Introduction to the Web

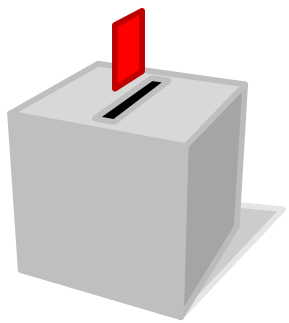
MPRI 2.26.2: Web Data Management

Antoine Amarilli

POLL: World Wide Web creation

The Web was invented...

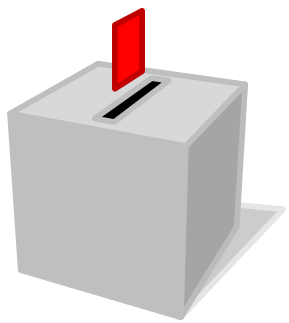
- **A**: About at the same as the Internet
- **B**: 2 years after the Internet
- **C**: 5 years after the Internet
- **D**: More than 5 years after the Internet



POLL: World Wide Web creation

The Web was invented...

- **A**: About at the same as the Internet
- **B**: 2 years after the Internet
- **C**: 5 years after the Internet
- **D**: **More than 5 years after the Internet**



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

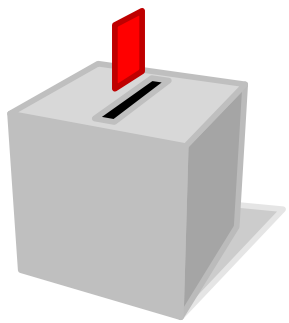
1998 Google are founded

2001 Wikipedia is created

POLL: Number of Internet domains

How many domain names exist on the Internet?

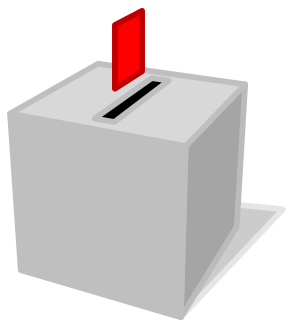
- **A**: 3 million
- **B**: 30 million
- **C**: 300 million
- **D**: 3 billion



POLL: Number of Internet domains

How many domain names exist on the Internet?

- **A**: 3 million
- **B**: 30 million
- **C**: **300 million**
- **D**: 3 billion



Statistics

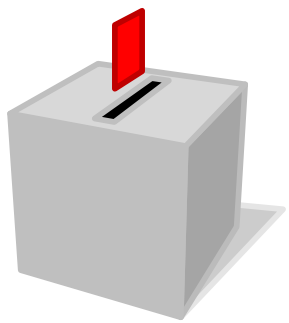
- Around **370 million** domains, including 150 million in `.com`[1]
- **64%** of content in **English** and **3%** in French[2]
- Google knows over one **trillion** (10^{12}) of unique URLs[3] and possibly hundreds of trillions
  - → The **same content** can live in many different URLs
  - → Parts of the Web are not indexable: the **hidden Web** or **deep Web**

[1] `https://www.verisign.com/en_US/domain-names/dnib/index.xhtml`
[2] `https://w3techs.com/technologies/overview/content_language/all`
[3] `https://googleblog.blogspot.fr/2008/07/we-knew-web-was-big.html`

POLL: Internet users

Which proportion of the world population is using the Internet

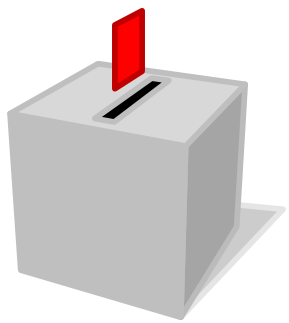
- **A**: Less than 1/3
- **B**: Between 1/3 and 1/2
- **C**: Between 1/2 and 2/3
- **D**: More than 2/3



POLL: Internet users

Which proportion of the world population is using the Internet

- **A**: Less than 1/3
- **B**: Between 1/3 and 1/2
- **C**: **Between 1/2 and 2/3**
- **D**: More than 2/3



