Introduction to the Web

MPRI 2.26.2: Web Data Management

Antoine Amarilli

Friday, December 7th

The old days

1969 ARPANET (ancestor of the Internet)

1974 TCP

1990 The World Wide Web, HTTP, HTML

1994 Yahoo! was founded

1995 Amazon.com, Ebay, AltaVista are founded

1998 Google are founded

2001 Wikipedia is created

Statistics

- Around **330 million** domains, including 130 million in `.com`[1]
- **54%** of content in **English** and **4%** in French[2]
- **48%** of people have Internet access, and **71%** of ages 15–24[3]
- Google knows over one **trillion** (10^{12}) of unique URLs[4]
  - → The **same content** can live in many different URLs
  - → Parts of the Web are not indexable: the **hidden Web** or **deep Web**

[1] `https://www.businesswire.com/news/home/20170914006386/en/Internet-Grows-331.9-Million-Domain-Registrations-Quarter`

[2] `https://w3techs.com/technologies/overview/content_language/all`

[3] `https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2017.pdf`

[4] `https://googleblog.blogspot.fr/2008/07/we-knew-web-was-big.html`

Table of Contents

Historical web browsers

Mosaic First common graphical browser, 1993–1997

- From 80% in 1994 to $<$ 10% in 1996

Historical web browsers

Mosaic First common graphical browser, 1993–1997

- From 80% in 1994 to $<$ 10% in 1996

Netscape Released in 1994, based on Mosaic

- From 80% in 1996 to $<$ 10% in 2001

Historical web browsers

Mosaic First common graphical browser, 1993–1997

- From 80% in 1994 to $<$ 10% in 1996

Netscape Released in 1994, based on Mosaic

- From 80% in 1996 to $<$ 10% in 2001

Internet Explorer Released in 1995

- Provided with Windows 95
- IE 6 released in 2001 and reaches 80% market share
- Leads to an antitrust lawsuit in the USA, 1998–2001

Historical web browsers

Mosaic First common graphical browser, 1993–1997

- From 80% in 1994 to $<$ 10% in 1996

Netscape Released in 1994, based on Mosaic

- From 80% in 1996 to $<$ 10% in 2001

Internet Explorer Released in 1995

- Provided with Windows 95
- IE 6 released in 2001 and reaches 80% market share
- Leads to an antitrust lawsuit in the USA, 1998–2001

Firefox Released in 2002 from Netscape (open-sourced in 1998)

- Free and open-source
- Popularized tabbed browsing
- Attacked IE 6's monopoly

Current Web browsers

IE IE 7 released in 2006, replaced by Microsoft Edge

Firefox Still actively developed

Safari Released in 2003, default Web browser on Mac OS X

Opera Released in 1996, initially commercial (until 2000) then ad-supported (until 2005), now free but proprietary.

