



SUGI27

12.17.01 SAS Group International
Orlando, Florida - April 14-17, 2002

Mining Web Server Logs: Tracking Users and Building Sessions

Don Koch
John C. Brocklebank, Ph.D.
Tom Grant
Richard Roach

Analytic ASP Development

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



Introduction

Web data mining relies heavily on accurate session re-construction for the analysis of user click-stream behavior.

This paper illustrates one particular problem with re-constructing user sessions from web server log data. We will discuss session head loss, its effects and how to prevent it.

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



Web Logs Are Bad

- They don't identify sessions or users
- Critical events (e.g. add to cart) are missing
- Form data is not revealed
- URL lack semantic information
- URL is not important anymore
- Not in a relational database
- Information is redundant
- They don't contain additional business data
- Cached data may not be logged

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



Web Logs Are Good

- Every web server uses them
- Most if not all of the bad can easily be circumvented
- Mine a site AS-IS
- Intrusive site instrumentation can be avoided
- Work well with application logging
- Work well with site instrumentation

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



Doing a Site Overhaul?

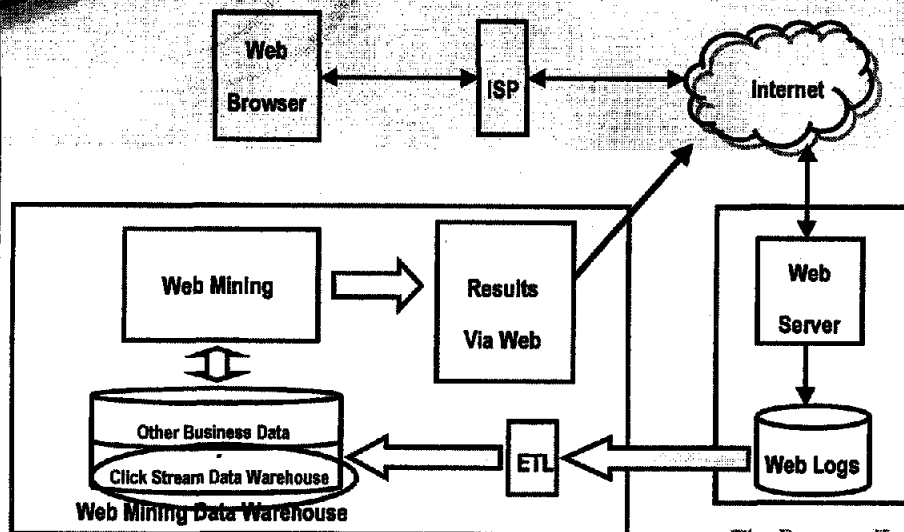
Don't forget the analysts!

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



Architecture Overview



The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.

What is Click-Stream Data?

- All web servers are capable of logging requests
- Each browser request is recorded in a web log
- A single "click" may result in many requests
- Not all requests are logged
- Not a perfect data source

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.

Web Mine Data Warehouse Click-Stream Data Sources

Transaction logs

- Web page servers
- Caching servers
- Secure servers
- Application servers
- Media servers

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



Web Mine Data Warehouse Business Data Sources

- Sales Data
- Add Campaigns
- Product Catalogs
- Customer Information
- Partners

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



What is a Session

The click stream for a single user during an arbitrary period of time.

- Try to capture user activity in one "sitting"
- Typical session time out is 30 minutes.
- Some application servers enforce a session time-out.
- Need Session Identification

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.

Fields of a Typical Web Log

Field	Example
Date Time	2002-03-26 00:00:01
Client IP	64.216.85.240
Method	GET POST HEAD
URI Stem	/default.asp
URI Query	DeptID=20078&CatID=20081
Cookie	SITESERVER=ID=df5f9ad4b03441df1678
User Name	Don
Referrer	http://www.google.com
User Agent	Mozilla/4.0+(compatible;+MSIE+5.12;+Mac_P
Status	200
Bytes	15643

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.