Users

- 63% of the world population uses the Internet
  - Gender imbalance: 57% of women and 62% of men
  - Age imbalance: 71% of people with ages 15–24
- The connectivity exists, however:
  - 95% of the world population have access to a mobile network
  - 88% have access to 4G

Source : ITU `https://www.itu.int/en/ITU-D/Statistics/Documents/facts/FactsFigures2021.pdf`

Table of Contents

Introduction

Web browsers

Other Web clients

Historical web browsers

Mosaic First common graphical browser, 1993–1997

Netscape Released in 1994, based on Mosaic

Internet Explorer Released in 1995, provided with Windows 95

- IE 6 released in 2001 and reaches 80% market share

Firefox Released in 2002 from Netscape

- Attacked IE 6's monopoly

Current Web browsers (desktop)

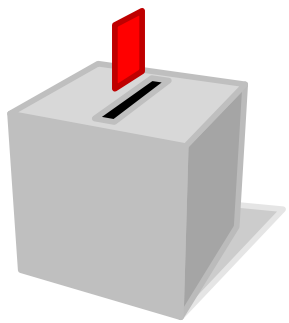
IE	IE 7 released in 2006, replaced by Microsoft Edge
Firefox	Still actively developed
Safari	Released in 2003, default Web browser on Mac OS X
Opera	Released in 1996, proprietary (niche)
Chrome	Released in 2008 by Google, with an open-source version (Chromium)

To check rendering on old browsers, use `browserstack.com`

POLL: Web Browser Market share (1/3)

Which is the most common Web browser nowadays?

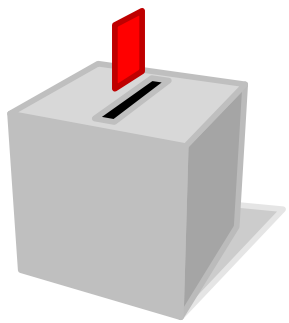
- **A**: Internet Explorer / Edge
- **B**: Mozilla Firefox
- **C**: Google Chrome
- **D**: Apple Safari



POLL: Web Browser Market share (1/3)

Which is the most common Web browser nowadays?

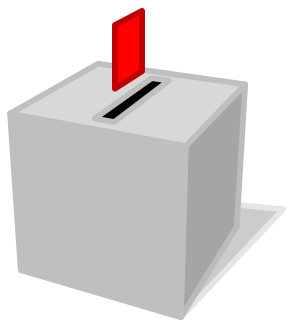
- **A**: Internet Explorer / Edge
- **B**: Mozilla Firefox
- **C**: **Google Chrome**
- **D**: Apple Safari



POLL: Web Browser Market share (2/3)

What is its main competitor?

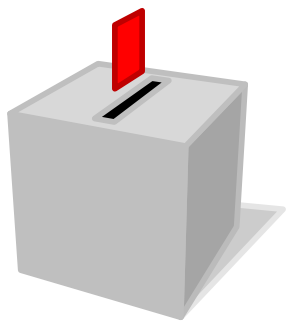
- **A**: Internet Explorer / Edge
- **B**: Mozilla Firefox
- **C**: Apple Safari
- **D**: A more obscure browser?



POLL: Web Browser Market share (2/3)

What is its main competitor?

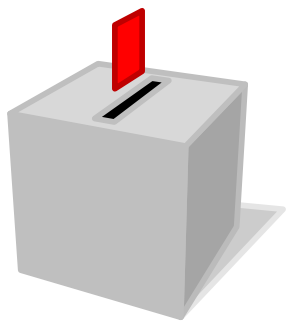
- **A**: Internet Explorer / Edge
- **B**: Mozilla Firefox
- **C**: **Apple Safari**
- **D**: A more obscure browser?



POLL: Web Browser Market share (3/3)

What is the market share of the main challenger (Safari)?