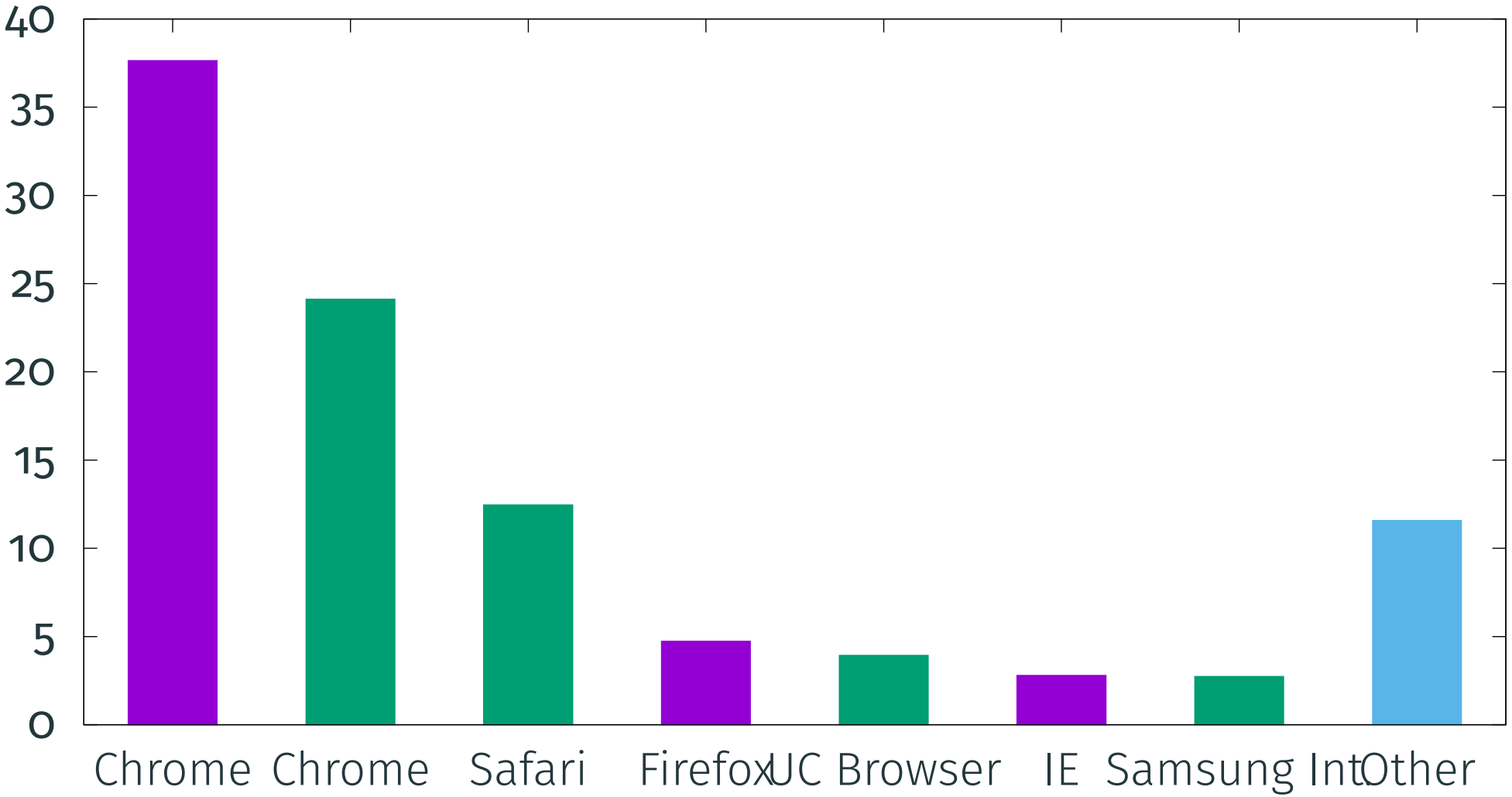
Chrome Released in 2008 by Google, with an open-source version (Chromium)

Mobile 55% market share according to `gs.statcounter.com`

- Safari (iOS), Android browser or Chrome (Android)
- Firefox Mobile, Blackberry, Opera, UC Browser...

