Web and Social Media Analytics

Web Analytics (Knowledge Gathering Part I)

Dr Philip Worrall

School of Computer Science and Engineering
115 New Cavendish Street
University of Westminster
London, W1W 6UW
worralph@westminster.ac.uk

LW1

Outline

- Module organisation
  - Assessment
  - Tutorial arrangements
- Introduction
- Success on the web
  - Web Metrics
- Principles of the web
  - Web Servers and Data Collection
  - Page Tagging
  - HTTP Protocol and Cookies
- Knowledge gathering

Introduction

- In this section of the module, we are going to attempt to understand what factors contribute towards a website's *success*.
- I'm going to suggest that once we are aware of such factors we can implement strategies to make a website more successful
- But, first lets consider success on the web

Success

- What do you believe characterises a successful website?
  - Think about websites you visit regularly
  - Why do you use them over others?

- What characterises an unsuccessful one?
  - Why might you decide to switch to using an alternative?

A successful website?

Metric	Value
No. of Visitors	HIGH
Time spent on the website	HIGH
No. people who buy something?	HIGH
No of people who return?	HIGH
Level of interactivity	HIGH
Level of trust	HIGH
Accuracy of information	HIGH
Amount of information	MEDIUM/HIGH
Level of advertisements	LOW

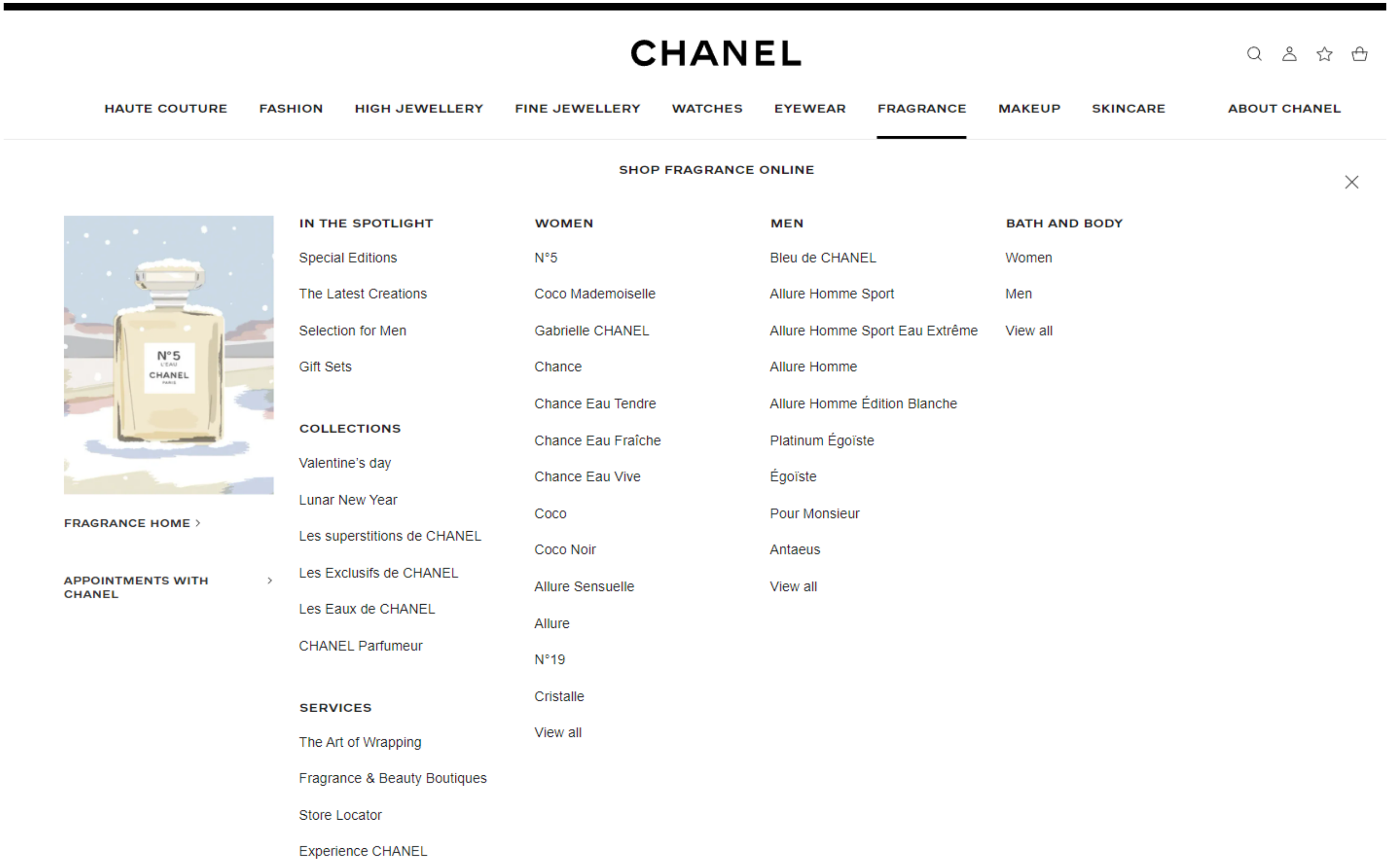
Measuring success

- Success can be measured in several ways
  - Nature of the website
  - E-retailer, charity vs. professional photographer

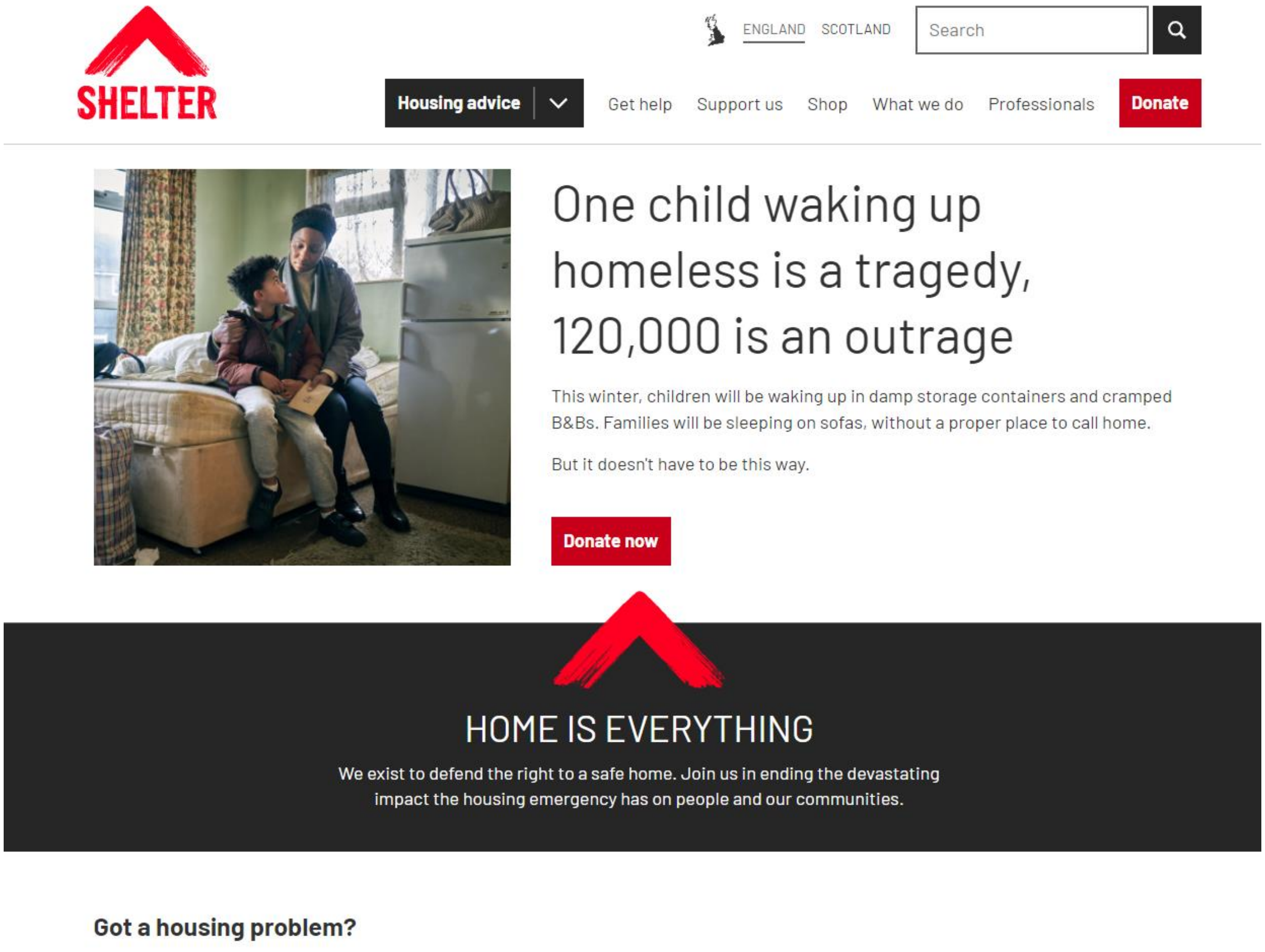
- We refer to these measures of success as web metrics

- In fact, depending on the website we may use a completely different set of metrics to measure success.
  - Even if it uses the same metrics, the weightings a website gives to a particular set of metrics may differ

How to measure success?



How to measure success?



One child waking up homeless is a tragedy, 120,000 is an outrage

This winter, children will be waking up in damp storage containers and cramped B&Bs. Families will be sleeping on sofas, without a proper place to call home.