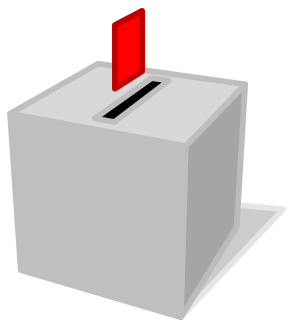
- **A**: 10%
- **B**: 20%
- **C**: 30%
- **D**: 40%



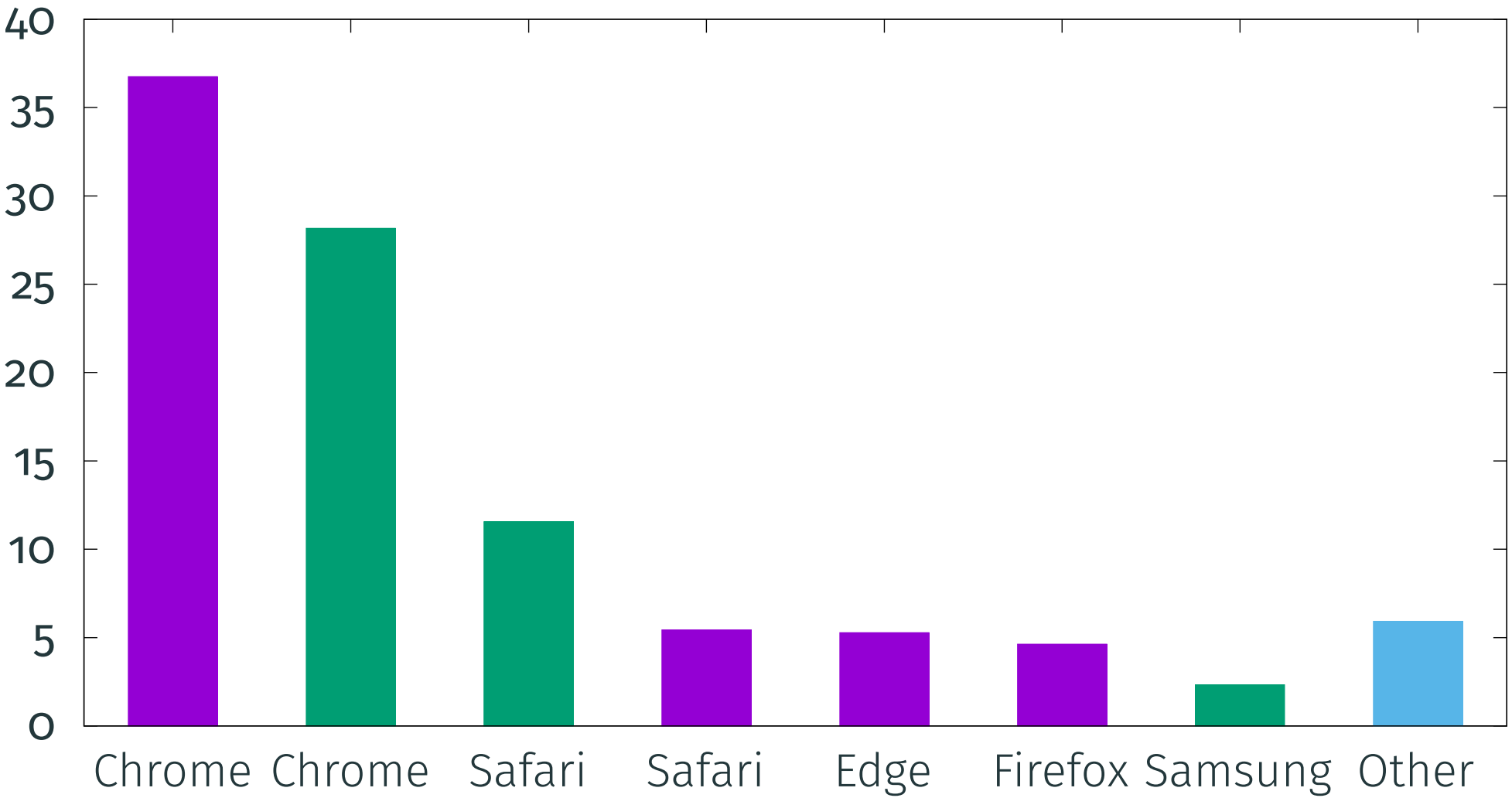
POLL: Web Browser Market share (3/3)

What is the market share of the main challenger (Safari)?

