Web and Social Media Analytics

Web Analytics (Knowledge Gathering Part II)

Dr Philip Worrall

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

LW2

Outline

- Recap of last week's material
- Clickstream data
  - Graphical Representations
  - Ash's (2012) Marketing Funnel
  - Heatmaps
- Surveys and polls
- Key Performance Indicators (KPI's)
- A/B Testing and Multivariate Testing

Recap of Last Week's material

Success on the Web

Web Metrics

Client Device

Page Tagging

Web Logs

HTTP Cookies

Web Server

Bounce Rate

Stateless Web

Advantages/Disadvantages

Web logging

ADVANTAGES

- Straightforward to collect
- Tracks non-HTML resources
- Captures status code of each request
- Doesn't depend on JavaScript

DISADVANTAGES

- Requires admin access to server
- Small performance impact
- Limited data collected about the user

Page Tagging

ADVANTAGES

- No changes required to the server
- Captures richer set of data
- Wealth of existing commercial and free tools

DISADVANTAGES

- Tracks HTML files only
- Impacts upon page load time
- Code must be added to all pages
- Coding errors can cause a page to fail to load correctly

Software tools

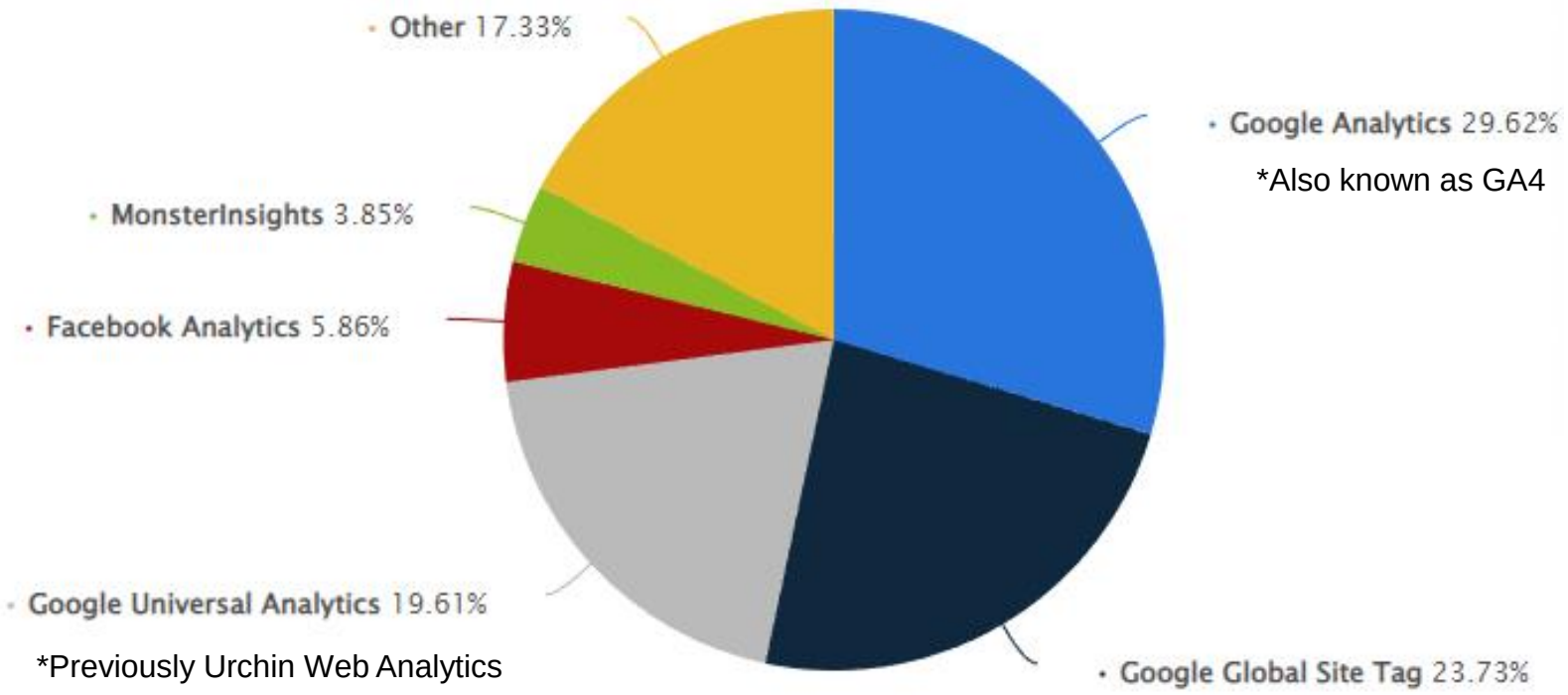
- For both **web logging** and **page tagging** there are a range of **software tools** for calculating and monitoring changes in **key web metrics.**

- Key areas of differentiation
  - Feature set (metrics calculated and functionality)
  - Data collection strategy (Page tagging vs. web server logs)
  - Data storage (stored by the third-party provider or kept locally, SQL)
  - Price (open source, commercial)

- It is estimated Google Analytics (LW5 tutorial) has the largest market share among the most visited sites on the web.

