrid/remote and see how they’d like to proceed.
  2) Don’t forget to Hit “record” !
- We’ll try help hybrid/remote folks do the same work as those in-person (but we’re learning as we go...)
  - Likely hybrid/remote people may need to send video for this to be useful
  - Remote video will default to being displayed in the lab, but I could turn off that screen for a bit if needed (video would still be sent to the WebRTC call though...)
- If the hybrid/remote experience sucks really really badly compared to being in the room, I’ll try organise a separate session for those who **cannot** make it to college (if you really “cannot” as opposed to “just didn’t”)

Today’s Goals

- Login to SCSS account and/or get your laptop online
- Fire up a “new” browser
- Explore settings
- Watch HTTP traffic (shift-ctrl-I)
- Find the “worst” site you can
  - NSFW disallowed, otherwise you define/justify “worst”

Stretch goal

- If we get the above done great, if not, we can do more next week
  - IOW – I’m very unsure how well this’ll go;-)
- Could be some people get to the stretch goal today or we look at it in future...
- Stretch goal: HTTP archive (.har) file generation and a bit of analysis

Login/get-online

- Desktop logins: no “domain”, use SCSS password, not your TCD password (if those differ), so e.g. if your TCD email is bloggsj14@tcd.ie then you enter “bloggsj14” as the username
- Get online: we’ll deal with things as they arise
- After you’re done: see if anyone else needs (1m distanced) help

I did check that username... :-)

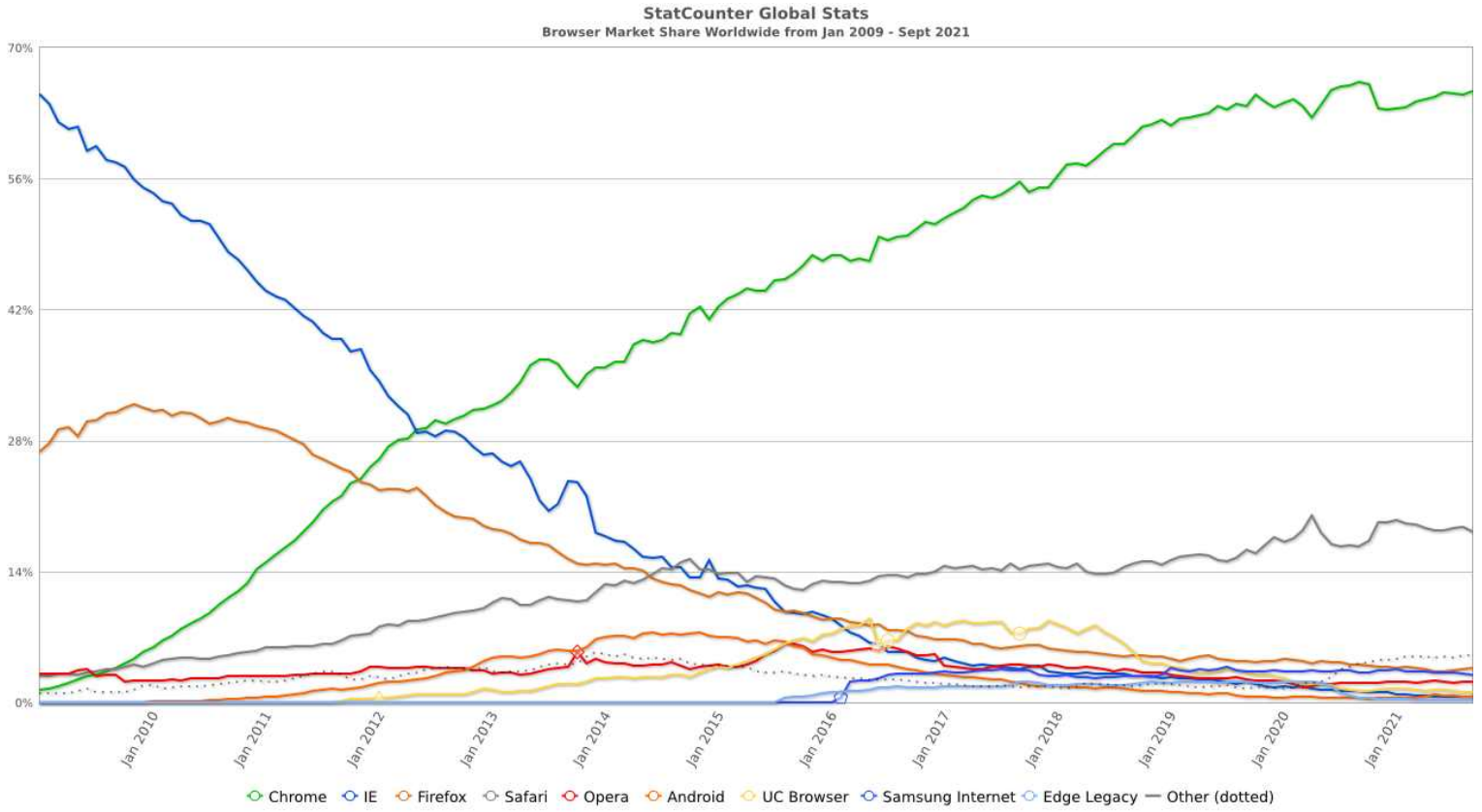
“Your message to bloggsj14@tcd.ie couldn't be delivered.
A custom mail flow rule created by an admin at tcdud.onmicrosoft.com has blocked your message.
5.1.1 The e-mail service at tcd.ie does not know this email address.”