- **A**: 10%
- **B**: **20%**
- **C**: 30%
- **D**: 40%

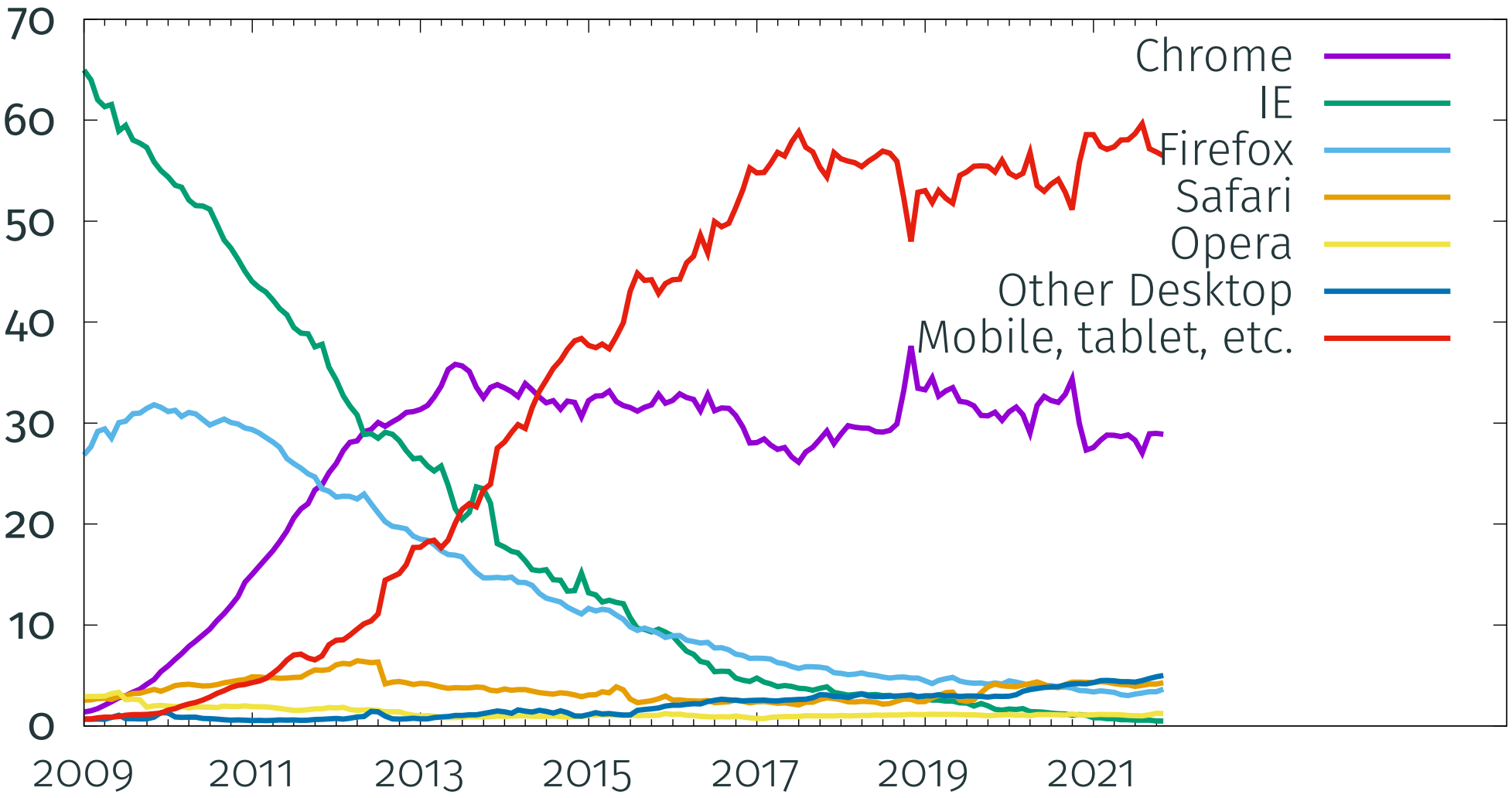


Recent market share



Source: `gs.statcounter.com` (November 2021)