Popular tools

Estimated market share of web analytics software products**.**



(Statistica, 2023)

AWStats

Statistics for:
destailleur.fr

Summary
When:
Monthly history
Days of month
Days of week
Hours
Who:
Countries
- Full list
- Regions
- Cities

Hosts
- Full list
- Last visit
- Unresolved IP Address

Authenticated users
- Full list
- Last visit

Robots/Spiders visitors
- Full list
- Last visit

Navigation:
Visits duration
File type
Downloads
- Full list

Viewed
- Full list
- Entry
- Exit

Operating Systems
- Versions
- Unknown

Browsers
- Versions

Last Update: 15 Feb 2022 - 00:00

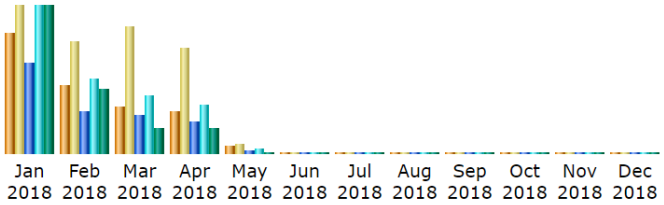
Reported period: Apr 2018 OK

Summary

Reported period	Month Apr 2018				
First visit	01 Apr 2018 - 00:00				
Last visit	30 Apr 2018 - 23:57				
	Unique visitors	Number of visits	Pages	Hits	Bandwidth
Viewed traffic *	**2,304**	**5,921** (2.56 visits/visitor)	**8,437** (1.42 Pages/Visit)	**12,886** (2.17 Hits/Visit)	**120.89 MB** (20.9 KB/Visit)
Not viewed traffic *			**65,278**	**76,730**	**177.54 MB**

* Not viewed traffic includes traffic generated by robots, worms, or replies with special HTTP status codes.

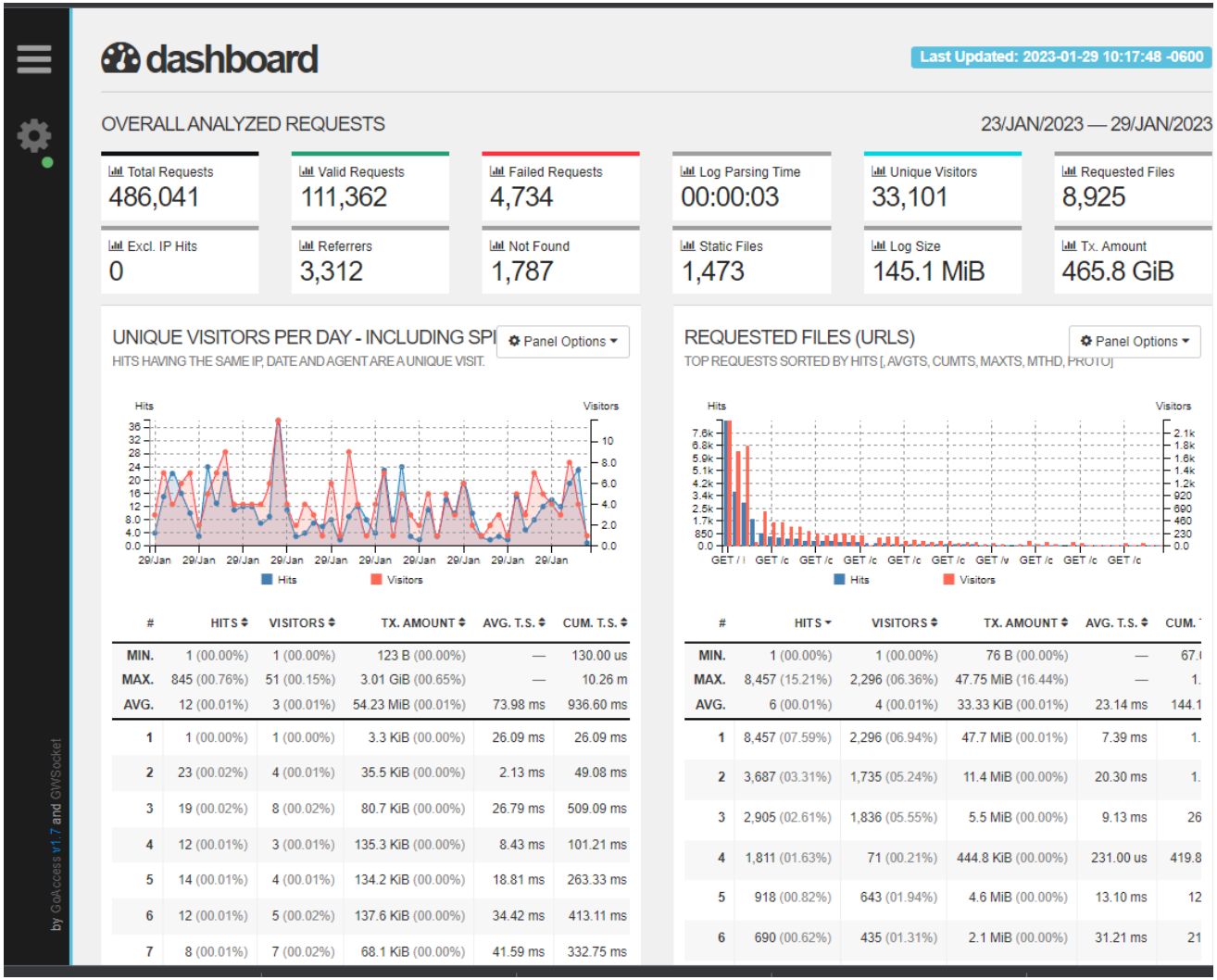
Monthly history



Month	Unique visitors	Number of visits	Pages	Hits	Bandwidth
Jan 2018	6,687	8,200	23,928	38,961	716.45 MB
Feb 2018	3,815	6,283	11,188	19,504	317.85 MB