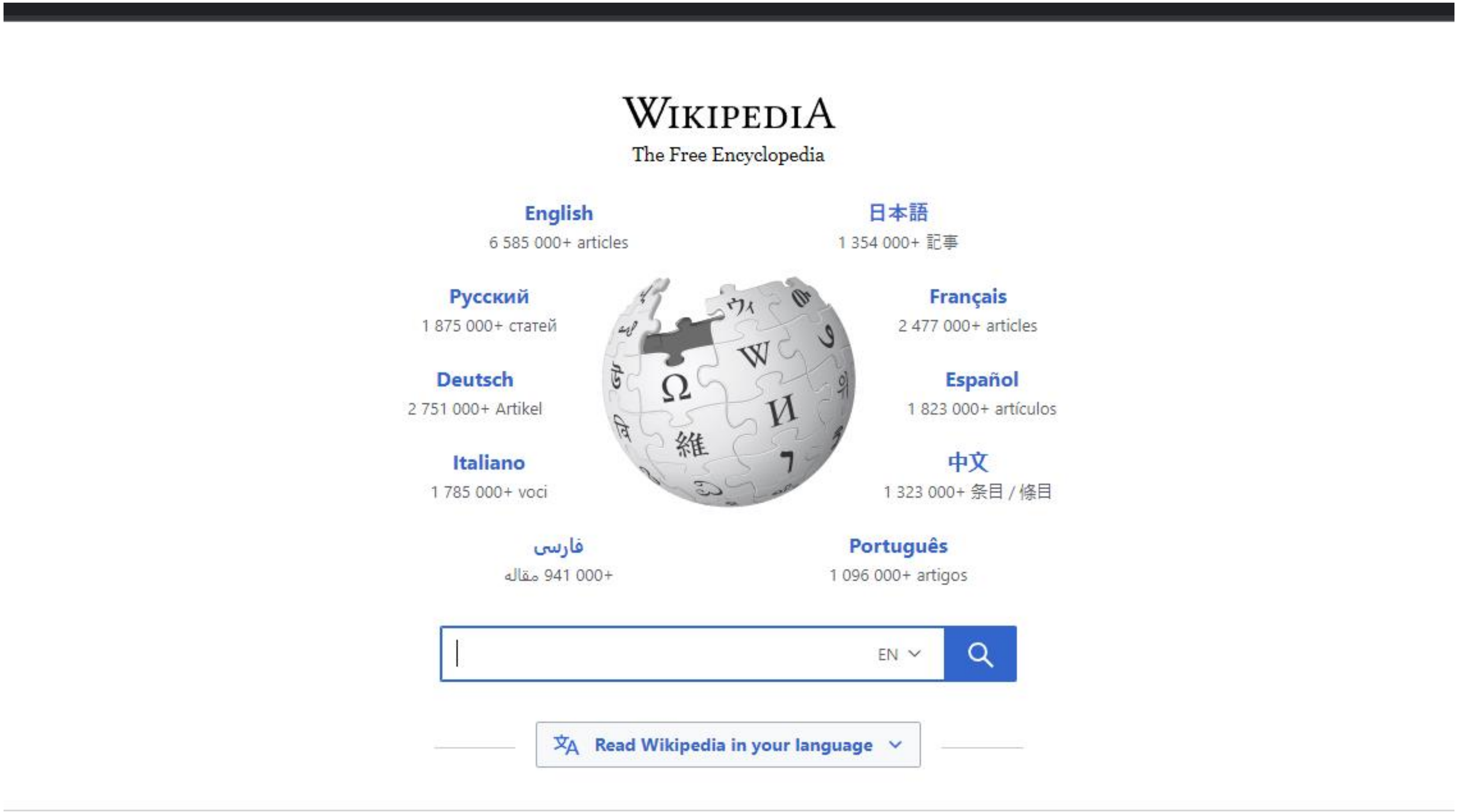
But it doesn't have to be this way.

HOME IS EVERYTHING

We exist to defend the right to a safe home. Join us in ending the devastating impact the housing emergency has on people and our communities.

Got a housing problem?

How to measure success?



WIKIPEDIA

The Free Encyclopedia

English
6 585 000+ articles

日本語
1 354 000+ 記事

Русский
1 875 000+ статей

Français
2 477 000+ articles

Deutsch
2 751 000+ Artikel

Español
1 823 000+ artículos

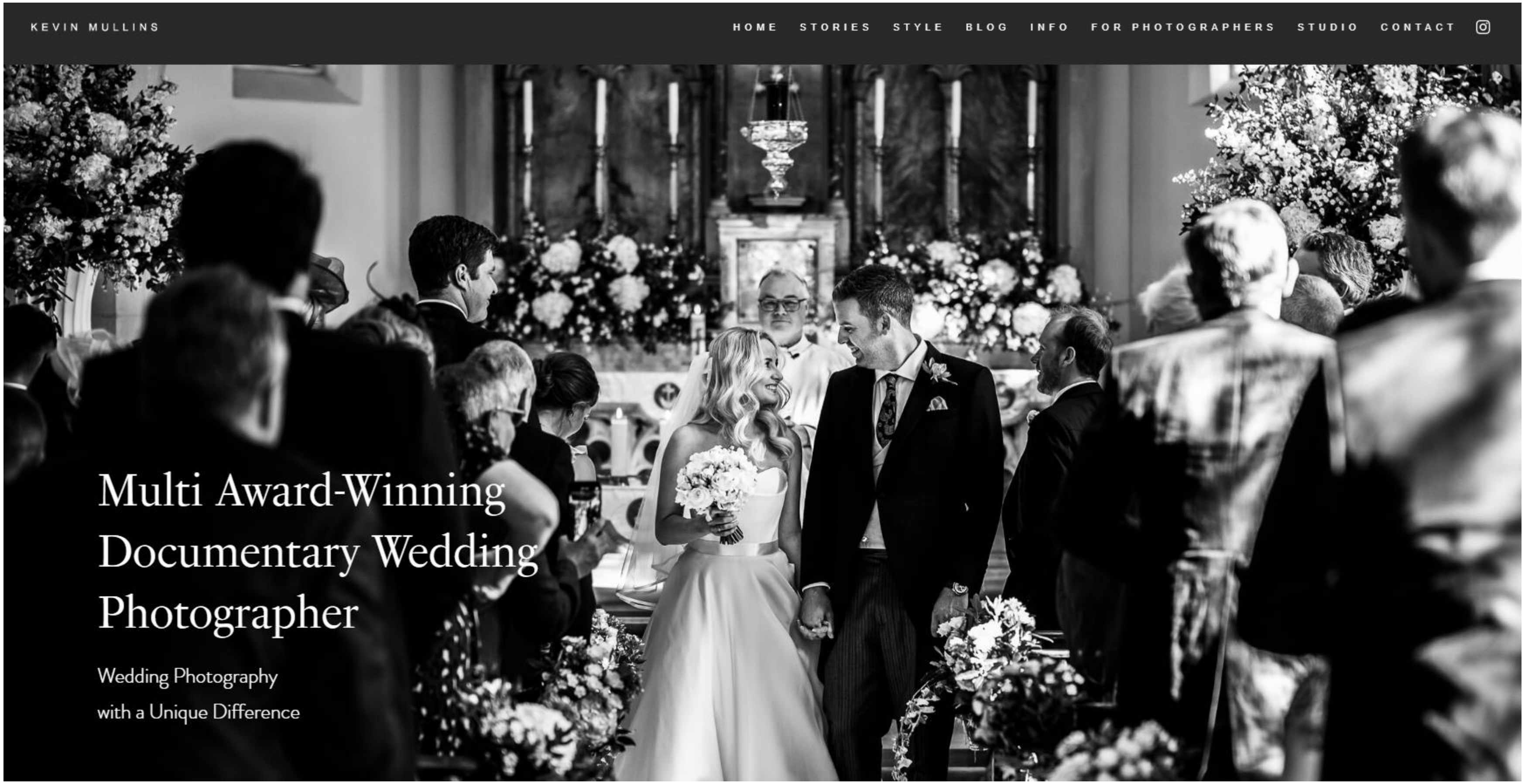
Italiano
1 785 000+ voci

中文
1 323 000+ 条目 / 條目

فارسی
941 000+ مقاله

Português
1 096 000+ artigos

How to measure success?



Multi Award-Winning Documentary Wedding Photographer

Wedding Photography
with a Unique Difference

Measuring success

- Whilst some metrics are fairly easy to measure, others can be far more problematic.
  - Wikipedia
    - May care more about how many people cite its articles.
    - How many people publish /edit an article
  - Cancer Research
    - May measure success by how many people go to get screened after checking their symptoms
  - BMW
    - May be more interested in the number of people who call to arrange a test drive.

- How might information about this be collected?

Principles of the web

- Web analytics can therefore be thought of a process to quantify the effectiveness of sites on the web.
  - Identifying relevant metrics to measure impact and reach
  - Collecting data on user engagement and activity
  - Implementing changes based on feedback and analysis

- We need to know something about how the web works at a technical level to appreciate the opportunities for data collection.

Phase 1: the internet (1/3)

- Collection of “freakish” ideas in 1960’s
- Prof. Joseph Carl Robnett Licklider proposed the idea of a “galactic computer network” whilst working at MIT
- Computers work together to solve problems
- In October 1962 he became the head of computer research for ARPA (now know as DARPA **Defense Advanced Research Projects Agency**)
- He continued to work with researchers at MIT and one in particular ,Lawrence G. Roberts, setup the first connection between two distant computers over a normal telephone line
- Between Massachusetts and California

Phase 1: the internet (2/3)

- Roberts eventually went on to work for ARPA
- Planned to solve big problems by distributing the various pieces of a problem among a large group of connected computers
  - tender put out to build the network (ARPANET)
- Universities connected
  - UCLA, UC Santa Barbara, University of Utah, Stanford
  - 1971, first email message sent over ARPANET
  - 1973, file transfer protocol (FTP)

Phase 1: the internet (3/3)

- Up until now ARPANET was mainly used to solve research and non-commercial problems
  - Ability to send and receive messages across the network

- 1979 Norway and the UK were connected

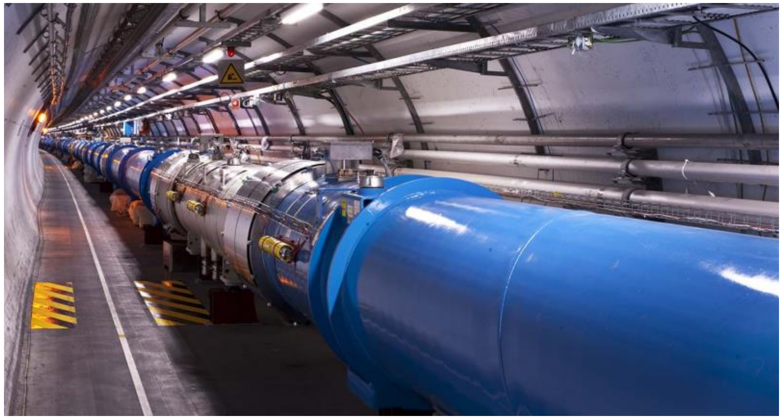
- From the 1980's the commercial sector began to get interested