Evolution

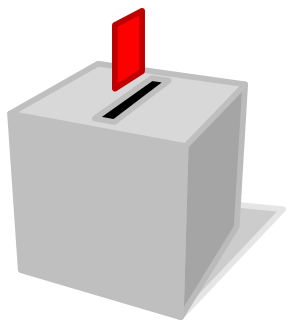


Source: `gs.statcounter.com`

POLL: Mobile traffic

Which proportion of Web traffic is on mobile?

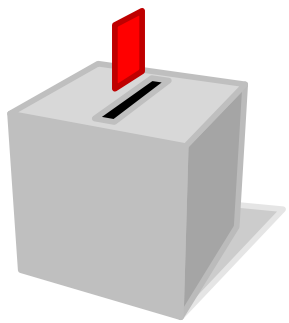
- **A**: Less than 1/3
- **B**: A bit less than 1/2
- **C**: A bit more than 1/2
- **D**: More than 2/3



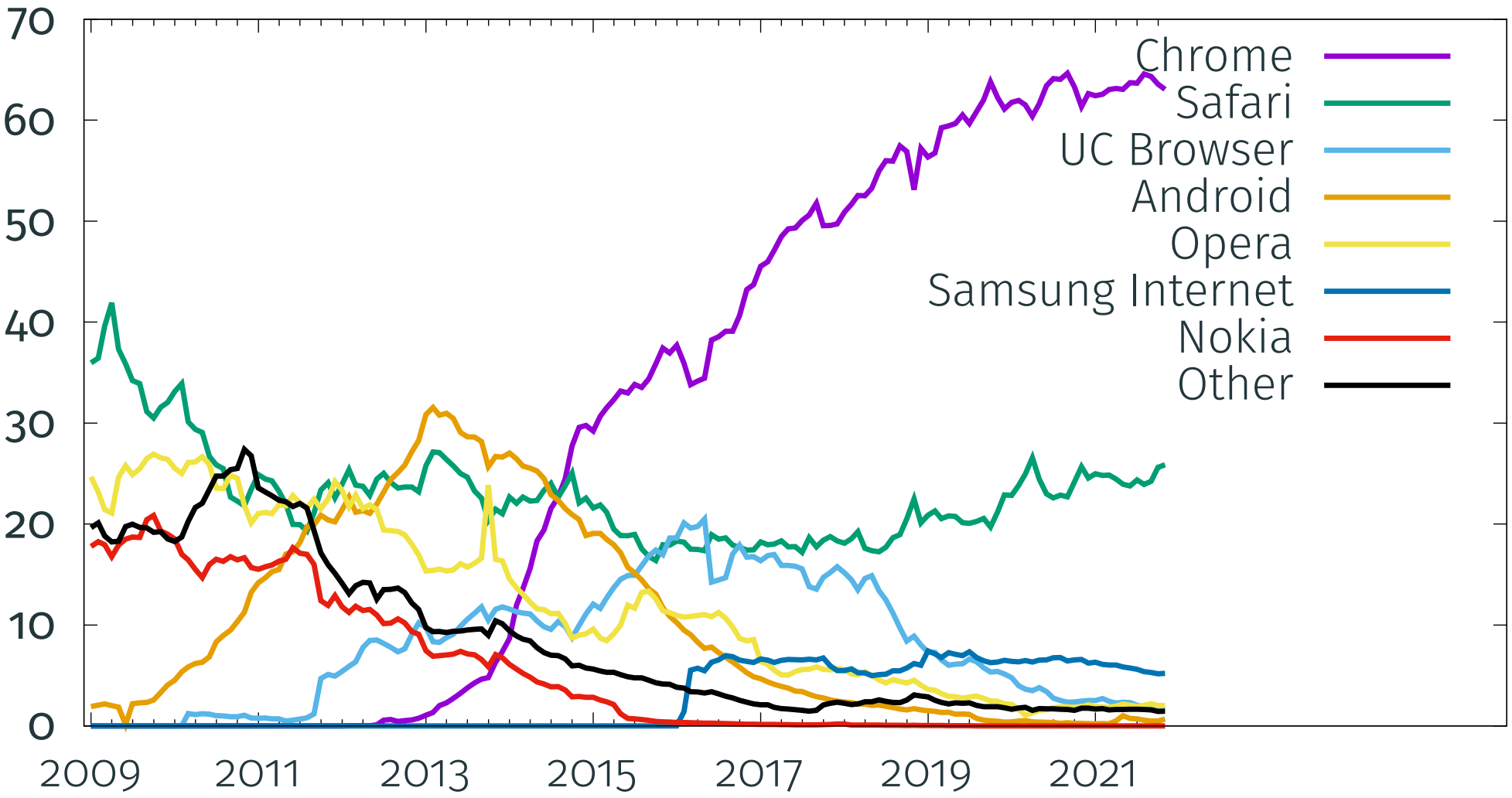
POLL: Mobile traffic

Which proportion of Web traffic is on mobile?

- **A**: Less than 1/3
- **B**: A bit less than 1/2
- **C**: **A bit more than 1/2**
- **D**: More than 2/3



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, Chromium since January 2020

Opera Originally Presto, then Blink

Others Dillo, KHTML, and other old/minimalistic engines

Summary and perspectives

- Webkit/Blink and Chrome/Chromium are dominant
- Main contenders: Safari (especially on mobile) and Firefox
- Blink is open-source but controlled by Google
- Different browsers using this rendering engine
  - Some minimalistic, e.g., `uzbl`
  - Some new browsers using Blink: Vivaldi, Brave

New topics about Web browsers (1)

- Ecosystem of extensions: Chrome Web Store, Mozilla Store, signing

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

New topics about Web browsers (1)

- Ecosystem of **extensions**: Chrome Web Store, Mozilla Store, signing
- **Ad blocking** (in the US, **47%** of desktop and **29%** of mobile users[4]), counter*-measures
  - **uBlock Origin** extension, based on **Easylist**
    `https://easylist.to/easylist/easylist.txt`
  - More generally, **JavaScript blockers**, e.g., uMatrix, NoScript, etc.