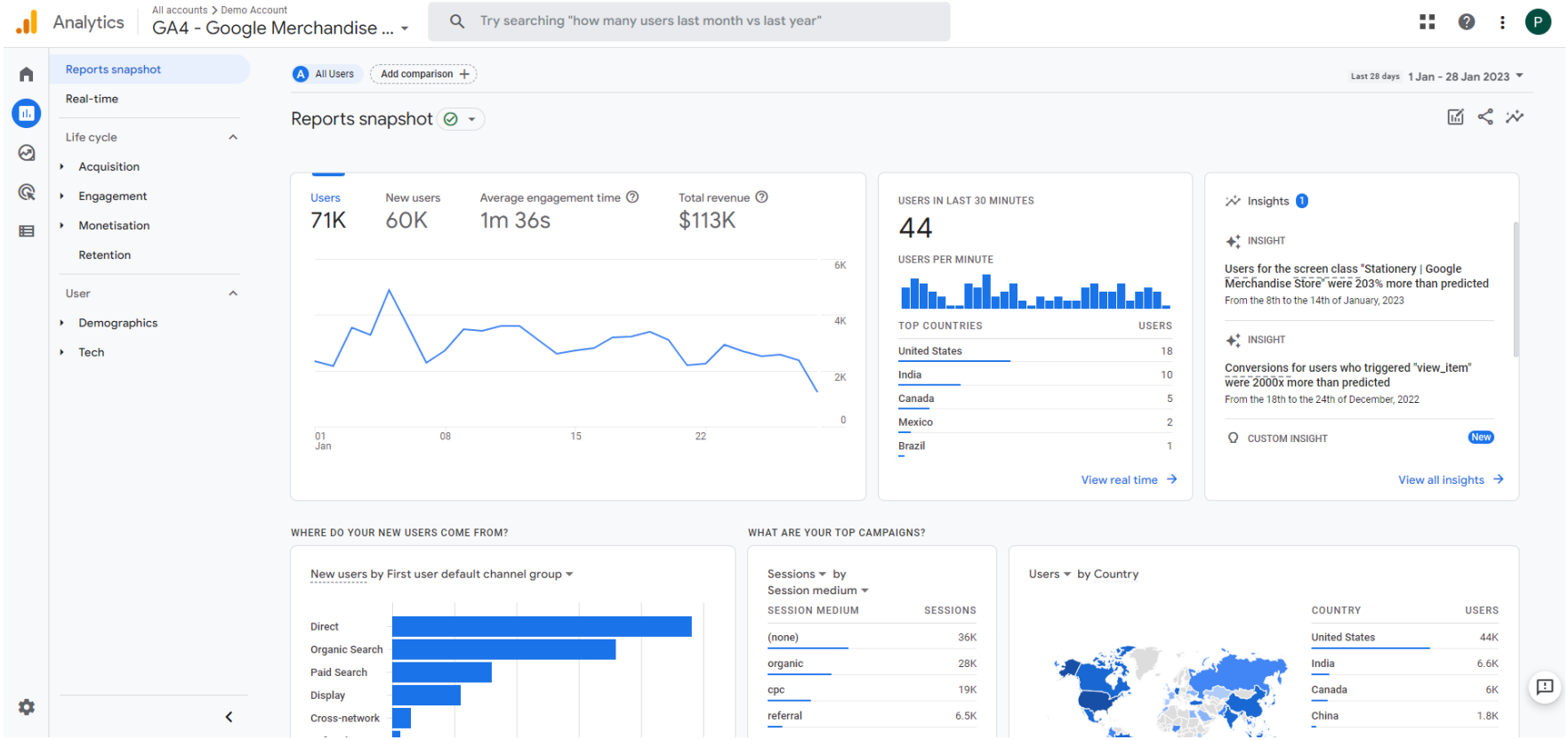
https://awstats.sourceforge.io

GoAccess



https://goaccess.io

Google Analytics (GA4)



https://analytics.google.com/analytics/

Clickstream data

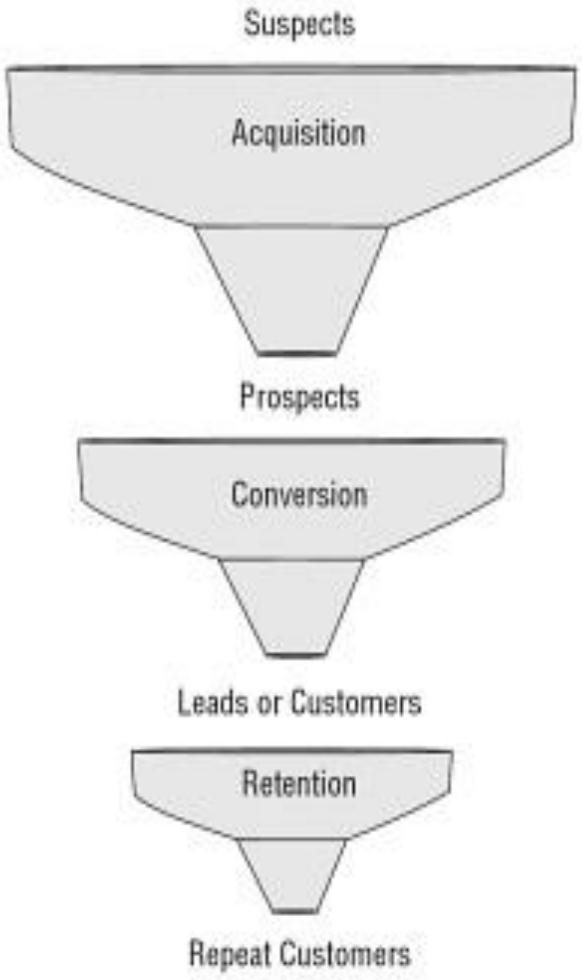
DEFINTION:

Data generated by the **recording** of an individuals clicking **behaviour** during an interaction with a web site.

Clickstream data helps us to understand:

- **How** individuals navigate through a site
- Pages viewed and critically the **sequence**
- Where ***exactly*** on a page users click
- The success and impact of **external** links
- Performance of **digital marketing** campaigns

Ash's (2012) marketing funnel



- **Digital Marketing** involves the use of digital approaches to increase demand for a company's products or services.
- The goal of a marketing campaign is to encourage people to complete an **action**, e.g. buy a product or sign-up to a free trial.
- The first stage involves **acquiring new prospects**, this can be through email, affiliate or social media campaigns.
- Clickstream data can help us to quantify the effectiveness of different campaigns by collecting information on the **acquisition method** of a new prospect.