Categorization of Fields

Field	Relationship	Presence	Lifetime
<i>IP Address</i>	Many to Many	Always	None
<i>User Agent</i>	One to Many	Usually	Browser
<i>Time Stamp</i>	Many to Many	Always	None
<i>Cookie</i>	One to One	Second Hit	Persistent
<i>Query String</i>	One to One	Second Hit	Session

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



Relationship Many to Many

- A browser may have more than one value during a session.
- A specific value may be used for more than one browser.
- Example: IP Address

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



Relationship One to Many

- A browser will have only one value during a session.
- A specific value may be used for more than one browser.
- Example: User Agent

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



Relationship One to One

- A browser will have only one value during a session.
- A specific value may be used for only one browser.
- Example: Server Cookie (GUID)

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



Presence

- Always
 - There is always a value (e.g. IP Address)
- Usually
 - There is usually a value (e.g. User Agent)
- Second Hit
 - Blank first visit, first request (e.g. Cookie)

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.

Lifetime

- **Session**
 - Value is maintained entire session (e.g. Query String)
- **Persistent**
 - Value maintained across sessions (e.g. Cookie)
- **Browser**
 - Value maintained for life of browser (e.g. User Agent)

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.

Categorization of Fields

Field	Relationship	Presence	Lifetime
<i>IP Address</i>	Many to Many	Always	None
<i>User Agent</i>	One to Many	Usually	Browser
<i>Time Stamp</i>	Many to Many	Always	None
<i>Cookie</i>	One to One	Second Hit	Persistent
<i>Query String</i>	One to One	Second Hit	Session

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



What is a Server Cookie?

An anonymous cookie generated by the web server software used for session tracking.

- Guaranteed to be unique.
- Can be persistent (best for repeat visitors).
- Usually used for session tracking.
- IIS uses an ASAPI filter to set SITESERVER=ID.
- Apache use mod_usertrack to set Apache cookie.

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



What is a Browser Cookie

An application-specific cookie generated by the browser using a client-side scripting language.

- Depends on application.
- Can be set anytime during the session
- Requires maintenance
- Usually used for application state

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



How is URL Re-Writing Used?

A session ID can be attached to the query string of every request.

- GET /basket.html?sessionID=2434
- Only for dynamically-generated pages.
- Used by application servers ASP, JSP.
- Can't be turned off as cookies can.
- Often used in conjunction with cookies.

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



Understanding Your Site

Understanding how your site works is THE most important task for the analyst.

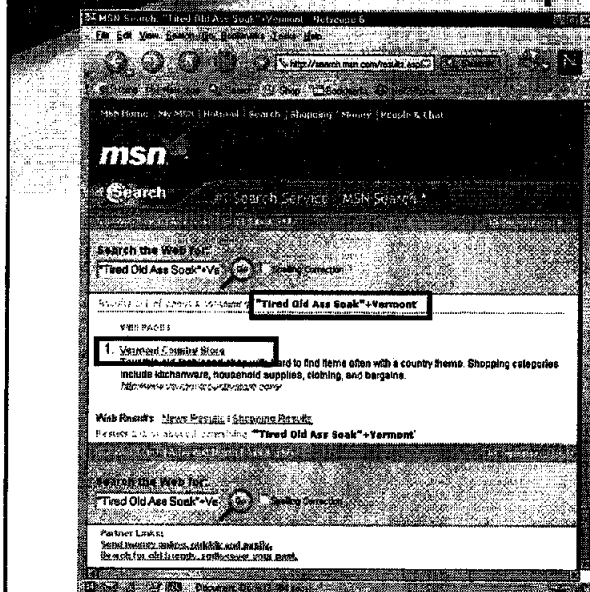
It's not always easy.

- Can't prevent browser from following redirects.
- Don't always know sequence of cookie sets.
- You may only have web logs to look at.
- Always have access to public site.

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.

A Sample Session



One result returned:
Vermont Country Store

The Power to Know.

A Sample Session



New Session Started
Vermont Country Store

The Power to Know.