Phase 2: the early web

- ***Tim Bernes-Lee*** is accreted with the invention of the world wide web (www) *circa* 1989
- Born in SW London
- Software engineer
  - background in telecommunications
- Consultant for CERN
  - European Organization for Nuclear Research, Geneva

Phase 2: the early web

- **Scientists would travel to Geneva to carry out large experiments at CERN**
- Lots of data/reports generated
  - Desire to have that data available to their organisation when they went back home
  - People wanted to read about the experiments performed by others
    - collaborate on ideas
    - people
    - groups

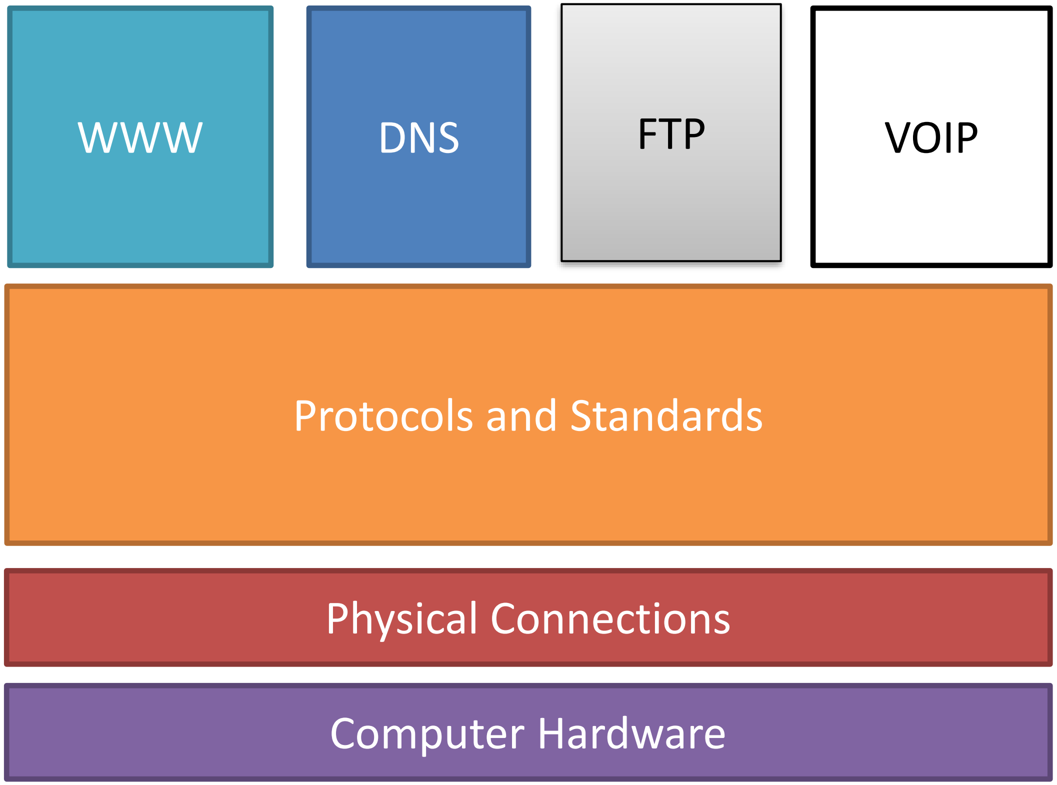


Phase 2: the early web

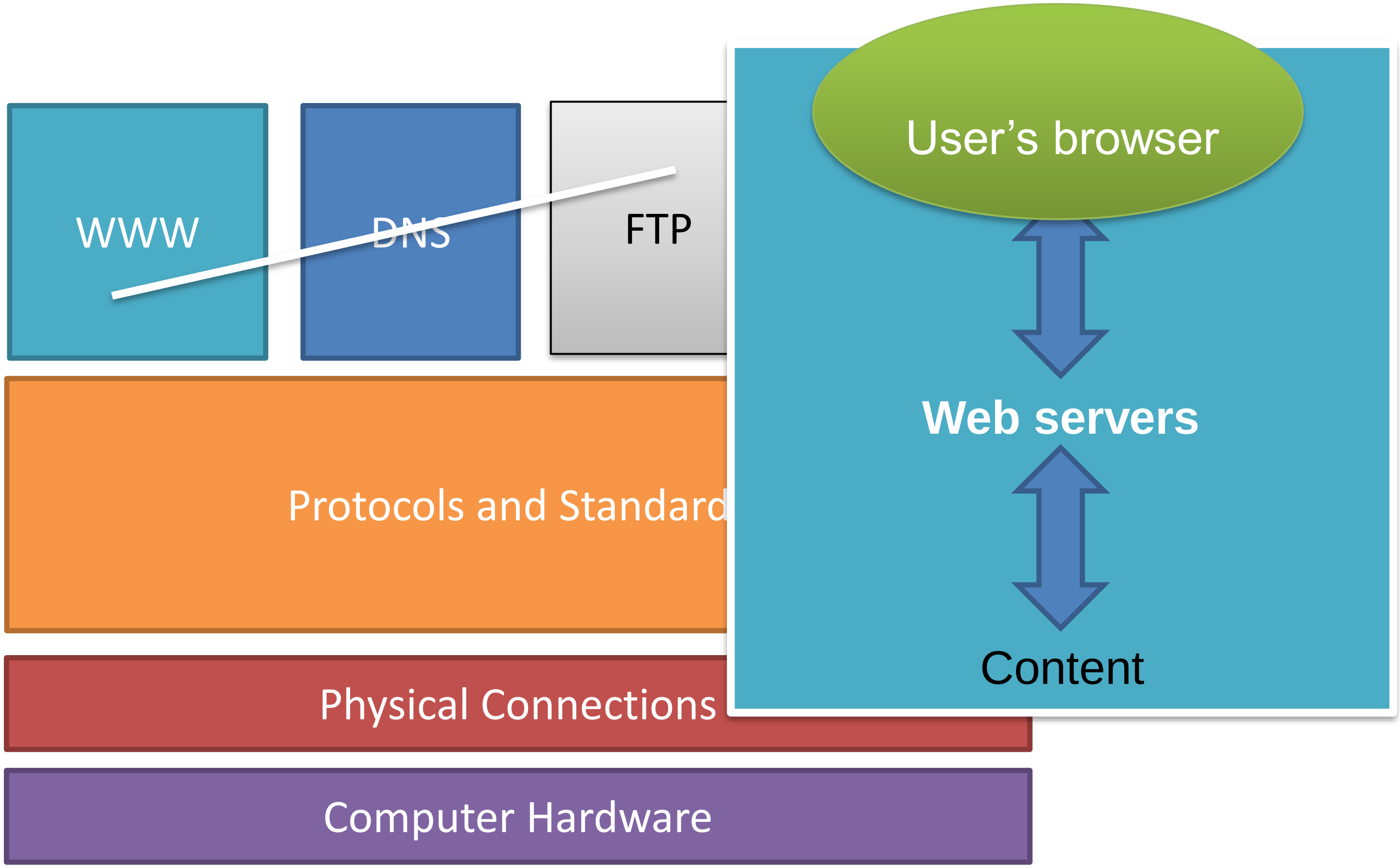
- **While consulting for CERN he put out a plan to exchange data using links between the various documents**
  - Easier for researchers to see what has already been done
- Hypertext (HTML), URI, HTTP
- Ability to view this information regardless of the type of computer you were using
- Tim's ideas were not initially accepted
- First web page viewed in 1990

Source: London 2012 Opening Ceremony

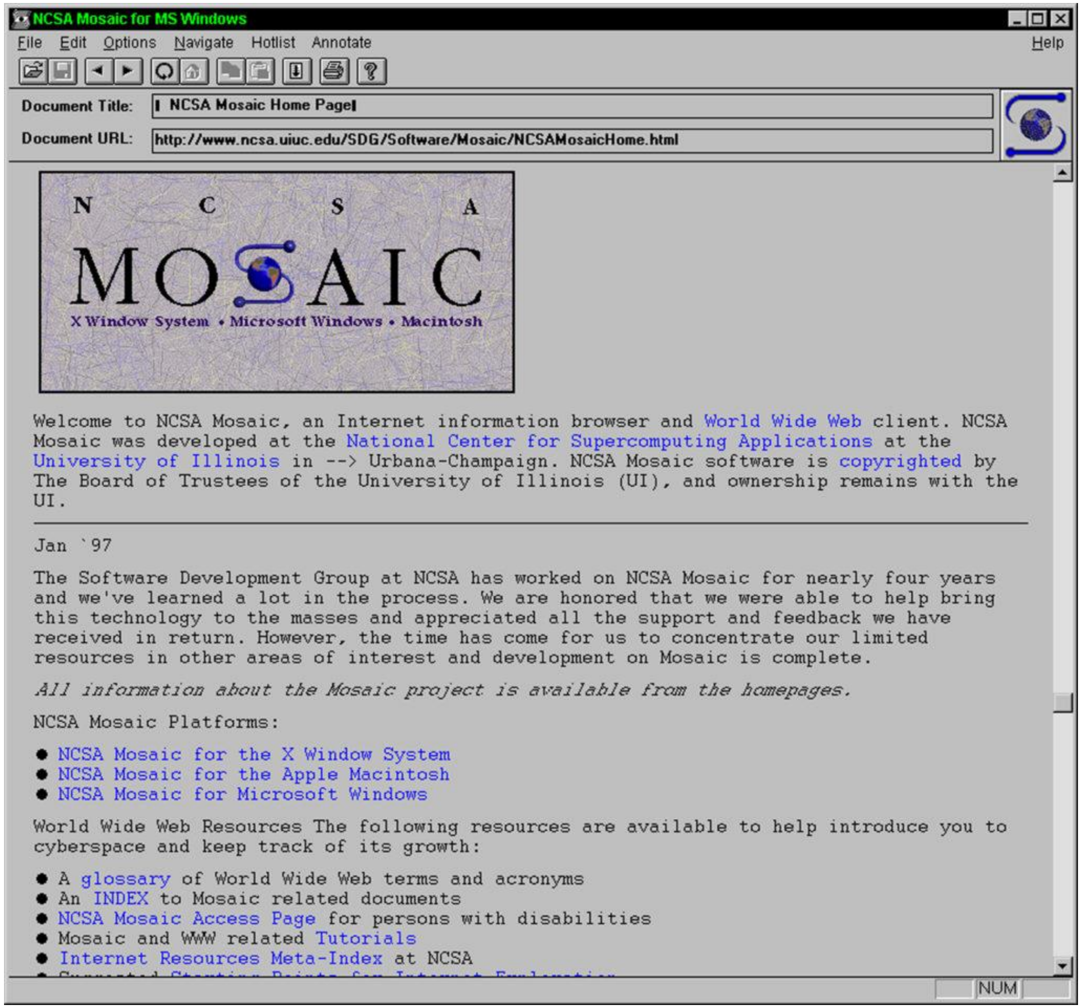
Phase 3: the modern internet



Phase 3: the modern internet



First browser to load pictures


NCSA Mosaic for MS Windows
File Edit Options Navigate Hotlist Annotate Help
N C S A
MOSAIC
X Window System • Microsoft Windows • Macintosh
NUM

Document Title: NCSA Mosaic Home Page

Document URL: http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/NCSAMosaicHome.html

Welcome to NCSA Mosaic, an Internet information browser and World Wide Web client. NCSA Mosaic was developed at the National Center for Supercomputing Applications at the University of Illinois in --> Urbana-Champaign. NCSA Mosaic software is copyrighted by The Board of Trustees of the University of Illinois (UI), and ownership remains with the UI.