An extract of a web log

Time	IPV4 Address	User Agent	Cookie	URI
09:01	182.1.1.21	Windows/Chrome	NULL	http://www.abc.com/
09:02	182.1.1.21	Windows/Chrome	NULL	http://www.abc.com/search.html
09:03	182.1.1.21	Windows/Chrome	NULL	http://www.abc.com/offers.html
09:03	182.1.1.21	Windows/Chrome	NULL	http://www.abc.com/cart.html
09:08	182.1.1.21	Windows/Chrome	NULL	https://www.abc.com/order.html
09:09	182.1.1.21	Windows/Chrome	NULL	http://www.abc.com/search.html
09:09	112.1.19.122	MacOS/Safari	A6AHX	http://www.abc.com/

Questions

How many visitors did the site receive?

How many page views?

Can we use this web log for clickstream data analysis?

Inspecting the URI

DEFINTION:

The **Uniform Resource Identifier** (URI) is a unique sequence of characters that uniquely identifies a resource available on the internet.

For example:

https://www.estore.com/depts/clothing/shoes?ob=price&m=desc

The **URI** consists of:

- The scheme
- The subdomain
- The domain name and top-level domain
- The path
- The query string

Which of the **6 components** of the URI do you think may be useful collecting clickstream data?

Having modified the URIs

Time	IPV4 Address	User Agent	Cookie	URI
09:01	182.1.1.21	Windows/Chrome	NULL	http://www.abc.com/
09:02	182.1.1.21	Windows/Chrome	NULL	http://www.abc.com/search.html?nav=top
09:03	182.1.1.21	Windows/Chrome	NULL	http://www.abc.com/offers.html?c=shoes
09:03	182.1.1.21	Windows/Chrome	NULL	http://www.abc.com/cart.html
09:08	182.1.1.21	Windows/Chrome	NULL	https://www.abc.com/order.html?nav=visa
09:09	182.1.1.21	Windows/Chrome	NULL	http://www.abc.com/search.html
09:09	112.1.19.122	MacOS/Safari	A6AHX	http://www.abc.com/promo/jan_4_email

Questions

Which offers was the first visitor interested in?

Which email campaign did the second visitor likely see?

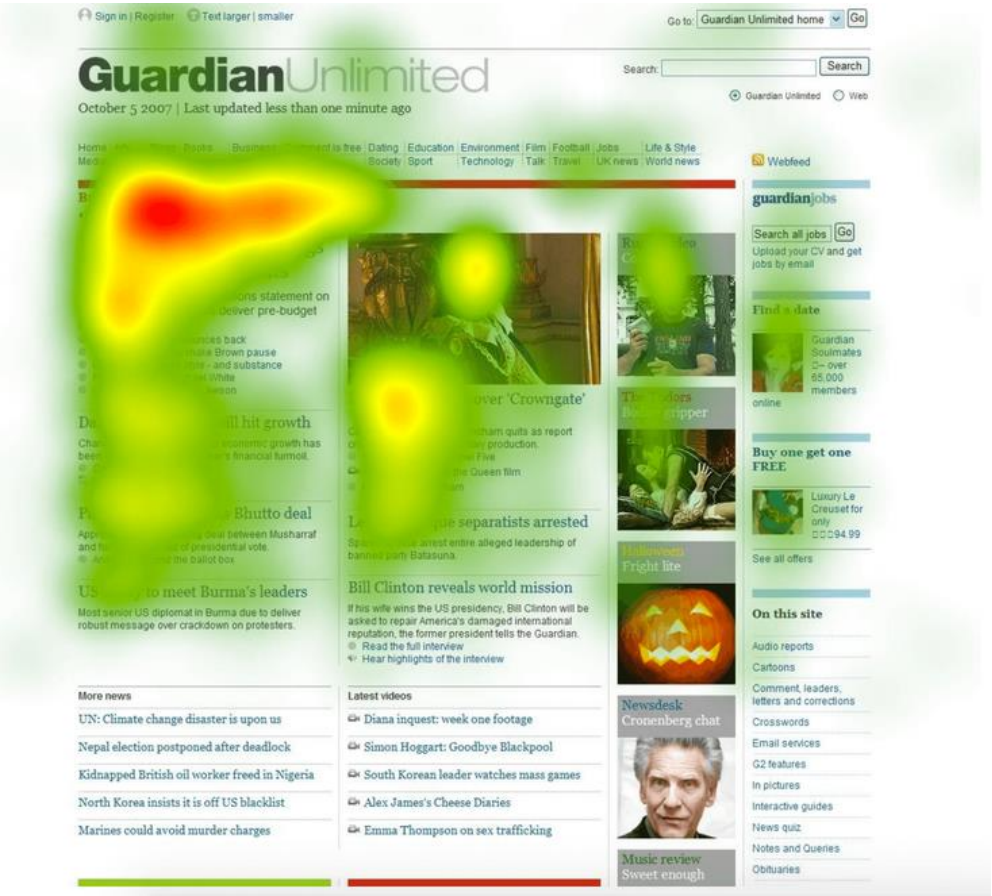
Which credit card did the first user intend to pay with?

Visualising Funnels

Clickstream data can be aggregated to show the **flow of users**.

They are also known as conversion funnels, path funnels or path exploration charts

Heatmaps (click maps)



(The ascent, 2023)

- Clickstream data can also be aggregated to form **heatmaps**
- Heatmaps highlight the **density** of clicks in a particular area
- Heatmaps can help support **conversion metrics**, since we can identify where users typically focus their attention
- Heatmaps can be created by most analytical tools, although they are sometimes referred to as **overlays** or **click maps**

Surveys and Polls

- **So far**, we have seen three **indirect** ways to measure **user behaviour**
  - Metrics gathered by web logging or page tagging
  - Using clickstream data to model the interaction process.

- **Surveys** and **polls** provide two other more **direct** ways to gather information about users. They can also provide a means to increase **user engagement**.