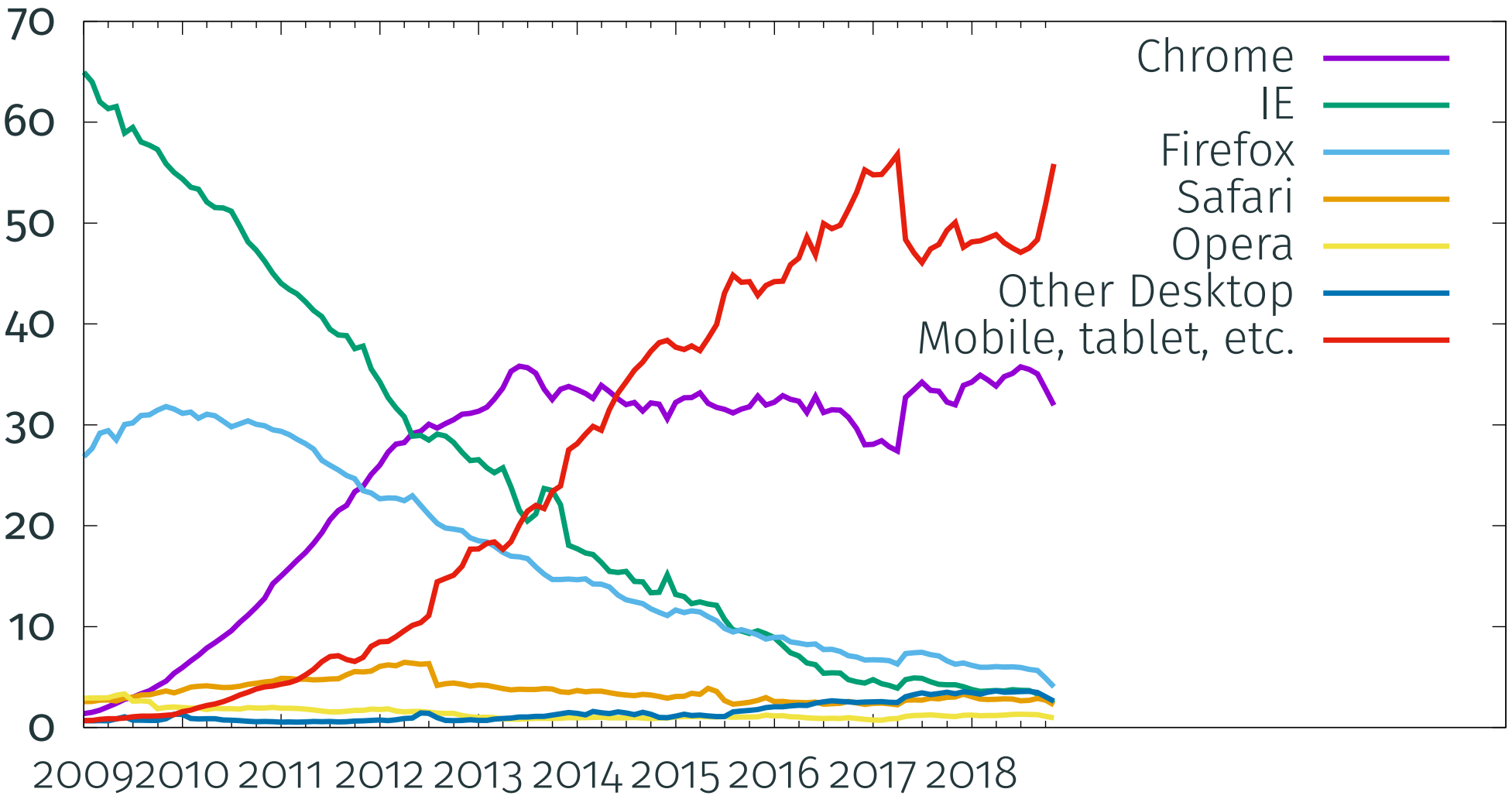
To check rendering on old browsers, use `browsershots.org`