GOTO “the web” slides 16-22

https://down.dsg.cs.tcd.ie/witidtm/lectures/2021-2022/200-web.pdf

Browser Hygiene

Browser worldwide market share



2009-2021

Chrome
Safari
Edge
Firefox
Opera
Android
UC browser
...
Brave
Vivaldi

https://gs.statcounter.com/browser-market-share#monthly-200901-202109

Overall browser landscape

- Browser defaults are chosen by browser implementers (Google, Mozilla, Microsoft, Apple, Handset vendors...)
  - Generally they allow Javascript and cookies, do telemetry, try get you to login, keep lots of state ...
- Historically, browser-makers seemed to care most about market share
  - Performance and rendering were their main concerns as they lose market share if they’re slower or sites don’t render (well)
- They started getting significantly better at security a while back (2013+)
- Some browser-makers are starting to get a bit better at privacy
- IMO they don’t behave as if they think you should be the one in control

Why browser hygiene matters...

- Developer of popular (300k installs) chrome ad blocking extension hasn’t time to keep maintaining that...
- Someone offers to buy the code and promises to maintain it...
- That someone adds malware to the code that steals cookies and session tokens, and maybe more...
  - That “someone” seems to be a repeat offender
- Result: 300k very unhappy people changing passwords all over and one very very embarassed original maintainer whose name is now mud (for some).
- Happened last year: https://github.com/jspenguin2017/Snippets/issues/2

My browser setup (1)

- Default browser: FF “nightly” + NoScript/Ghostery & disallowing cookies, with some white-listed sites, and search via DuckDuckGo (“!g” works too, if needed:-)
  - This is the only browser that saves logins, but not for sensitive things (we’ll consider passwords later)
  - Some sites don’t work with the above; mostly: screw ‘em
- Opera for managing home network
- Tor Browser: If searching for anything sensitive (e.g. medical info)
- If-need-be: vivaldi or chromium/incognito with no write-access to disk and so that it shoots it’s own brain out on exit (at least I hope so;-)
  - Use that e.g. for airline/hotel bookings
- If-all-else-fails: Brave

My browser setup (2)

- If I really have to watch some crappily DRM protected video, sometimes googlechrome (which differs from chromium on linux).
- On phone: Sailfish OS (not Apple and not Android) - sailfish browser with no JS/no cookies and 2ndary open-kimono browsers if-need-be (Webcat/Web pirate)
  - Or a 2nd phone phone (android, yuk!) with Brave or FF
- Recommend you figure out some browser-hygiene you consider ok and follow that
  - Requires some self-discipline!
  - Be willing to help others do the same!

Lab: play with “new” browser

- Don’t use one that has e.g. stored credentials for some account you care about – basically don’t muck up your daily-driver setup
- Lab machine browsers can probably be reset easily enough (TBC)
- On own laptop: install one you’ve not used before
  - Possibles: Firefox, edge, vivaldi, brave, opera...
  - More exist, but start being careful if you go beyond the above as esp. less widely used browser downloads have been known to contain malware from time to time (but mostly on phones)

Lab: watch http traffic

- Open browser
- Type shift-ctrl-I (or equivalent) to open developer interface
- Re-size screens to taste
- Choose “network” tab in developer pane
- Try loading a few sites and watch what happens
  - DO NOT load NSFW sites!
- Say which site is the “worst” from your POV and why
  - Just yell/put up hand when you have a “worst” to nominate
  - We’ll pick a winner if we’ve time – Prize == applause:-)

Shift-ctrl-I for macs...

- Macs differ:
  - Firefox: Option + Command + I
  - Safari: Option + Command + C
  - Chrome: Option + Command + C
- Access to developer tools in Safari has to be activated in the settings first. If anyone has problems with that they can find detailed information on how to do this here: https://support.apple.com/en-ie/guide/safari/sfri20948/mac
- Thanks to Luca Schäfer (2021 student) for the above

Lab: stretch goal

- Figure out how to save an HTTP archive file (.har)
- Figure out how to view .har files
- Figure out how to diff .har files
- See what changes between seemingly identical browser sessions
  - ...any tracking?