A Sample Session



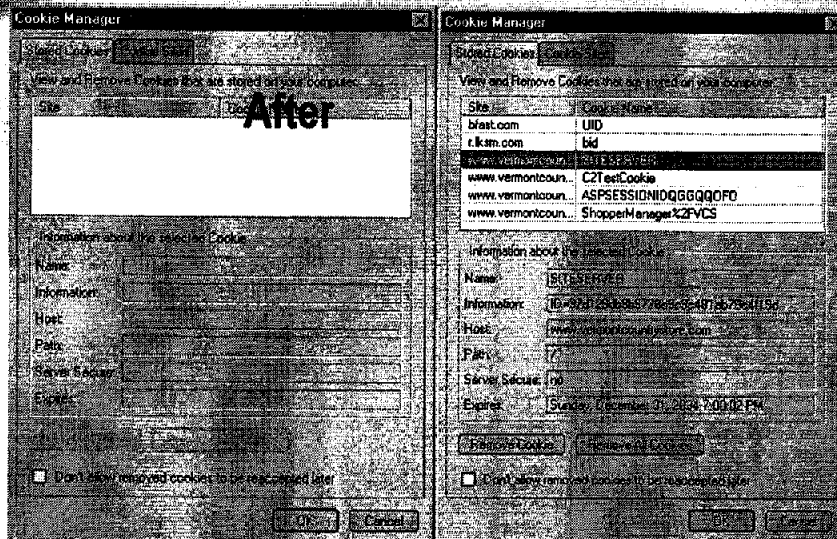
Second Session "Click"
Many Requests Made

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



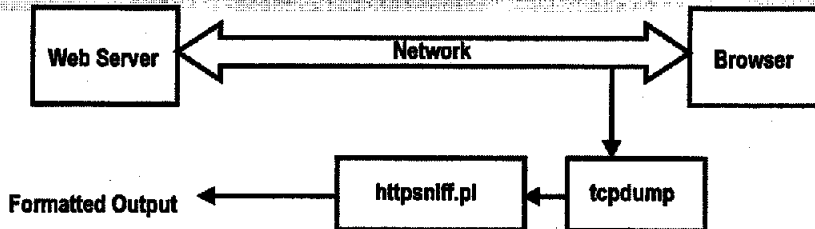
Cookies



Copyright © 2001, SAS Institute Inc. All rights reserved.

What's Going on Between Browser and Web Server?

Listening to Browser and Server Conversations



```

TIME=16:47:14.127027 TO=10.40.12.75.1338 FROM=10.16.149.76.80
=====
GET http://search.msn.com/results.asp?q=%22Tired+Old+Ass+Soak%22%2BVermon
x=0&y=0&origq=&RS=CHECKED&FORM=SHORT&v=1&cfq=SHCINITIAL&ncap=the
Host: search.msn.com
User-Agent: Mozilla/5.0 (Windows; U; WinNT4.0; en-US; rv:0.9.4.1) Gecko/20030314 Netscape/6.2.2
Referer: http://search.msn.com/
  
```

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.

Formatted Output Browser Request

```

TIME=16:47:14.127027 TO=10.40.12.75.1338 FROM=10.16.149.76.80
=====
GET http://search.msn.com/results.asp?q=%22Tired+Old+Ass+Soak%22%2BVermon
Host: search.msn.com
User-Agent: Mozilla/5.0 (Windows; U; WinNT4.0; en-US; rv:0.9.4.1) Gecko/2
Accept: text/xml, application/xml, application/xhtml+xml, text/html;q=0.9
Accept-Language: en-us
Accept-Encoding: gzip, deflate, compress;q=0.9
Accept-Charset: ISO-8859-1, utf-8;q=0.66, *;q=0.66
Keep-Alive: 300
Proxy-Connection: keep-alive
Cookie: lang=en-us; MC1=V=2&GUID=99a45246370f43d095f97c7f53491faf; mh=MSR
Referer: http://search.msn.com/
  
```

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



Formatted Output Server Response

TIME=16:47:52.499488 TO=10.16.149.76.80 FROM=10.40.12.75.1335

HTTP/1.0 200 OK

Server: Microsoft-IIS/5.0

Date: Wed, 10 Apr 2002 20:47:51 GMT

Content-Length: 39392

Content-Type: text/html

Set-Cookie: ShopperManager=J8B7MR8844A68KCWA926CXMV3KH47HAC;
expires=Fri, 01-Jan-2010 05:00:00 GMT; path=/

Cache-Control: private

X-Cache: MISS from inetgw01.unx.sas.com

X-Cache-Lookup: MISS from inetgw01.unx.sas.com:80

Proxy-Connection: keep-alive

Web Content Goes Here

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



What is Session De-Heading?

