

# Outline

- Definition
- Examples
- Theory and Essential Construct
- Data Collection
- Method
- Discussion

# What is Web log analysis?

Jim Jan...  
College of Informa... Technology  
The P... University  
[jan@acm.org](mailto:jan@acm.org)

Let's make this a discussion!

# Web log analysis is part of the domain of ...

- ... Web analytics
- The Web Analytics Association (WAA) defines **Web analytics** as *the measurement, collection, analysis, and reporting of Internet data for the purposes of understanding and optimizing Web usage* (<http://www.webanalyticsassociation.org/>)
- Shares common **theoretical** and **methodology** characteristics with all forms of log analysis (e.g., Intranet logs, systems logs, OPAC logs, search logs, etc.)



```
W3C Extended Log Format
File Edit Format View Help
#Software: Microsoft Internet Information Services 5.1
#Version: 1.0
#Date: 2002-08-12 00:23:05
#Fields: time c-ip cs-username s-ip s-port cs-method cs-uri
cs(User-Agent)
00:23:05 127.0.0.1 - 127.0.0.1 80 GET /iisstart.asp 302 0
(compatible;+MSIE+6.0;+windows+NT+5.1;+.NET+CLR+1.0.3705)
00:23:05 127.0.0.1 - 127.0.0.1 80 GET /localstart.asp 401
(compatible;+MSIE+6.0;+windows+NT+5.1;+.NET+CLR+1.0.3705)
00:23:05 127.0.0.1 BL-UIITS-OSIRIS\jcausey 127.0.0.1 80 GET
(compatible;+MSIE+6.0;+windows+NT+5.1;+.NET+CLR+1.0.3705)
00:23:06 127.0.0.1 BL-UIITS-OSIRIS\jcausey 127.0.0.1 80 GET
Mozilla/4.0+(compatible;+MSIE+6.0;+windows+NT+5.1;+.NET+CL
00:23:06 127.0.0.1 - 127.0.0.1 80 GET /winxp.gif 200 0 Moz
(compatible;+MSIE+6.0;+windows+NT+5.1;+.NET+CLR+1.0.3705)
00:23:06 127.0.0.1 BL-UIITS-OSIRIS\jcausey 127.0.0.1 80 GET
(compatible;+MSIE+6.0;+windows+NT+5.1;+.NET+CLR+1.0.3705)
00:23:06 127.0.0.1 BL-UIITS-OSIRIS\jcausey 127.0.0.1 80 GET
(compatible;+MSIE+6.0;+windows+NT+5.1;+.NET+CLR+1.0.3705)
00:23:06 127.0.0.1 - 127.0.0.1 80 GET /mmc.gif 200 0 Mozil
(compatible;+MSIE+6.0;+windows+NT+5.1;+.NET+CLR+1.0.3705)
00:23:06 127.0.0.1 - 127.0.0.1 80 GET /help.gif 200 0 Mozil
(compatible;+MSIE+6.0;+windows+NT+5.1;+.NET+CLR+1.0.3705)
00:23:07 127.0.0.1 - 127.0.0.1 80 GET /print.gif 200 0 Moz
(compatible;+MSIE+6.0;+windows+NT+5.1;+.NET+CLR+1.0.3705)
00:23:07 127.0.0.1 BL-UIITS-OSIRIS\jcausey 127.0.0.1 80 GET /iis
help/iis/misc/default.asp 200 0
Mozilla/4.0+(compatible;+MSIE+6.0;+windows+NT+5.1;+.NET+CLR+1.0.3705)
00:23:07 127.0.0.1 - 127.0.0.1 80 GET /iis/help/iis/misc/navbar.asp 200 0 Mozilla/4.0+
(compatible;+MSIE+6.0;+windows+NT+5.1;+.NET+CLR+1.0.3705)
00:23:07 127.0.0.1 BL-UIITS-OSIRIS\jcausey 127.0.0.1 80 GET /iis
help/iis/misc/contents.asp 200 0
Mozilla/4.0+(compatible;+MSIE+6.0;+windows+NT+5.1;+.NET+CLR+1.0.3705)
00:23:07 127.0.0.1 - 127.0.0.1 80 GET /iis/help/iis/htm/core/iitop.htm 200 0 Mozilla/4.0+
(compatible;+MSIE+6.0;+windows+NT+5.1;+.NET+CLR+1.0.3705)
00:23:07 127.0.0.1 BL-UIITS-OSIRIS\jcausey 127.0.0.1 80 GET /iis
help/iis/misc/ismhd.gif 200 0
Mozilla/4.0+(compatible;+MSIE+6.0;+windows+NT+5.1;+.NET+CLR+1.0.3705)
00:23:07 127.0.0.1 - 127.0.0.1 80 GET /iis/help/iis/misc/navpad.gif 200 0 Mozilla/4.0+
(compatible;+MSIE+6.0;+windows+NT+5.1;+.NET+CLR+1.0.3705)
00:23:07 127.0.0.1 - 127.0.0.1 80 GET /iis/help/iis/misc/MS_logo.gif 200 0 Mozilla/4.0+
```

W3C Extended Log Format - Variety of fields for examining visitors to Web sites.

Other common format is **NCSA Separate Log** that is composed of three logs ( **Common log** – actions on the server, **Referral log** – where they came from, and **Agent log** – stuff about the client computer)



## Theoretical Foundations

- Part of the **behaviorism paradigm**
- **Behaviorism** – an approach focused on the outward **behavioral aspects** of thought and emphasizes the **observed behaviors**
- Behaviorism – Pavlov, Watson, & Skinner



Ivan Petrovich Pavlov



John B. Watson



Burrhus Frederic Skinner

## Behaviorism Characteristics

- **Inductive, data-driven** and characterized by **empirical** observation of measurable behavior
- Grounded on **somebody** doing **something** in a **situation** (*all* the environmental and situational features are embedded behaviors)
- **Critics** of behaviorism as a psychological theory have issues with **rejection of mental processes**. **I agree** - people are more than “**mediators between behavior and the environment**” (Skinner, 1993, p 428)

## What is a Behavior?

... an **observable activity** of a person, animal, team, organization, or system.