Jan `97

The Software Development Group at NCSA has worked on NCSA Mosaic for nearly four years and we've learned a lot in the process. We are honored that we were able to help bring this technology to the masses and appreciated all the support and feedback we have received in return. However, the time has come for us to concentrate our limited resources in other areas of interest and development on Mosaic is complete.

All information about the Mosaic project is available from the homepages.

NCSA Mosaic Platforms:

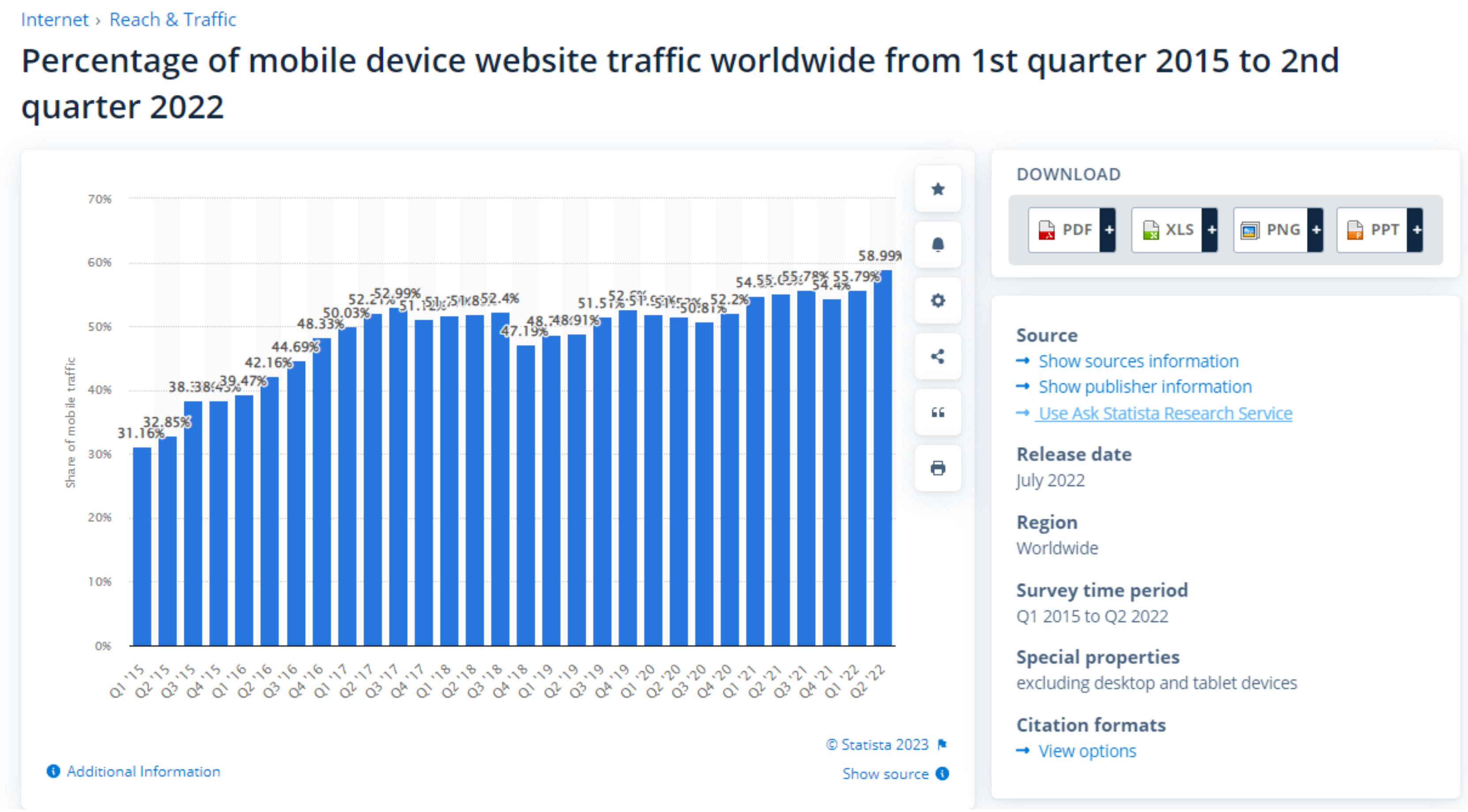
- NCSA Mosaic for the X Window System
- NCSA Mosaic for the Apple Macintosh
- NCSA Mosaic for Microsoft Windows

World Wide Web Resources The following resources are available to help introduce you to cyberspace and keep track of its growth:

- A glossary of World Wide Web terms and acronyms
- An INDEX to Mosaic related documents
- NCSA Mosaic Access Page for persons with disabilities
- Mosaic and WWW related Tutorials
- Internet Resources Meta-Index at NCSA

Source: University of Illinois (2021)

Web Clients



Internet › Reach & Traffic

Percentage of mobile device website traffic worldwide from 1st quarter 2015 to 2nd quarter 2022

Source

→ Show sources information
→ Show publisher information
→ Use Ask Statista Research Service

Release date
July 2022

Region
Worldwide

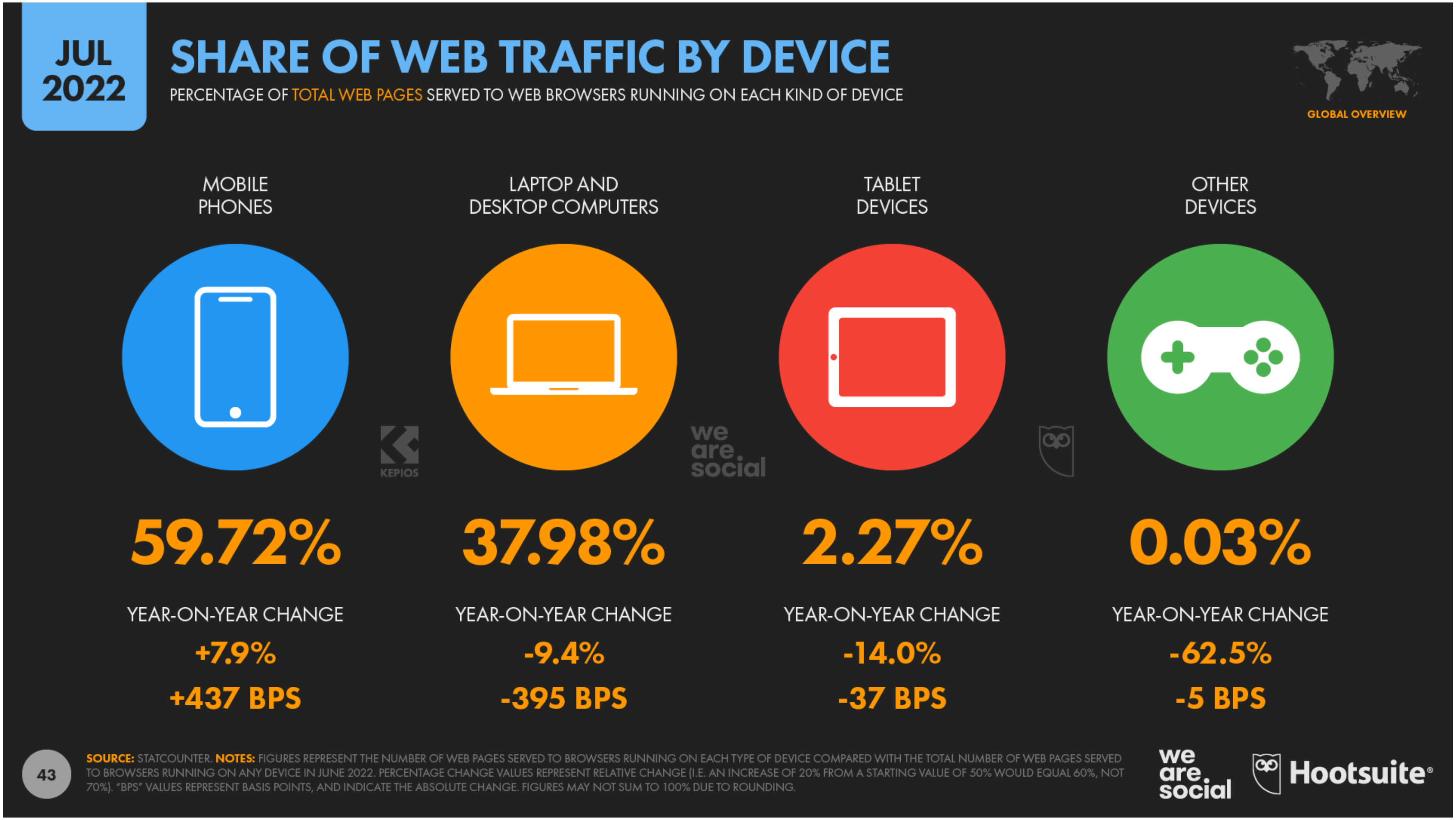
Survey time period
Q1 2015 to Q2 2022

Special properties
excluding desktop and tablet devices

Citation formats
→ View options

Source: Statistica (2023)