Recent market share



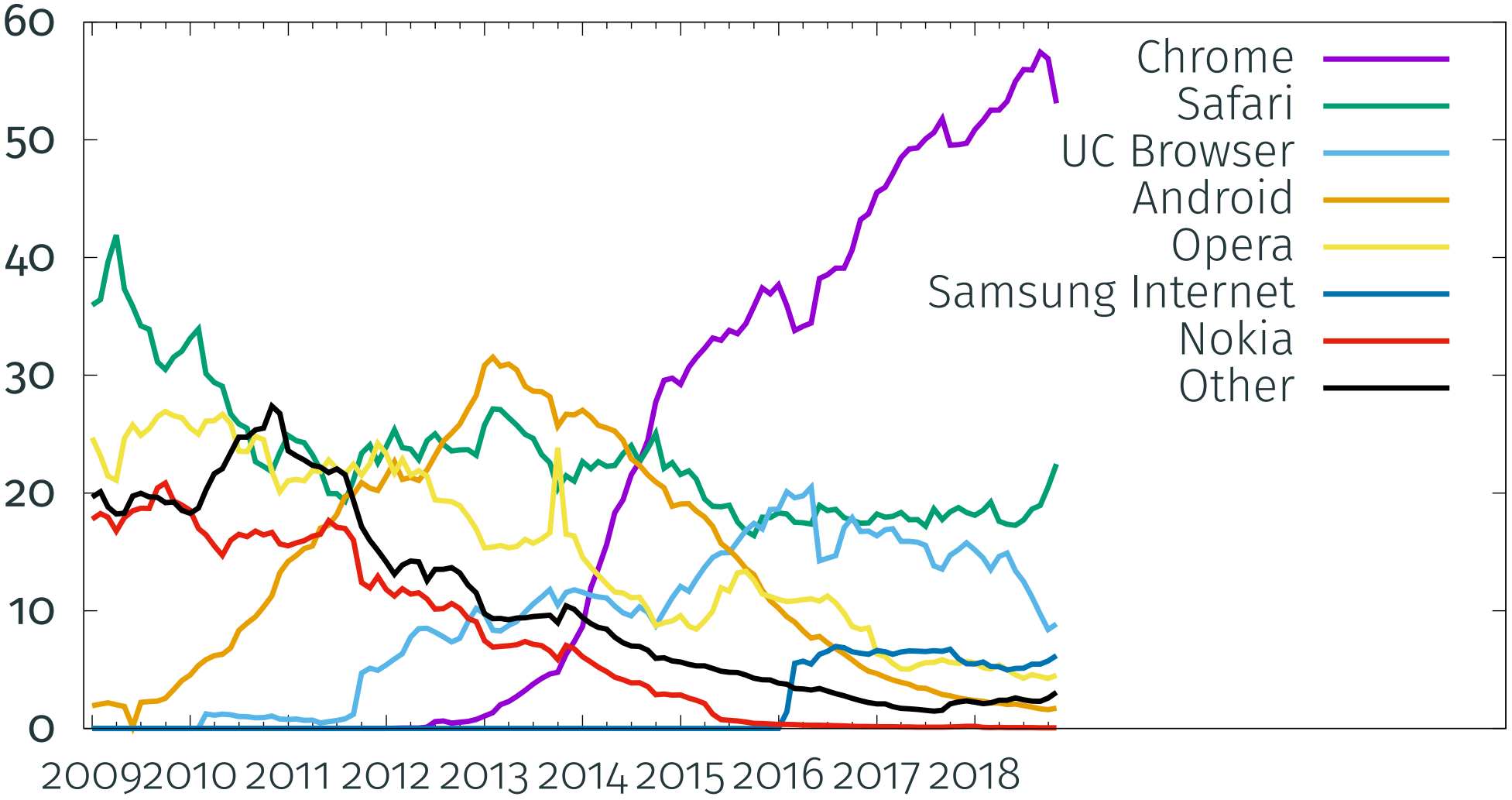
Source: `gs.statcounter.com` (November 2018)

Evolution



Source: `gs.statcounter.com`

Evolution (mobile)



Source: `gs.statcounter.com`

Rendering engine

Firefox Gecko, and (work-in-progress) Servo, using Rust

Safari WebKit engine

Chrome Blink (fork of Webkit, in April 2013)

IE Originally Trident, then EdgeHTML, soon Chromium[5]

Opera Originally Presto, then Blink

Others Dillo, KHTML, and other old/minimalistic engines

[5] `https://www.windowscentral.com/microsoft-building-chromium-powered-web-browser-windows-10`, Dec 3, 2018

Summary and perspectives

- Webkit/Blink and Chrome/Chromium are dominant
- Only serious challenger: Firefox (but around 5% globally)
- Blink is open-source but controlled by Google
- Different browsers using the same rendering engine
  - Some minimalistic, e.g., `uzbl`
  - Some variants of existing browsers, e.g., Pale Moon, Basilisk (Firefox forks)
  - Some new browsers using Blink: Vivaldi, Brave (use Blink)

New Web browser themes

- Ecosystem of **extensions**: Chrome Web Store, Mozilla Store
  - Firefox: Mandatory **signing**, and **backward-incompatible** changes

[6] `https://www.statista.com/topics/3201/ad-blocking/`

New Web browser themes

- Ecosystem of **extensions**: Chrome Web Store, Mozilla Store
  - Firefox: Mandatory **signing**, and **backward-incompatible** changes
- **Ad blocking** (**27%** of users[6]), counter*-measures
  - **uBlock Origin** extension, based on **Easylist**

    `https://easylist.to/easylist/easylist.txt`
  - More generally, **JavaScript blockers**, e.g., uMatrix, NoScript, etc.