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

New topics about Web browsers (2)

- Filtering out bots: robots exclusion standard, CAPTCHAs
  - reCAPTCHA: now volunteer work for Google

New topics about Web browsers (2)

- Filtering out bots: robots exclusion standard, CAPTCHAs
  - reCAPTCHA: now volunteer work for Google
- Security: site isolation, one process/site
  - Available in Chrome, experimental in Firefox (Project Fission)

New topics about Web browsers (2)

- Filtering out bots: robots exclusion standard, CAPTCHAs
  - reCAPTCHA: now volunteer work for Google
- Security: site isolation, one process/site
  - Available in Chrome, experimental in Firefox (Project Fission)
- Web browser fingerprinting `https://panopticlick.eff.org/`

New topics about Web browsers (2)

- Filtering out bots: robots exclusion standard, CAPTCHAs
  - reCAPTCHA: now volunteer work for Google
- Security: site isolation, one process/site
  - Available in Chrome, experimental in Firefox (Project Fission)
- Web browser fingerprinting `https://panopticlick.eff.org/`
- In-browser cryptocurrency mining (cryptojacking)

New topics about Web browsers (2)

- Filtering out bots: robots exclusion standard, CAPTCHAs
  - reCAPTCHA: now volunteer work for Google
- Security: site isolation, one process/site
  - Available in Chrome, experimental in Firefox (Project Fission)
- Web browser fingerprinting `https://panopticlick.eff.org/`
- In-browser cryptocurrency mining (cryptojacking)
- Tor and Tor hidden services