Web Clients



JUL 2022

SHARE OF WEB TRAFFIC BY DEVICE

PERCENTAGE OF TOTAL WEB PAGES SERVED TO WEB BROWSERS RUNNING ON EACH KIND OF DEVICE

GLOBAL OVERVIEW

MOBILE PHONES	LAPTOP AND DESKTOP COMPUTERS	TABLET DEVICES	OTHER DEVICES
59.72%	37.98%	2.27%	0.03%
YEAR-ON-YEAR CHANGE	YEAR-ON-YEAR CHANGE	YEAR-ON-YEAR CHANGE	YEAR-ON-YEAR CHANGE
+7.9%	-9.4%	-14.0%	-62.5%
+437 BPS	-395 BPS	-37 BPS	-5 BPS

43

SOURCE: STATCOUNTER. **NOTES:** FIGURES REPRESENT THE NUMBER OF WEB PAGES SERVED TO BROWSERS RUNNING ON EACH TYPE OF DEVICE COMPARED WITH THE TOTAL NUMBER OF WEB PAGES SERVED TO BROWSERS RUNNING ON ANY DEVICE IN JUNE 2022. PERCENTAGE CHANGE VALUES REPRESENT RELATIVE CHANGE (I.E. AN INCREASE OF 20% FROM A STARTING VALUE OF 50% WOULD EQUAL 60%, NOT 70%). "BPS" VALUES REPRESENT BASIS POINTS, AND INDICATE THE ABSOLUTE CHANGE. FIGURES MAY NOT SUM TO 100% DUE TO ROUNDING.

Source: Datareportal (2022)

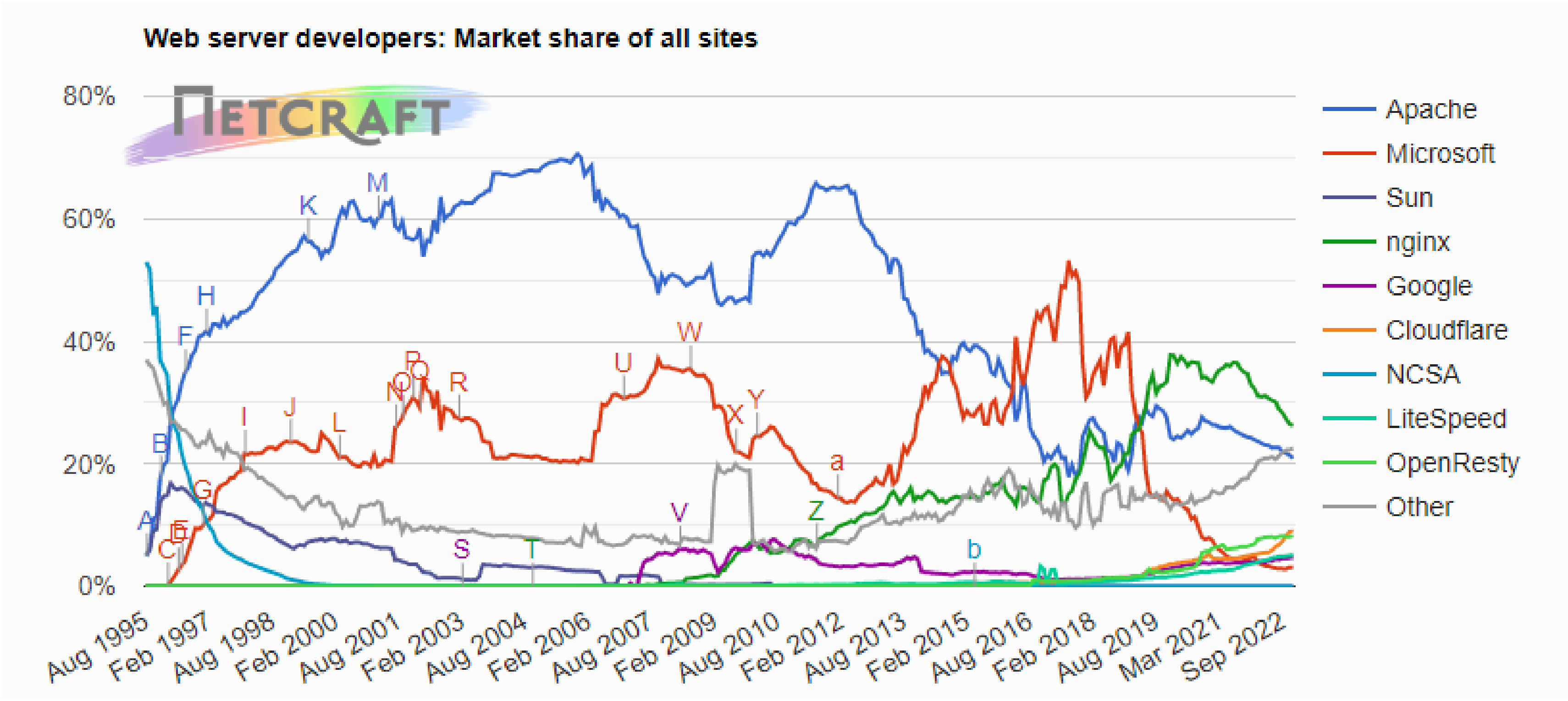
Web servers

- Responsible for handling a user’s request for a particular resource (e.g. web page)

- User enters an address -> responsible web server located -> request forwarded to web server -> web server **serves** the correct response

- Domain Name System (DNS)

Web server market share



Source: Netcraft (2022)

Web server logging

- A *convenient* property of nearly all web servers is that they log/record almost all activity
    - Who requested which resources and when
    - Name and version of their browser
    - Type of device
    - Their locale (language and time zone)
    - ***Failed requests***

Web server logging

- Each request is logged (written to a file)
  - format is semi-standardised (NCSA, W3C, IIS)
- 1:N mapping between page and no. requests
  - N >= 1 (example)
- We can read back these "logs" to determine how a website is used
  - Sometimes referred to as web activity data
  - We can use this data to gather *knowledge*

Web activity data

- All websites have goals and we can *sometimes* observe whether a particular goal is reached
  - website receives an order
  - think back to our *conversion* metrics
- However, not every visitor to a site “converts”
  - free to browse
  - competition
  - users passive
  - anonymous

Web activity data

- Using web activity can study the behaviour of users of our site, regardless of whether or not they “buy” something or not

- We can record when someone makes an order on our website but we are not always sure why people fail to complete the checkout process.

- Data that facilitates this type of analysis can be collected using;
  1. *web server log files*
  2. *page tags*

Page tags

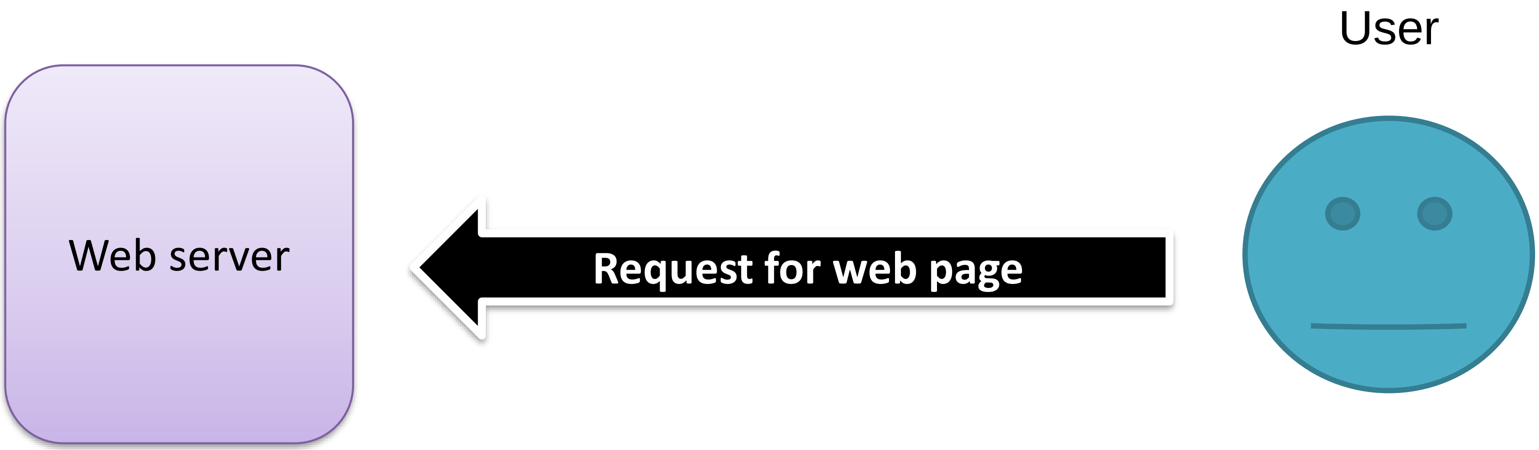
- *Page tags* collect information from the user’s web browser
- They usually consist of small snippets of JavaScript™ code placed on a web page
- The user’s web browser executes them
- They have the ability to disclose more information about the user to the original web server or a third party
- They are a little bit like a ***Trojan*** horses