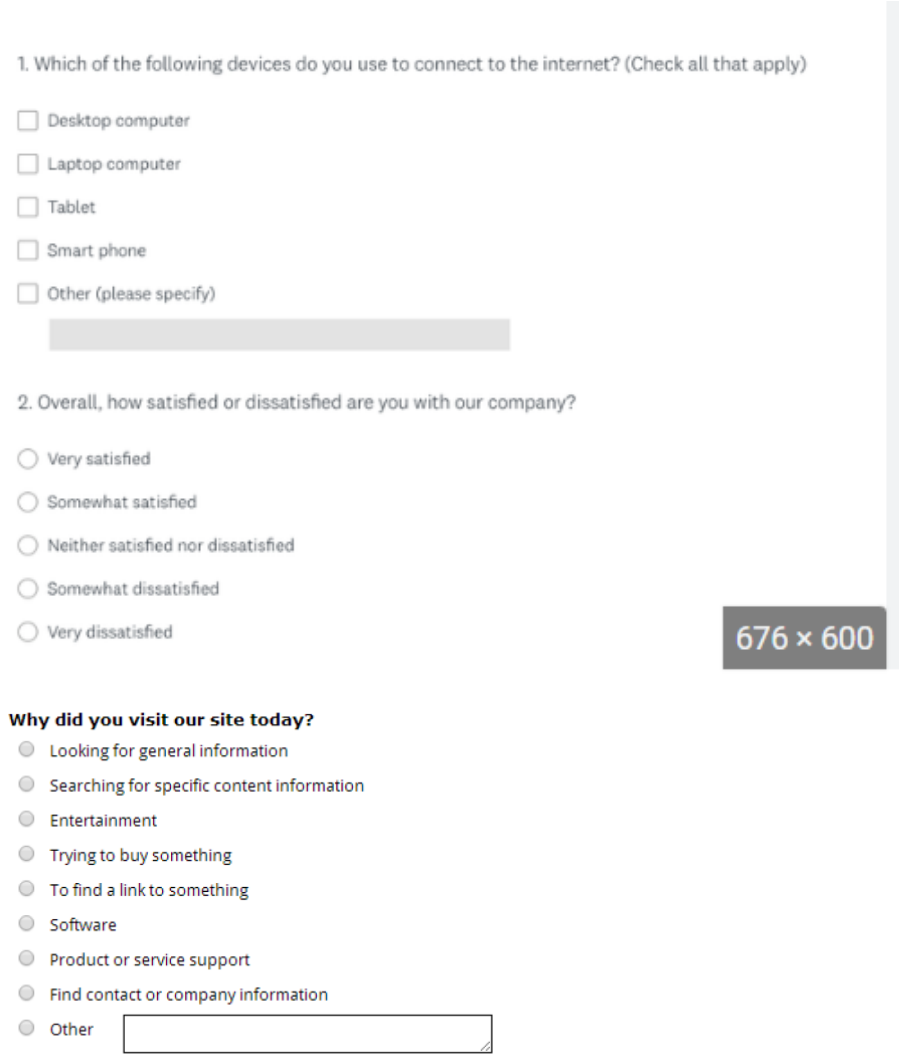
Surveys

Sample the website's audience by inviting users to take part in a questionnaire. Our intention is to draw statistical inferences about the users on a site, their likes/dislikes and preferences.

Polls

Sample the website's audience by inviting users to answer a multiple choice style question. Our intention is to gauge the opinions of users on a site.

Surveys and Polls



- **Surveys** are characterised by having more than one question.
- Surveys will also typically provide a way to users to enter **free text** or provide an extended answer.
- The +1 or Like feature on social media sites in as example of a **Poll**.
- Like their offline counterparts, surveys and polls can be subject to **bias,** this includes**:**
  - Sampling bias
  - Response bias
  - Confirmation bias

Web analytics and KPIs

"Most people are using web analytics as a benchmark: how did we do yesterday, and how are we doing today? Smart people are actually analyzing to optimize their website. The advanced people are using Web data to optimize all of their marketing."

Jim Sterne, founding director and chairman of the Web Analytics Association

Everything we have discussed so far is **useless** if it is not translate to **actions**.

Web analytics and KPIs

- Software tools (like Google Analytics, AWStats, GoAccess etc.) generate reports
- Reporting is **NOT** analysis
- Reports → provide data
- Analysis → provides insights
- You cannot expect to get significant insights from any analytics tool simply by implementing it
- In any case, a lot of reports and analysis may not be relevant or a priority to the organisation in question (many important metrics are missing)
- Implementing the analytics tool is only the start