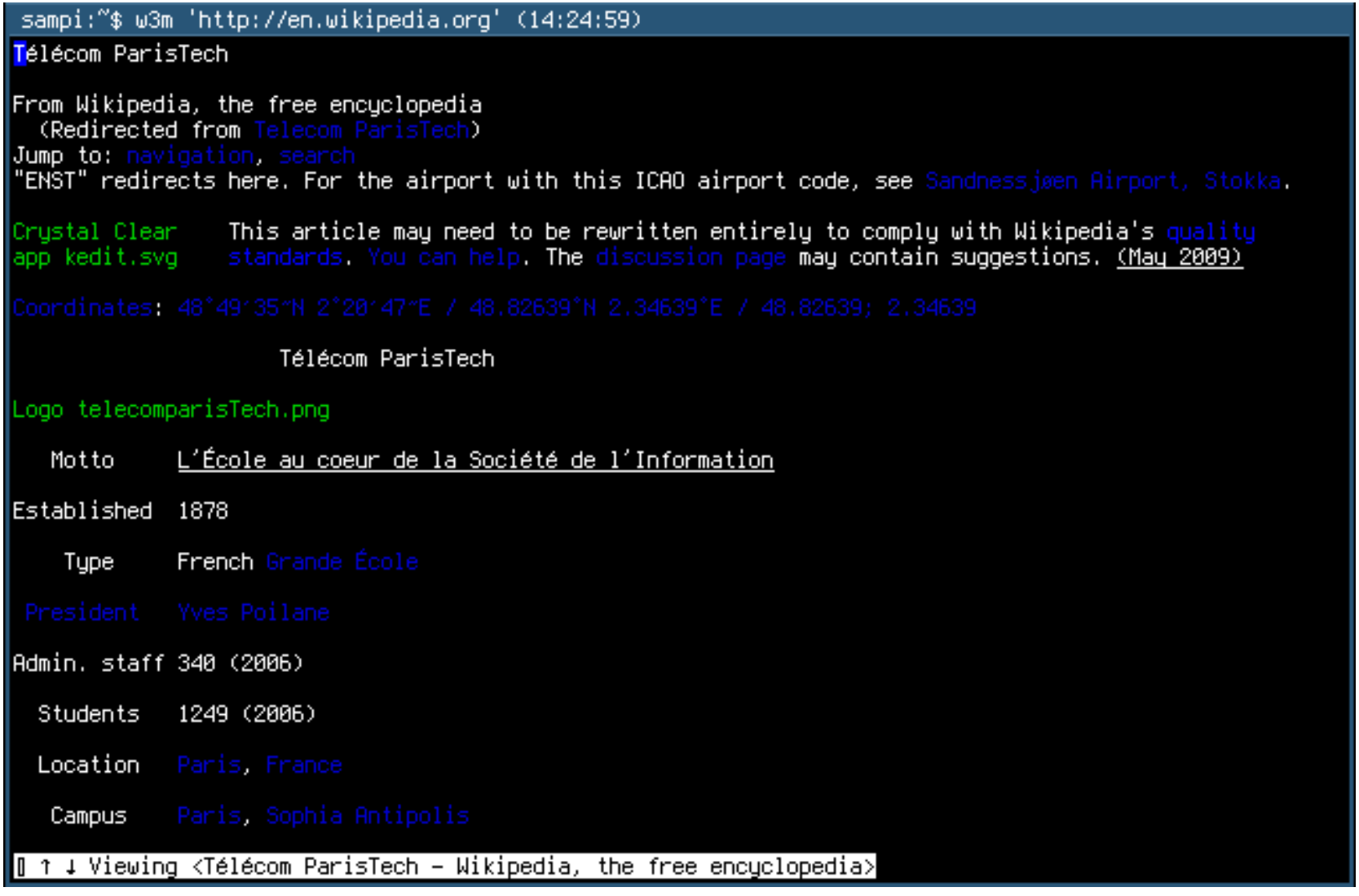
Table of Contents

Introduction

Web browsers

Other Web clients

Textual Web browsers



- lynx (still maintained), w3m, elinks
- Also: screen readers for visually impaired users

Robots

Many automated programs on the Web:

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

Credits

- Course material inspired by course notes by Pierre Senellart