What is the Internet doing to me?
(witidtm 2021/2022 - TEU00311)

Lab Session #2

Stephen Farrell
stephen.farrell@cs.tcd.ie

https://github.com/sftcd/witidtm
https://down.dsg.cs.tcd.ie/witidtm

Today’s goals

- Find the location of an image
- Consider what facial recognition means for us

- DO NOT use any image that has a reasonable probability of upsetting anyone

Data vs. Metadata

- Typically we talk about the “data” as being the main thing being processed or communicated or stored...
  - E.g: the bits of an image or video, the content of an email, the messages in a text chat or the audio packets in a phone call
- So-called “metadata” is also data but is “about” the above rather than part of the above
  - E.g. the timing of a communication, the sender/receiver IP addresses, the size of data, etc
- Even if data is well-protected, metadata can leak separately (or be deliberately stored/exposed) so meta-data creates risk
- For someone surveilling, metadata can be more attractive than data, e.g. law enforcement may benefit more from building a social graph of criminals compared to seeing the content of a few messages, or, facebook might learn enough from whatapp metadata that they no longer need to see the content to sell advertising
- Metadata can also be a little unexpected, e.g. author information in documents, or, in images...

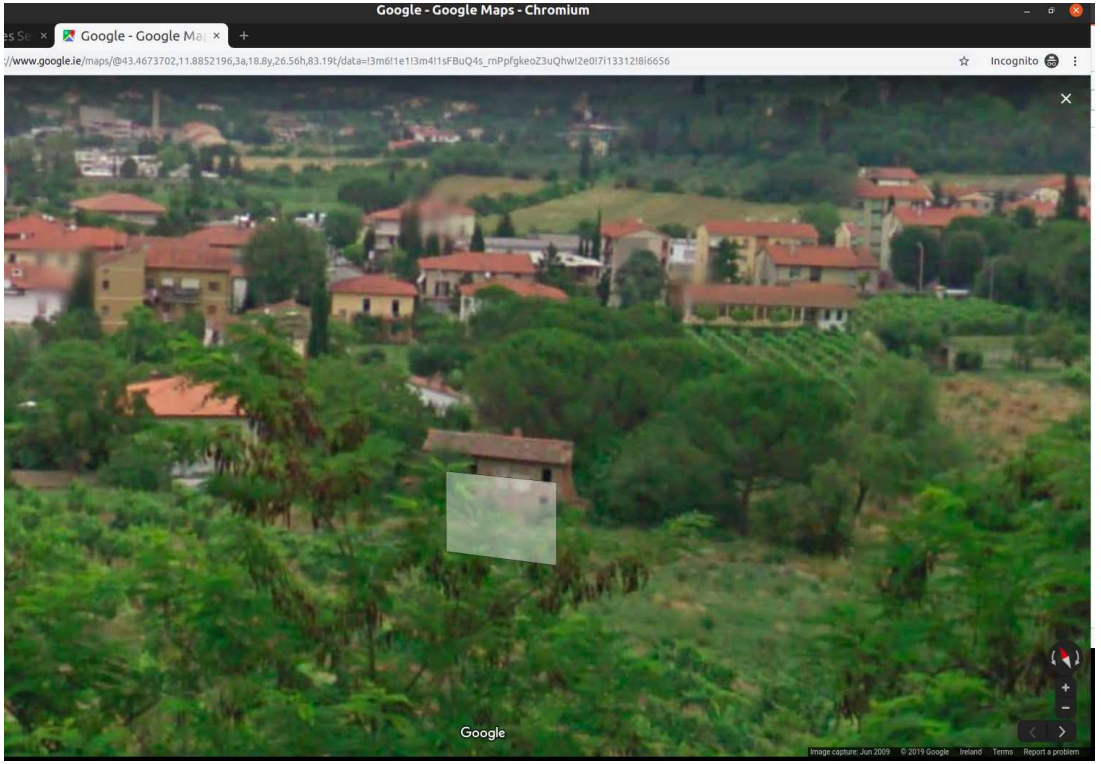
Image Metadata: EXIF

https://photographylife.com/what-is-exif-data
https://helpdeskgeek.com/how-to/how-to-remove-exif-data-from-your-photos/

```
exif DSCN0010.jpg
EXIF tags in 'DSCN0010.jpg' ('Intel' byte order):
--------------------
+-------------------------------------------------------------
Tag                |Value
--------------------
+-------------------------------------------------------------
Image Description  |
Manufacturer       |NIKON
Model              |COOLPIX P6000
Orientation        |Top-left
X-Resolution       |300
Y-Resolution       |300
Resolution Unit    |Inch
Software           |Nikon Transfer 1.1 W
Date and Time      |2008:11:01 21:15:07
YCbCr Positioning  |Centered
Compression        |JPEG compression
X-Resolution       |72
Y-Resolution       |72
Resolution Unit    |Inch
Exposure Time      |1/75 sec.
F-Number           |f/5.9
Exposure Program   |Normal program
ISO Speed Ratings  |64
Exif Version       |Exif Version 2.2
Date and Time (Origi|2008:10:22 16:28:39
Date and Time (Digit|2008:10:22 16:28:39
Components Configura|Y Cb Cr -
Exposure Bias      |0.00 EV
Maximum Aperture Val|2.90 EV (f/2.7)
Metering Mode      |Pattern
Light Source       |Unknown
Flash              |Flash did not fire, compulsory flash mode
Focal Length       |24.0 mm
Maker Note         |3298 bytes undefined data
User Comment       |
FlashPixVersion    |FlashPix Version 1.0
Color Space        |sRGB
Pixel X Dimension  |640
Pixel Y Dimension  |480
File Source        |DSC
Scene Type         |Directly photographed
Custom Rendered    |Normal process
Exposure Mode      |Auto exposure
White Balance      |Auto white balance
Digital Zoom Ratio |0.00
Focal Length in 35mm|112
Scene Capture Type |Standard
Gain Control       |Normal
Contrast           |Normal
Saturation         |Normal
Sharpness          |Normal
Subject Distance Ran|Unknown
North or South Latit|N
Latitude           |43, 28, 2.81400000