The web analytics process

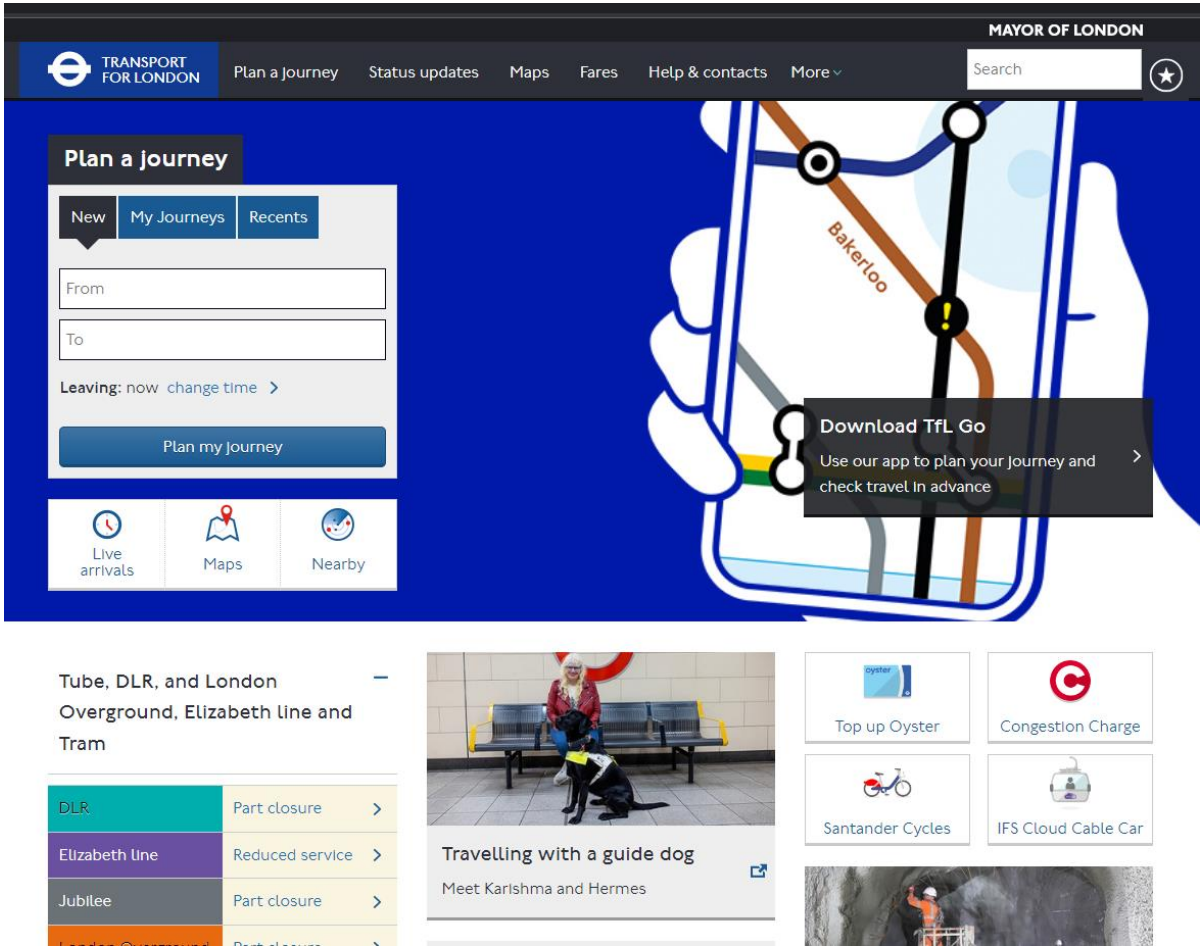
1. Identification of organisational **issues**
2. Develop an understanding of the organisation's **purpose, strategy** and **goals**
3. Determine what **measures of success** might be relevant and convert them into **web metrics**
4. Implement **data collection** and any required tools
5. Use the longer term-aims and objectives of the organisation to create shorter term **Key Performance Indicators** (KPI)
6. Monitor these KPIs over time. Make **decisions** and take **actions**.
7. Re-evaluate **progress** and make appropriate **adjustments**.

Structure of a KPI

#1 KPI Name (author, date created)

	RAW	PROGRESS	CHANGE
Example	10 new users	10% of all visitors	+10% more donations
Source of data	Does the data already exist or do we need to collect it? How far back does the data go? Is there an existing Google Analytics report that we can use? Is the data stored locally?		
Frequency	Daily? Weekly? Monthly? Quarterly? Yearly? How often should we monitor our progress?		
Target	What is the target we are trying to hit? How can we track progress graphically? All KPIs must be temporal. What is the deadline?		
Alignment	How does the KPI relate to the organisation's long-term strategy and goals?		

Case study: Transport for London



Question? What might the long-term goals of the organisation be? How might it measure them?

Case study: Transport for London

Lets consider that one of the long-term goals of TFL is be ranked in the **top 10 for customer support**.

However, suppose that calls to its helpline have been increasing and in many cases people have to wait **more than 30 minutes** to get a response.

Many of such calls are to ask about planned engineering works taking place over the weekend.

This information is already provided on the TFL website but apparently not many people read it.

A potential KPI for TFL…

#1 TFL, Increase proportion of visitors that see planned engineering information

	RAW	PROGRESS	CHANGE
Example		20% of all visitors	
Source of data	Web log data, collected locally and available daily. The raw data is stored on a MySQL database and queried using SQL. The URI is present in the log as is COOKIE_ID.		
Frequency	Weekly		
Target	20% of visitors to view the planned engineering information page before the start of the next financial year. A graphic should be created to plot the live percentage of visitors that have viewed this page on a weekly basis.		
Alignment	It directly contributes towards improved customer service by reducing the amount of calls to the customer service number.		

Conducting the experiment…

- With the KPI in place, we now brainstorm potential **actions** we could take to ensure the KPI is met by the specified deadline.

- One possible approach would be to **experiment with different versions** of the homepage and test whether we can encourage a greater number people to click on the link to view the weekend closures page.

Definition

A/B testing is the process of analysing the performance of two different versions of a web page, where pages differ with respect to a single element, so as to decide upon the best version.

Setting up the test…

- We will **randomly assign** users to one of two different versions of the homepage. This is to ensure the experiment is **fair**.

- Page A is the original and Page B has been **modified** to make the link more prominent.

Obtaining the results…

Suppose we implemented these changes and recorded the following results (drawn in a contingency table);

Actual	Page	Clicked	Didn't Click	
	A	37	33	70
	B	52	18	70
		89	51	140

Question:

How many visitors took part in the experiment?

Which page tends to encourage more visitors to click on the link?