Page tags – example

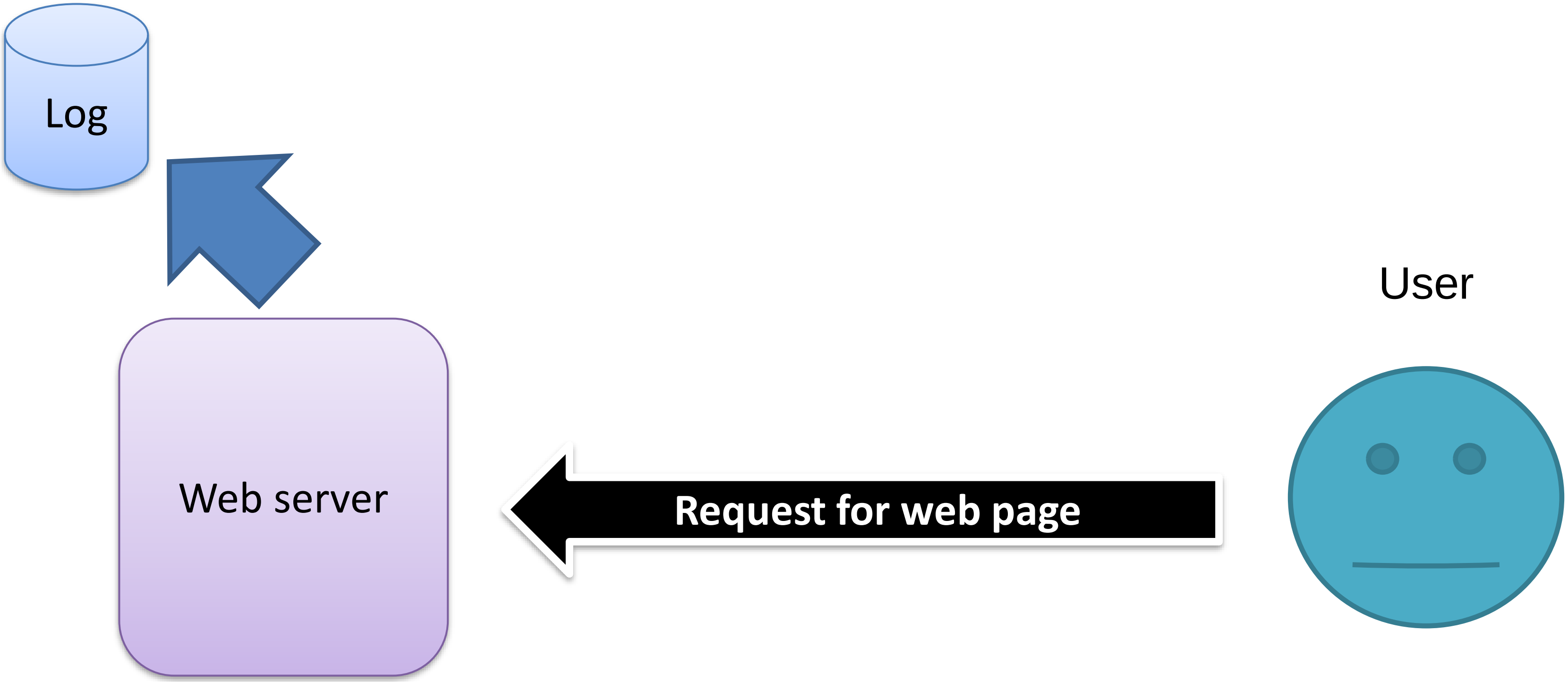
Web server

User

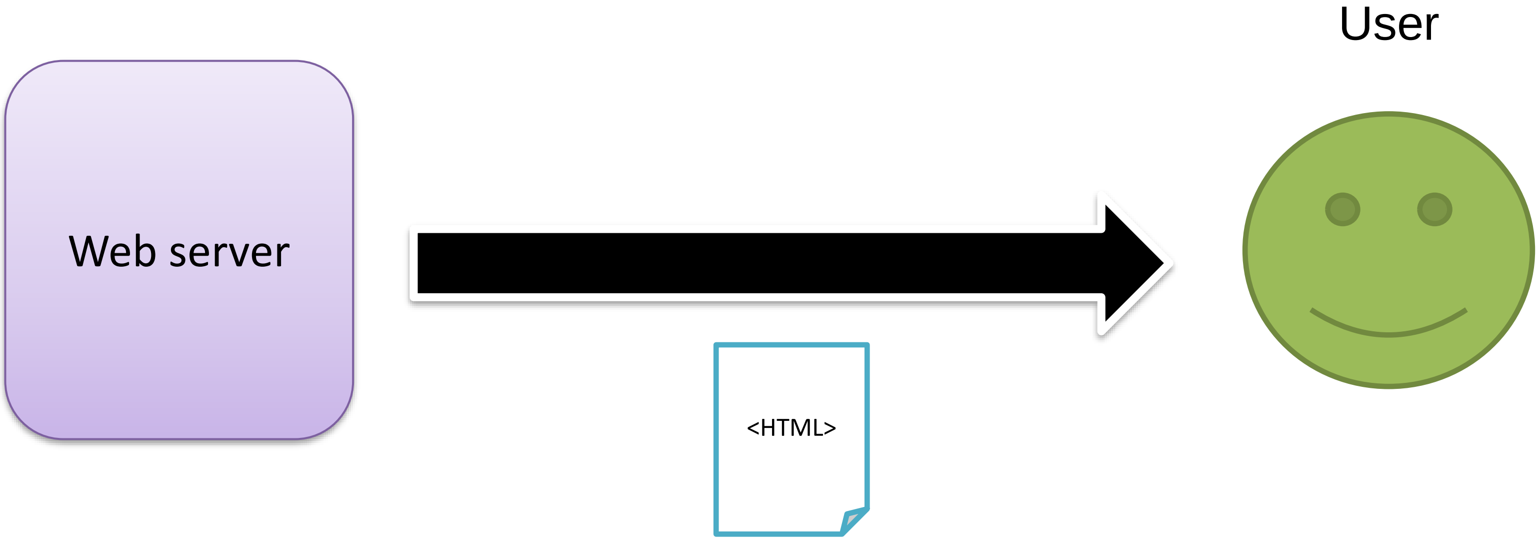
Page tags – example



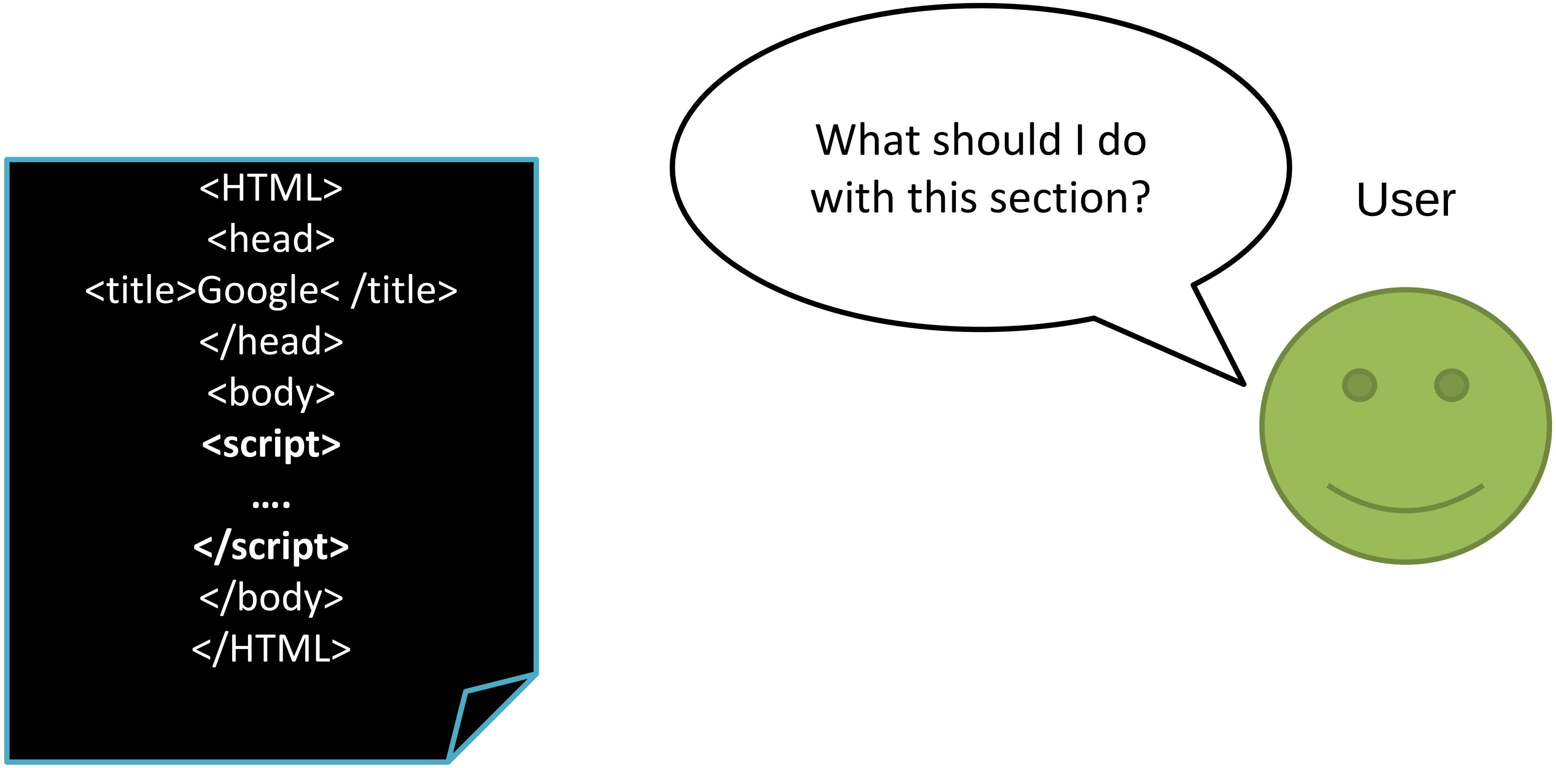
Page tags – example



Page tags – example

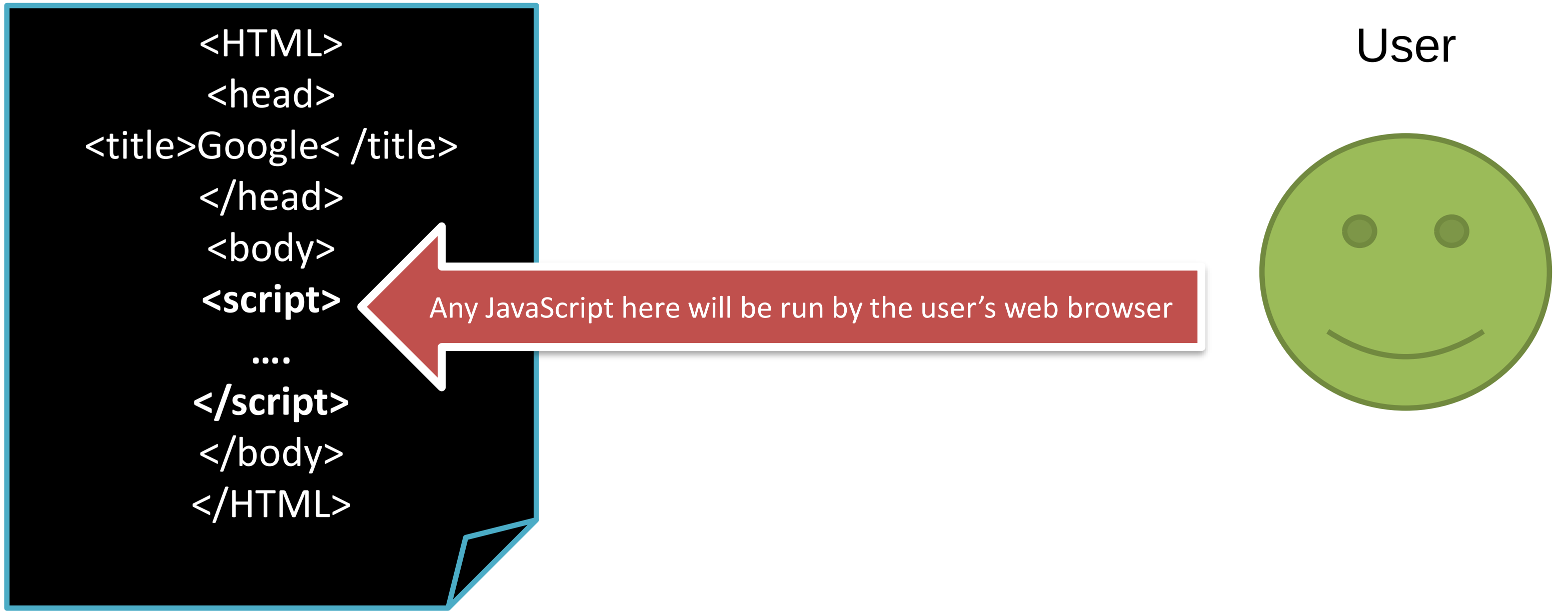


Page tags – example



```
<HTML>
<head>
<title>Google< /title>
</head>
<body>
<script>
....
</script>
</body>
</HTML>
```

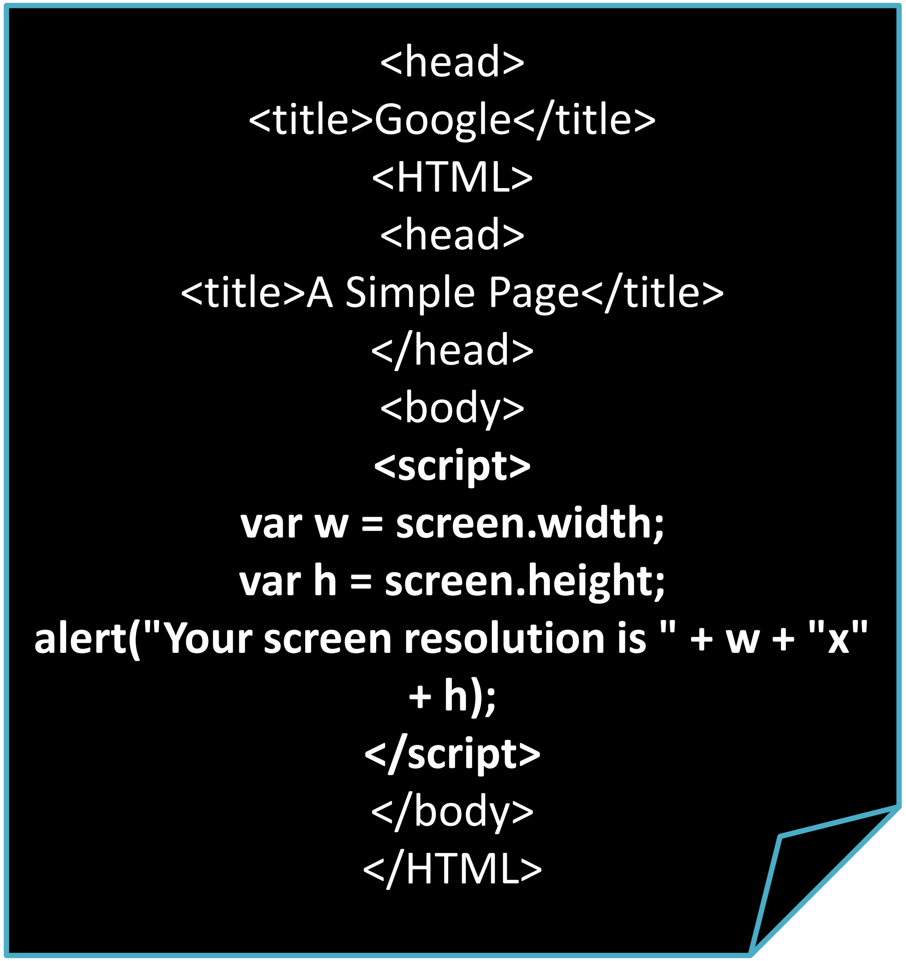
Page tags – example



```
<HTML>
<head>
<title>Google< /title>
</head>
<body>
<script>
....
</script>
</body>
</HTML>
```

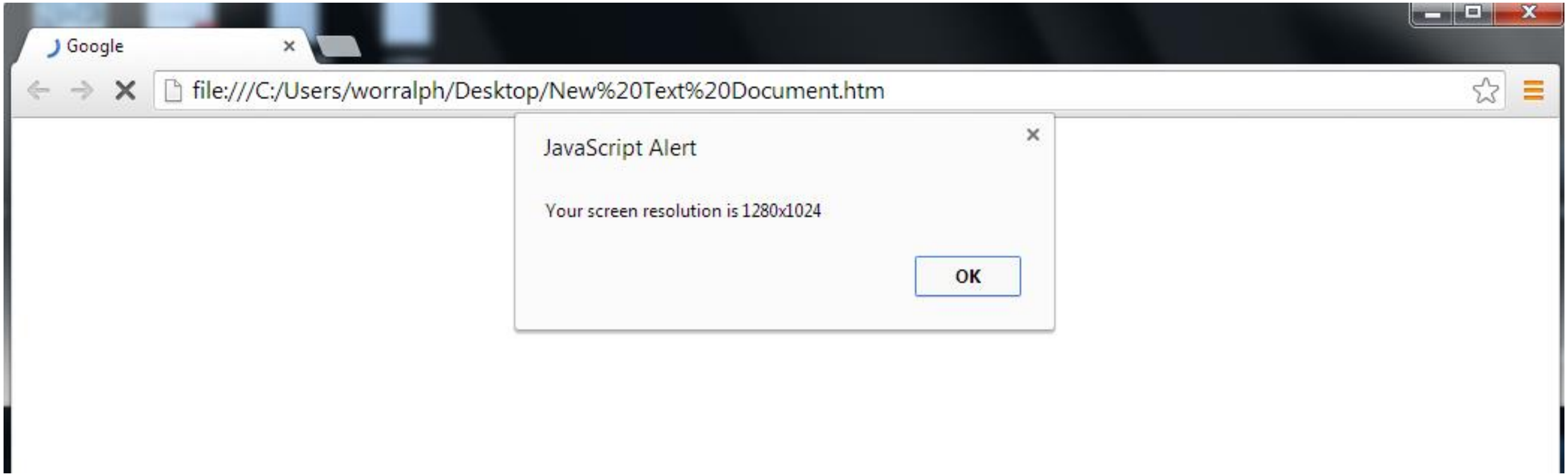
Any JavaScript here will be run by the user’s web browser

Page tags – example



```
<head>
<title>Google</title>
<HTML>
<head>
<title>A Simple Page</title>
</head>
<body>
<script>
var w = screen.width;
var h = screen.height;
alert("Your screen resolution is " + w + "x"
+ h);
</script>
</body>
</HTML>
```

The result



file:///C:/Users/worralph/Desktop/New%20Text%20Document.htm

JavaScript Alert

Your screen resolution is 1280x1024

How does this help?

- PageTags enable the site owner to run additional business logic on the user’s browser
- Through the use of PageTags the owner of the site can collect a much broader range of data about each user’s interaction with the site (e.g. Mouse events like scrolling or hovering)
- The collected data can be sent back to the webserver without the user having to directly interact with the site

An example using AJAX (asynchronous JavaScript + XML)

```
<script>
var http = new XMLHttpRequest();
var url = “http://localhost:12345";
var w = screen.width;
var h = screen.height





</script>
```

An example using AJAX

```
<script>
var http = new XMLHttpRequest();
var url = “http://localhost:12345";
var w = screen.width;
var h = screen.height
var params = “width=“ + w + “&height=“ + h;
http.open("POST", url, true);
http.setRequestHeader("Content-type", "application/x-www-form-urlencoded");
http.setRequestHeader("Content-length", params.length);
http.setRequestHeader("Connection", "close");




</script>
```

An example using AJAX

```
<script>
var http = new XMLHttpRequest();
var url = “http://localhost:12345";
var w = screen.width;
var h = screen.height
var params = “width=“ + w + “&height=“ + h;
http.open("POST", url, true);
http.setRequestHeader("Content-type", "application/x-www-form-urlencoded");
http.setRequestHeader("Content-length", params.length);
http.setRequestHeader("Connection", "close");
http.onreadystatechange = function() {
        if(http.readyState == 4 && http.status == 200) {
                alert(http.responseText); }
        }
 http.send(params);
</script>
```