East or West Longitu|E
Longitude          |11, 53, 6.45599999
Altitude Reference |Sea level
GPS Time (Atomic Clo|14:27:07.24
GPS Satellites     |06
GPS Image Direction |
Geodetic Survey Data|WGS-84
GPS Date           |2008:10:23
Interoperability Ind|R98
Interoperability Ver|0100
-------------------+---------------------------------------------------------
EXIF data contains a thumbnail (6702 bytes).
```

https://raw.githubusercontent.com/ianare/exif-samples/master/jpg/gps/DSCN0010.jpg

52100 Arezzo, Province of Arezzo, Italy



- Took about 5 minutes to find this in Google street view
- Most of that was finding a way to map degree, minutes, seconds to fractional Lat,Long
- All **automatable**, could easily produce location history from a set of images
- How could such a “leak” be damaging to you or to someone else in your images?

https://www.google.ie/maps/@43.4673702,11.8852196,3a,75y,23.31h,83.05t/data=!3m6!1e1!3m4!1sFBuQ4s_rnPpfgkeoZ3uQhw!2e0!7i13312!8i6656

Viewing EXIF Data

- Local: Right-click and “properties”
- Better local: install something allowing you to scan multiple images
  - E.g. “sudo apt install exif” in Linux
- On web: save image to local then GOTO above

Your EXIF task...

1) Find some image(s) online or locally
2) Determine if they contain EXIF location data
3) Find the location of that image in e.g. Google street view
4) If time remains: GOTO 1

- What can you infer from the above?
- What could you infer if you did the above for a number of images of related subjects?

Facial recognition

- A kind of “biometric” (more later on the imperfections of biometrics:-)
- Nice overview, including tricky issues at: https://en.wikipedia.org/wiki/Facial_recognition_system (accessed 20210927)
- Basic idea: program analyses image bits, search for pattern that looks like a face (eyes, nose, mouth, ...), classifies that (based on machine learning using image collections), compare results from two images – if close enough, declare match
  - False positives and negatives will happen
- Note: this is not my area of expertise!

Facial recognition (ab)uses

- Find a photo of “this person”
  - Find local pics of your mum, organise your image gallery
- Find people with outstanding arrest warrants in a crowd
  - Recognition of faces in moving crowd is harder than individually, but likely, not that much harder
- Determine ethnicity of people using public transport
  - Critics may say things like the above - proponents might talk about improving efficiency but build systems that have this effect

Your facial recognition task

- GOTO https://www.kairos.com/demos
  - I’ve no opinion of that system, other than that it offers the comparison I wanted for the lab
  - Hopefully it doesn’t stop working on us (e.g. because we used it too much;-)
- Play with various image pairs, with/without the same person visible, to try understand how well/badly this particular face verification works
  - Hint: a web search for images of a well known figure (politician, musician, ...) should produce a fairly good range of images of the same subject
- If you can do such comparisons some other way, great, but do tell us about it

- What do you infer about images uploaded to web sites or “the cloud”?
- What do you infer about images **you** capture or upload?