Calculating the expected results…

But we expected the following results;

Actual	Page	Clicked	Didn't Click	
	A	37	33	70
	B	52	18	70
		89	51	140
Expected	Page	Clicked	Didn't Click	
		=(C20/E20)*E18		70
	B	44.5	25.5	70
		89	51	140

The expected results…

But we expected the following results;

Expected	Page	Clicked	Didn't Click	
	A	44.5	25.5	70
	B	44.5	25.5	70
		89	51	140

Statistical significance

- Although **page B** looks to be **more optimal** by the no of clicks this result could have occurred simply by **random chance**
- e.g., It's quite possible to throw a coin and get 10 heads in a row
- But if it happened a 1000 times, we would begin to get **suspicious** as to whether the coin was in fact **fair**
- Several statistical tests exist to help us establish whether a particular set of results are **significant** or not
- Hence, we can test the **probability** of these results occurring by simple random chance alone

Fishers Exact Test

- In this case, and because the expected frequencies are **relatively small (<100 in each group)**, the appropriate statistical test is the two-tailed Fisher's Exact Test

Definition

Fishers Exact Test tests the significance of the association between two different categorical variables. The null hypothesis is that there is no relationship between the variables i.e. the page we show the user doesn't affect whether they click on the link or not.

- It's a **two tailed** test because we care whether there is a positive or negative influence
- At a very high level, it computes the **probability distribution** of the observed results relative to the expected ones

Interpreting the results…

We can calculate it in a number of ways, including in SPSS, but here I have used GraphPad ;

	Clicked	Didn't Click	Total
Page A	37	33	70
Page B	52	18	70
Total	89	51	140

Fisher's exact test

The two-tailed P value equals 0.0136
The association between rows (groups) and columns (outcomes) is considered to be statistically significant.

Question:
Is the result significant?
How do we interpret the results in the context of the original problem?

Choosing the appropriate test…

- **Categorical->Categorical**
  - *Does the PAGE increase the likelihood of an event occurring?*
  - Fishers Exact Test (for smaller samples)
  - Chi-Squared Test (for larger samples)

- **Categorical->Numerical**
  - *Does the PAGE increase the amount of time or amount of money spent?*
  - Wilcoxon Mann-Whitney Test (for smaller samples)
  - Two samples T-Test (for larger samples)

- **Categorical->Numerical (2 or more groups)**
  - *Does the PAGE cause the amount of time, amount of money spent, conversion rate or bounce rate across multiple groups to vary?*
  - *ANOVA (analysis of variance)*
  - *Compares the variation between groups to the variance within them*

Example ANOVA

- Lets consider an example whereby the site owner wishes to increase **time spent on a page** (engagement)
- The site owner carried out an experiment and obtained the following results.

Page Version	Time Spent on page (seconds)	Average time
A	56	
A	12	
A	33	
A	2	
A	64	33.4
B	14	
B	65	
B	24	
B	32	
B	7	28.4
C	45	
C	5	
C	21	
C	3	
C	43	23.4

Inputting this data into R…

```
> pageA = c(56, 12, 33, 2, 64)
> pageB = c(14, 65, 24, 32, 7)
> pageC = c(45, 5, 21, 3, 43)
>
```

Question:
What data type will be used to store the variable “pageB”?

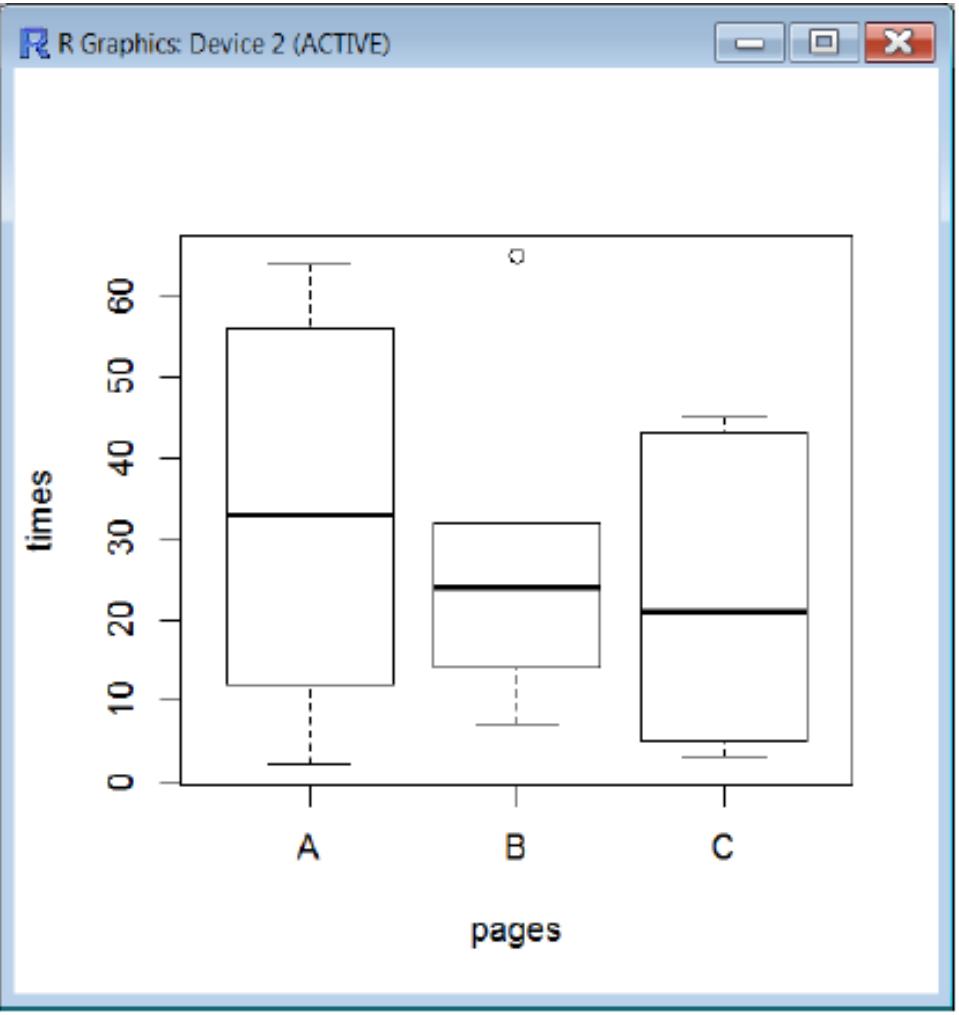
Placing the data into a data frame…

```
> times = c(pageA, pageB, pageC)
> pages = c(rep("A", 5), rep("B", 5), rep("C", 5))
> pages
 [1] "A" "A" "A" "A" "A" "B" "B" "B" "B" "B" "C" "C" "C" "C" "C"
> times
 [1] 56 12 33  2 64 14 65 24 32  7 45  5 21  3 43
> page_time = data.frame(times, pages)
> page_time
   times pages
1     56     A
2     12     A
3     33     A
4      2     A
5     64     A
6     14     B
7     65     B
8     24     B
9     32     B
10     7     B
11    45     C
12     5     C
13    21     C
14     3     C
15    43     C
>
```