[6] `https://www.statista.com/topics/3201/ad-blocking/`

New Web browser themes

- Ecosystem of **extensions**: Chrome Web Store, Mozilla Store
  - Firefox: Mandatory **signing**, and **backward-incompatible** changes
- **Ad blocking** (**27%** of users[6]), counter*-measures
  - **uBlock Origin** extension, based on **Easylist**

    `https://easylist.to/easylist/easylist.txt`
  - More generally, **JavaScript blockers**, e.g., uMatrix, NoScript, etc.
- Filtering out **bots**: robots exclusion standard, CAPTCHAs
  - **reCAPTCHA**: now volunteer work for Google

[6] `https://www.statista.com/topics/3201/ad-blocking/`

New Web browser themes

- Ecosystem of **extensions**: Chrome Web Store, Mozilla Store
  - Firefox: Mandatory **signing**, and **backward-incompatible** changes
- **Ad blocking** (**27%** of users[6]), counter*-measures
  - **uBlock Origin** extension, based on **Easylist**

    `https://easylist.to/easylist/easylist.txt`
  - More generally, **JavaScript blockers**, e.g., uMatrix, NoScript, etc.
- Filtering out **bots**: robots exclusion standard, CAPTCHAs
  - **reCAPTCHA**: now volunteer work for Google
- **Security**: site isolation, one process/site (in Chromium)

[6] `https://www.statista.com/topics/3201/ad-blocking/`

New Web browser themes

- Ecosystem of **extensions**: Chrome Web Store, Mozilla Store
  - Firefox: Mandatory **signing**, and **backward-incompatible** changes
- **Ad blocking** (**27%** of users[6]), counter*-measures
  - **uBlock Origin** extension, based on **Easylist**

    `https://easylist.to/easylist/easylist.txt`
  - More generally, **JavaScript blockers**, e.g., uMatrix, NoScript, etc.
- Filtering out **bots**: robots exclusion standard, CAPTCHAs
  - **reCAPTCHA**: now volunteer work for Google
- **Security**: site isolation, one process/site (in Chromium)
- Web browser **fingerprinting** `https://panopticlick.eff.org/`

[6] `https://www.statista.com/topics/3201/ad-blocking/`

New Web browser themes

- Ecosystem of **extensions**: Chrome Web Store, Mozilla Store
  - Firefox: Mandatory **signing**, and **backward-incompatible** changes
- **Ad blocking** (**27%** of users[6]), counter*-measures
  - **uBlock Origin** extension, based on **Easylist**
    `https://easylist.to/easylist/easylist.txt`
  - More generally, **JavaScript blockers**, e.g., uMatrix, NoScript, etc.
- Filtering out **bots**: robots exclusion standard, CAPTCHAs
  - **reCAPTCHA**: now volunteer work for Google
- **Security**: site isolation, one process/site (in Chromium)
- Web browser **fingerprinting** `https://panopticlick.eff.org/`
- In-browser cryptocurrency mining: **CoinHive**
  - Test: 30 hash/s, difficulty 47G, block reward 3.5 XMR = 185.5 USD

  → **0.01 USD/day**, vs about **0.10 USD/day** of electricity (for 50W)

[6] `https://www.statista.com/topics/3201/ad-blocking/`

New Web browser themes

- Ecosystem of **extensions**: Chrome Web Store, Mozilla Store
  - Firefox: Mandatory **signing**, and **backward-incompatible** changes
- **Ad blocking** (**27%** of users[6]), counter*-measures
  - **uBlock Origin** extension, based on **Easylist**