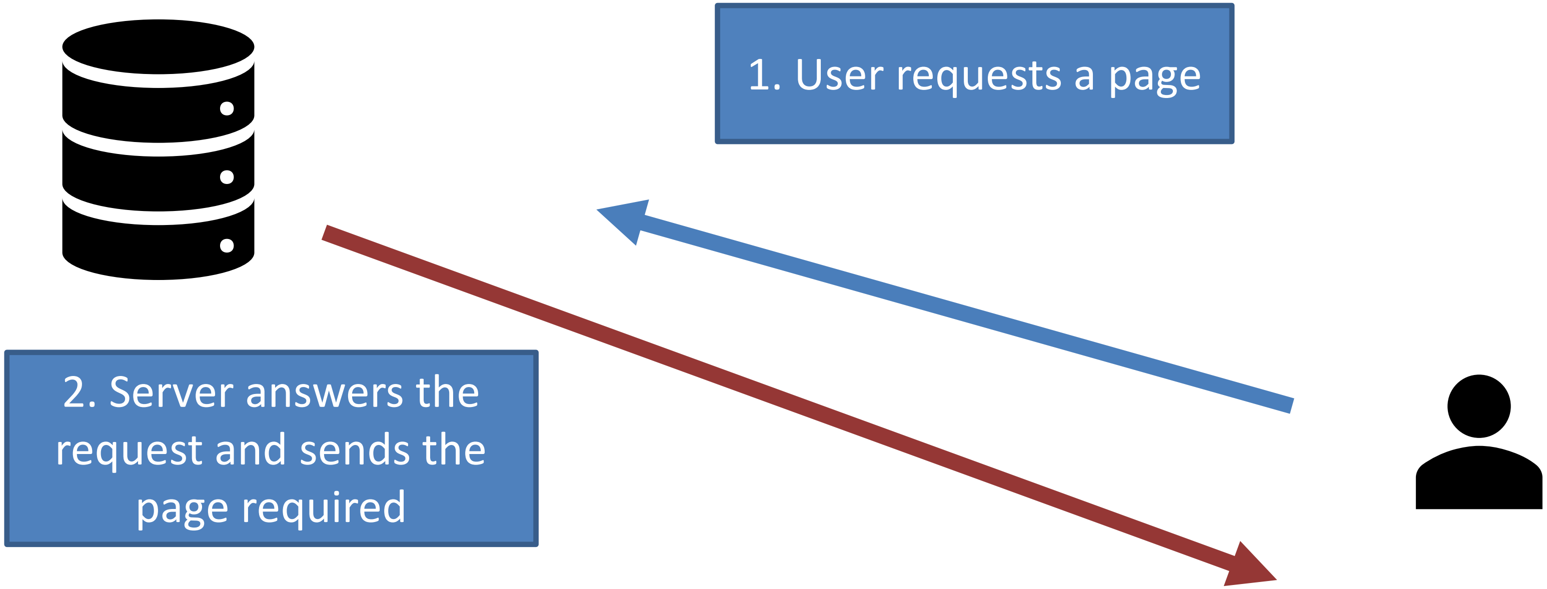
Third party page tags

- Many organisations have developed a set of standard page tags for use with their own analytics service
    - Google Analytics

- They are programmed to collect a significant amounts of information about the user (including screen resolution)

- More often than not, “HTTP cookies” are used alongside page tags to enable identification of individual users of a site

HTTP Protocol basics



1. User requests a page

2. Server answers the request and sends the page required

Server doesn’t store any information about previous requests, every request is considered a new request

Pages cannot be personalised because the server cannot distinguish between users

Stateless web

- The nature of the web is that it is stateless and thus characterised by two important features.

1. No memory of the past
    - No persistence

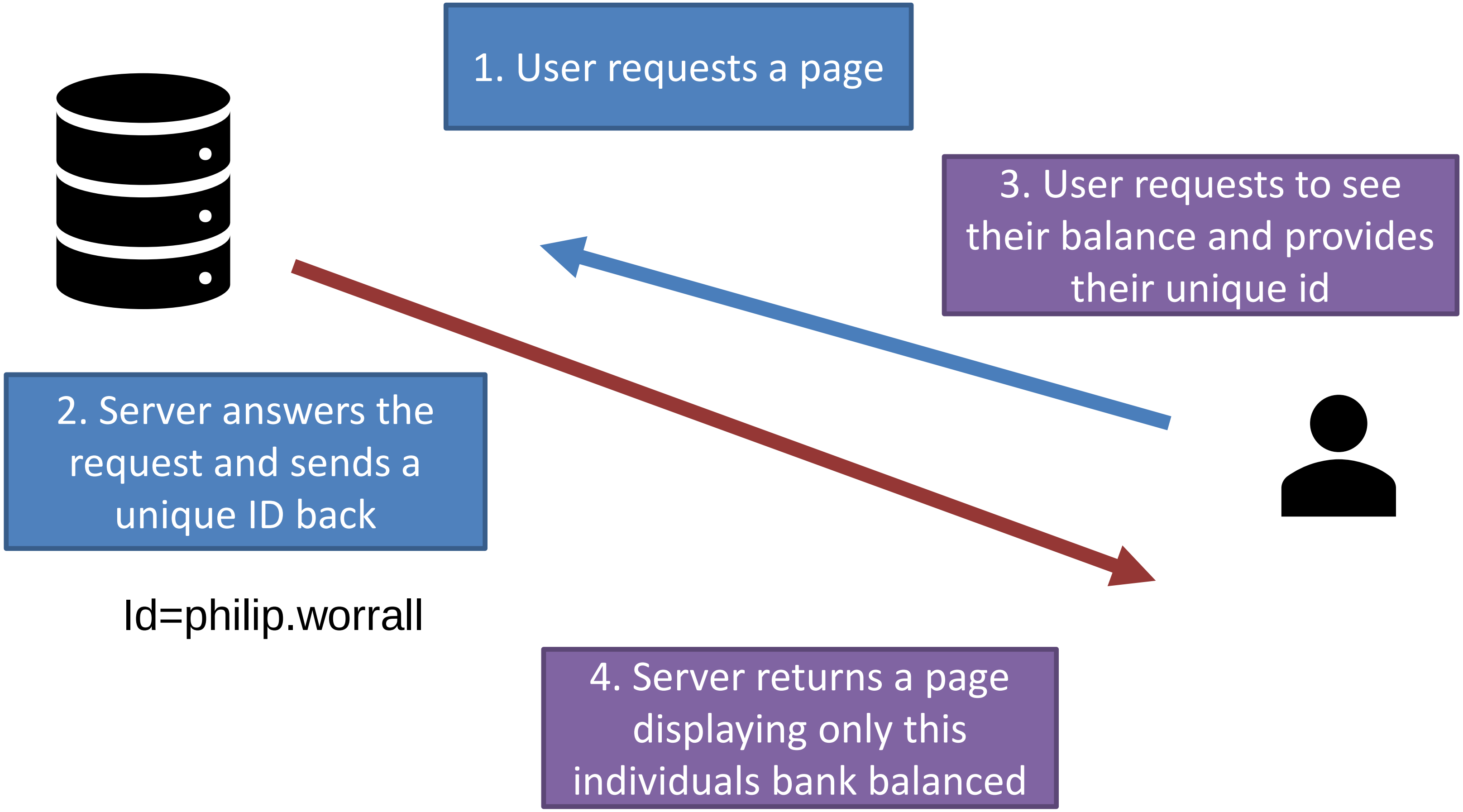
1. Each request is handled in isolation
    - No ordering
    - No dependencies

- Think GP. vs Pharmacy

HTTP Cookies

- HTTP Cookie
  - A data storage facility used by the web server to preserve some state between requests
  - The server tells each client to remember a unique value
  - Each time the user requests a page the client sends the unique value to the server
  - The server knows which unique values map to particular clients so for example you can only see the bank balance of your own account

HTTP Protocol (with cookies)



1. User requests a page

2. Server answers the request and sends a unique ID back

Id=philip.worrall

3. User requests to see their balance and provides their unique id

4. Server returns a page displaying only this individuals bank balanced

Sources of data

- Today we have identified two different sources of data that we can use in web analytics
  - web server logs
  - page tags

- Other sources of data including packet sniffing, UX and UI experiments and clickstream data.
  - Packet sniffing involves the monitoring of data across a network, this can be used on internal websites where both sides of the data exchange can be observed.
  - UX and UI experiments can be conducted with testers.
  - We will look at clickstream data next week.

Web metrics

- Using the data collected either through web logs or page tags we can calculate a range of performance metrics to determine a website’s effectiveness
- There is *no standard set of metrics* and in many respects the ones a particular website may be interested in will depend on its chosen goals
  - think of new site vs. an established site.
- There are however a range of standard metrics that most websites are interested in

On-site vs. off-site metrics

- *On-site* metrics are calculated using data collected from a single web site
  - web server log file
  - page tags

- Off-site metrics are calculated by third parties using data from several websites
  - Alexa Rank™
  - Google PageRank™
  - Shares
  - Likes (+1s)

Common on-site metrics

- Hits
  - total number of requests sent to the web server
- Page views
  - total number of requests for web pages (html)
- Bounce rate (%)
  - proportion of visits that request only a single page
- Exit rate (%)
  - page level
  - proportion of people who leave after viewing specific page
    - *checkout page*

Some metrics can be broken down

- Visits
  - new (first) visitors
  - unique visitors
  - returning visitors
  - average pages visit
  - average length of visit / visit duration

- User views 10 pages in the morning and 5 pages at 9pm.
  - How many visits?
    - 1, 2, 15?

Challenges to measuring

- There are a number of issues with how some on-site metrics are calculated
  - we have already seen the problem of counting visits
- First visits are usually important to a site but in general are often over estimated
  - many users delete their cookies
  - access website from a different machine
    - IP address
  - change web browser
- Visit duration?

Dynamic nature of metrics

- The ever changing web calls for a more dynamic view of web metrics whereby we consider the evolution of site metrics overtime

- For this reason, when we report web metrics we often quote the figures from previous periods and show how it has improved/worsened
  - Relative vs. absolute performance

In Summary

- Today we have built up our collective understanding of the web and how it works.
- Web servers are responsible for delivering the necessary images, text and video that make up a modern web page to the user's browser.
- Web servers record user activity in the form of log files, such data is often useful to web analysts as it enables us to study user behaviour.
- Page Tags are an alternative source of data to web server log files and consist of small snippets of JavaScript™ code embedded on a web page.
- Web metrics are different measures of a website's performance that can be generated using data collected from web server logs or page tags.
- Metrics relevant to an organisation may depend on their individual objectives, goals and the industry they operate in.
- The stateless nature of the web requires HTTP cookies to be used in order to associate a user with a particular set of requests.

Learning week 2

- In LW2 we will look at an alternative source of web activity data and how such data can be used to analyse web performance using different statistical measures.

- **A reminder to**
  - Sign up for a Twitter Developer account
  - Familiarise yourself with the module handbook
  - Complete the activities in the weekly tutorial packs

End