Creating a Box Plot…

R Console

```
> plot(times ~ pages, data=page_time)
>
```



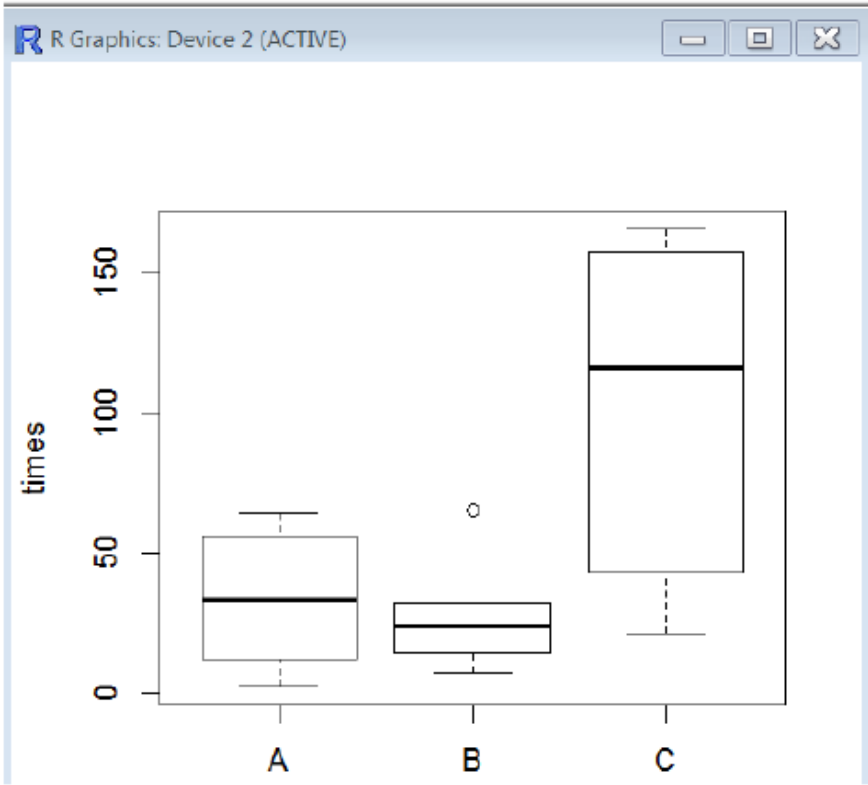
Conducting the ANOVA…

Can we reject the null hypothesis?

```
>
>
> plot(times ~ pages, data=page_time)
> results = aov(times ~ pages, data=page_time)
> summary(results)
             Df Sum Sq Mean Sq F value Pr(>F)
pages         2    250     125   0.229  0.798
Residuals    12   6540     545
> |
```

P-value = 0.798

What if the time spent on page C was actually this?



R Console

```
>
>
>
>
>
>
>
>
>
>
>
>
>
> pageC = c(165, 157, 21,116, 43)
> times = c(pageA, pageB, pageC)
> page_time = data.frame(times, pages)
> plot(times ~ pages, data=page_time)
> results = aov(times ~ pages, data=page_time)
> summary(results)
            Df Sum Sq Mean Sq F value Pr(>F)
pages        2  16163    8082   4.379 0.0373 *
Residuals   12  22148    1846
```

P-value = 0.0373

Tukey's Honest Significant Difference

- Recall that ANOVA cannot tell us which group's means are significantly different to one another, only that at least **TWO** do
- **Tukey's HSD** test compares all possible means with one another to test the extent to which they could have come from the same underlying distribution.
- We study the reported p-values to check whether the **difference in means** of any **TWO** groups are significant or not.
- When we carry out this test, we specify a **confidence level**, typically 95%, which controls how much we allow the means to differ before it becomes significant
- A higher confidence level corresponds with a **stricter** test

Conducting Tukey's HSD in R…

Which groups differ significantly?

```
> TukeyHSD(results, conf.level = 0.95)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = times ~ pages, data = page_time)

$pages
    diff          lwr       upr     p adj
B-A   -5 -77.4880369  67.48804 0.9815331
C-A   67  -5.4880369 139.48804 0.0710902
C-B   72  -0.4880369 144.48804 0.0516029
```

In Summary

- Today, we have reviewed important distinctions between different data collection methods along their advantages and disadvantages.
- We have highlighted some of the most prominent web analytics tools and key areas of differentiation.
- We explored the role of clickstream data, how it can be collected through the use of specially crafted URI's which pass encoded data between the client and the web server.
- Funnels and Heatmaps are two possible approaches to visualise clickstream data and can be used to support digital marketing campaigns.
- The practice of web analytics should be though of as a process, involving the development of a strong understanding of an organisations long term goals and objectives.
- KPIs (Key Performance Indicators) can be used to formulate and steer actions in the shorter term.
- A/B Testing is a useful tool in helping to conduct experiments to statistically test whether our actions are having the desired impact.

Learning week 3

- In LW3 we shift our attention on **Search Engines**, the largest and most well-known examples of Information Retrieval (IR) Systems.
- We will study their importance on the web and for the practice of web analytics.

- **A reminder to**
  - Complete the activities in the weekly tutorial packs
  - There is an extension exercise (optional) using a larger web log file

End