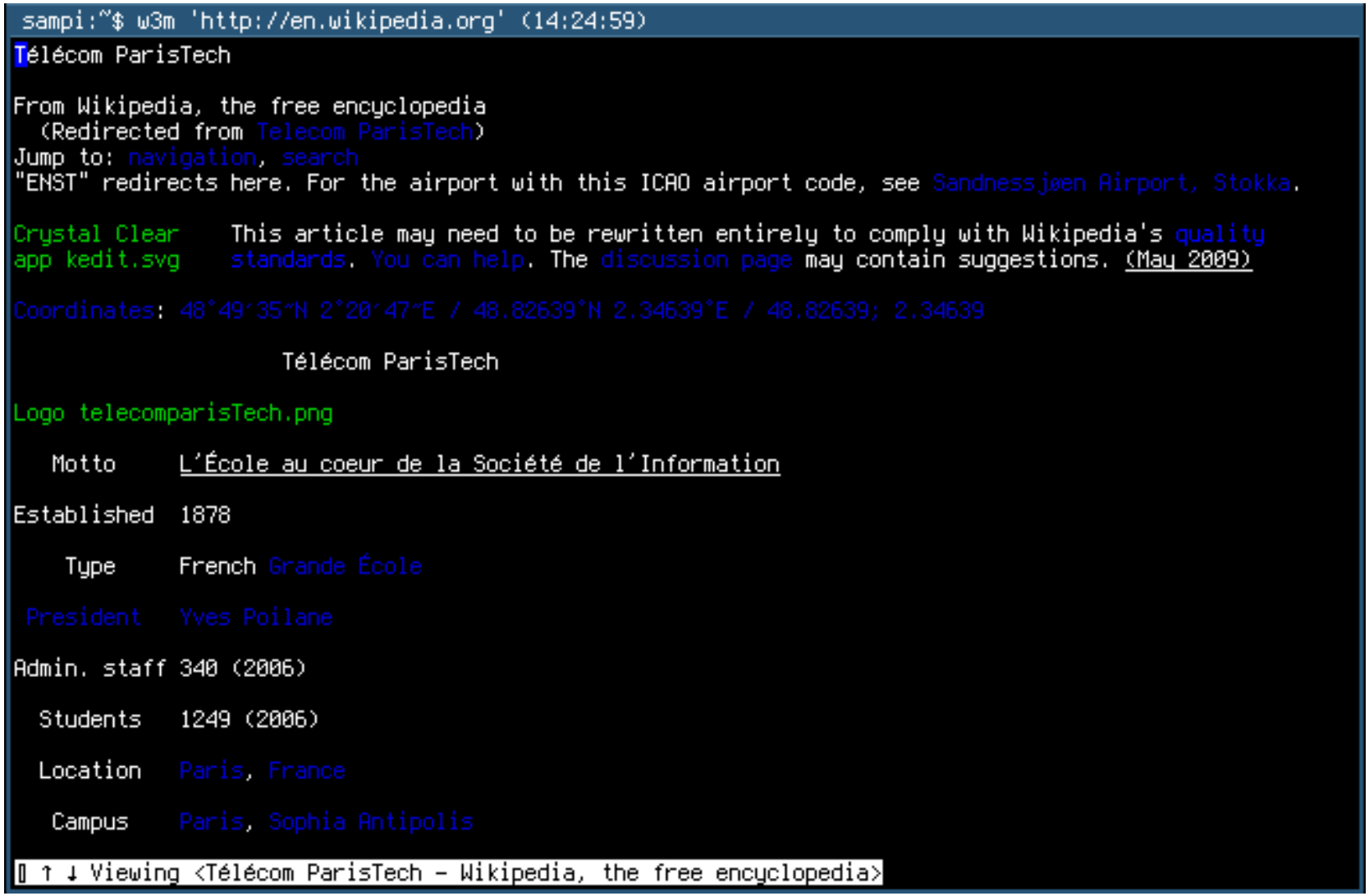
    `https://easylist.to/easylist/easylist.txt`
  - More generally, **JavaScript blockers**, e.g., uMatrix, NoScript, etc.
- Filtering out **bots**: robots exclusion standard, CAPTCHAs
  - **reCAPTCHA**: now volunteer work for Google
- **Security**: site isolation, one process/site (in Chromium)
- Web browser **fingerprinting** `https://panopticlick.eff.org/`
- In-browser cryptocurrency mining: **CoinHive**
  - Test: 30 hash/s, difficulty 47G, block reward 3.5 XMR = 185.5 USD

  → **0.01 USD/day**, vs about **0.10 USD/day** of electricity (for 50W)
- **Tor** and **Tor hidden services**

[6] `https://www.statista.com/topics/3201/ad-blocking/`

Table of Contents

Textual Web browsers



- lynx (still maintained), w3m, elinks

Robots

Many automated programs on the Web:

- Search engine crawlers: see Pierre's class
- RSS readers and aggregators
- Email harvesters (spammers)
- API consumers

Table of Contents

URL

- Anatomy of a URL:

`https://en.wikipedia.org/wiki/Telecom_ParisTech#History`

protocol machine path fragment

- Uniform Resource Locator
- Identifies a **resource** on the Web

Credits

- Course material inspired by course notes by Pierre Senellart