One can classify **behaviors** into three general categories. Behaviors are

- something that one can **detect** and **record**
- **actions** or specific goal-driven **events** with some purpose other than the specific action that is observable
- **reactive responses** to environmental stimuli

## What is a Behavior?

- Behavior is the **essential construct** of the behaviorism and of **log research**
- Logs record **behaviors** of users and systems (records behavior but can't tell **affective, cognitive, or situational** aspects)
- A behavior is the key **variable** (i.e., an **entity** representing a **set of events** where each event may have a **different value**)



# Ethograms

- a **taxonomy** or index of behavioral patterns
- details the **different forms** of behavior that an user exhibits
- categories of behavior are **objective, discrete, not overlapping**. This makes the definitions of each behavior (and category of behaviors) clear, detailed and distinguishable from each other

Example of an Ethogram	
Behavior	Description
<b>View results</b>	Interaction in which the user viewed or scrolled one or more pages from the results listing. If a results page was present and the user did not scroll, we counted this as a View Results Page.
<i>With Scrolling</i>	<i>User scrolled the results page.</i>
<i>Without Scrolling</i>	<i>User did not scroll the results page.</i>
<i>but No Results in Window</i>	<i>User was looking for results, but there were no results in the listing.</i>
<b>Selection</b>	Interaction in which the user makes a selection in the results listing.
<i>Click URL (in results listing)</i>	<i>Interaction in which the user clicked on a URL of one of the results in the results page.</i>
<i>Next in Set of Results List</i>	<i>User moved to the Next results page.</i>
<i>Previous in Set of Results List</i>	<i>User moved to the Previous results page.</i>
<i>GoTo in Set of Results List</i>	<i>User selected a specific results page.</i>
<b>View document</b>	Interaction in which the user viewed or scrolled a particular document in the results listings.
<i>With Scrolling</i>	<i>User scrolled the document.</i>
<i>Without Scrolling</i>	<i>User did not scroll the document.</i>

Behavior

Description of the behavior

What about the data collection method?

## Data Collection: Trace Data

- **Computer storage media** collected in log files as **trace data**
- **Physical remains** such as the **wear on a carpet** induce wear, or **reduce** the **physical remains**.
- **Computer storage media** are the **physical remains**.



Wear on a carpet



Trash heap



Computer storage media

## Trace Data

- In the past, trace data was often **time consuming** to gather and process, making such data costly.
- **logging software** makes collecting trace data **easy and cheap**
- Log data is **controlled accretion data**, where the researcher or some other entity alters the environment in order to create the accretion data
- With the user of client apps (such as desktop search bars), the **collection of data is nearly unlimited** from a technology perspective

What is **cool** about **trace data** for researchers?

## Data Collection

Log data has **significant advantages** as a data collection approach for the study and investigation of behaviors, including:

- **Scale**: not a limiting factor as in lab user studies
- **Power**: large sample size for inference testing; in fact, so large must account for the size effect
- **Scope**: naturalistic; researchers can investigate range of interactions in a multi-variable context
- **Location**: can collect in distributed environments
- **Duration**: collect log data over an extended period

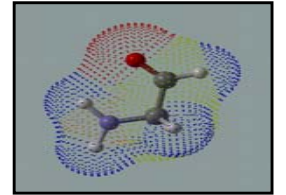
## Methodological Foundations

Use of **logs** to collect **trace data** is an unobtrusive methods (a.k.a., non-reactive or low-constraint). **Unobtrusive methods** ...

- allows data collection **without directly** interfering into the context and
- does **not require a direct response** from participants



Customer Behavior (video)



Chemistry (surface marking)

## Methodological Foundations

Three **justifications** for unobtrusive methods:

- Example: ethnography studies (where the researcher “bird dogs” a study participant)
- Example: no one searches for porn in a lab study of Web searching
- Example: is why medical trials are double blind rather than single blind

Trace data helps in **overcoming** the **Uncertainty principle**, **Observer effect**, and **Observer bias** in the data collection. Note for **data collection** but **not data analysis**

## Methodological Foundations

Inherent **characteristics** in the method of log data collection; Web analytics has issues to address as a result:

- **Abstraction** – how does one relate low-level data to higher-level concepts?
- **Selection** – how does one separate the necessary from unnecessary data?
- **Reduction** – how does one reduce the complexity and size of the data set?
- **Context** – how does one interpret the significance of events?
- **Evolution** – how can one collect data without impacting application deployment or use?

## Recap of Web Analytics

Type of Data	Trace	
--------------	-------	---

## Research

- Book: Jansen, B. J., Spink, A., and Taksa, I. (2009) Handbook of Research on Web Log Analysis, Hershey, PA: Idea Group Publishing
  - First chapter on theory of log analysis is free!
- Lecture: Jansen, B. J. (Forthcoming) *Understanding User – Web Interactions via Web Analytics*. Morgan-Claypool Lecture Series. Gary. Marchionini (Ed). Morgan-Claypool: San Rafael, CA.
  - manuscript about Web Analytics, soup to nuts

Thank you!  
(open for questions and further discussion)

**Jim Jansen**  
College of Information Sciences and Technology  
The Pennsylvania State University  
[jjansen@acm.org](mailto:jjansen@acm.org)