The first one or more clicks of a user's session cannot be associated with the rest of the user's session.

Only effects first-time visitors.

Can have a dramatic impact on click-stream analysis.

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.

Session Without Re-Direct

Browser

Web Server

GET http://www.vcs.com/
Referer: http://search.msn.com/...

HTTP/1.0 200 OK
Set-Cookie: SITESERVER-ID=39f...
<content>

GET http://www.vcs.com/css/style.asp
Cookie: SITESERVER-ID=39f...
Referer: http://www.vcs.com/

HTTP/1.0 200 OK
<content>

GET http://www.vcs.com/search.asp
Cookie: SITESERVER-ID=39f...
Referer: http://www.vcs.com/

HTTP/1.0 200 OK
<content>

Rest of user session

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.

Web Log from Session

cs-method/cs-uri-stem cs(Cookie) cs(Referer)

GET / - http://search.msn.com/

GET /css/style.asp SITESERVER-ID=39f... http://www.vcs.com/

GET /search.asp SITESERVER-ID=39f... http://www.vcs.com/

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



A Real Session Log

```
00:00:06 GET /navDefault.asp 302 ←
sourceid=00388381575067034674
http://myconsumerguide.com/vermont_country_store.htm

00:00:06 GET /no_cookies.asp 302 ←
referer=/navDefault.asp?sourceid=00388381575067034674&showWarning=False
SITESEVER=ID=2c5a8dc836b5ee0298ab2e40f2037b8c
http://myconsumerguide.com/vermont_country_store.htm

00:00:07 GET /navDefault.asp 200 ←
sourceid=00388381575067034674&flgCookiesEnabled=TRUE
SITESEVER=ID=2c5a8dc836b5ee0298ab2e40f2037b8c
http://myconsumerguide.com/vermont_country_store.htm

00:00:07 GET /include/stylesheets/style.asp 200 ←
SITESEVER=ID=2c5a8dc836b5ee0298ab2e40f2037b8c
http://www.vermontcountrystore.com/navDefault.asp?sourceid=00388381575067
```

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



Why is Session De-Heading a Problem?

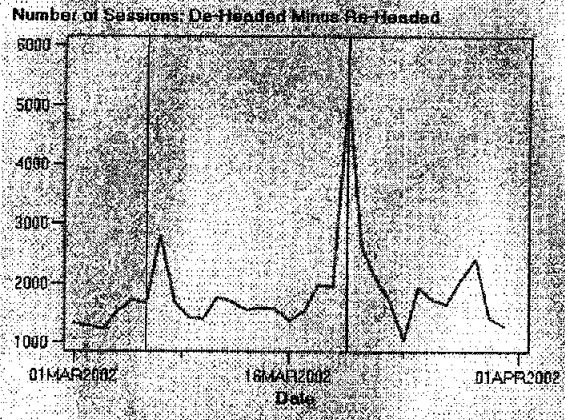
Referrer information can be lost to the session.
Exacerbated by campaigns designed to attract new users.

Over-estimate number of sessions. Any statistic based on number of sessions will be under-estimated.

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.

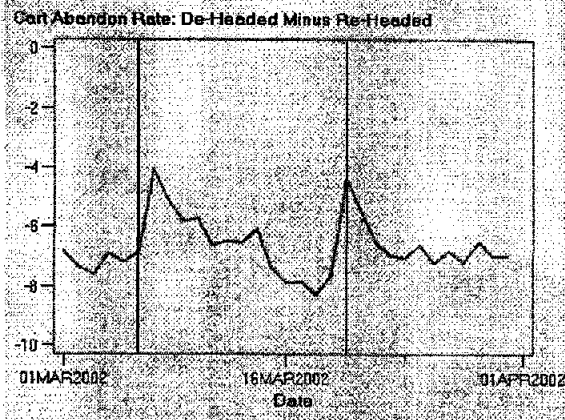
Session Counts Are Higher



The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.

Statistics are Lower Rate of Cart Abandonment

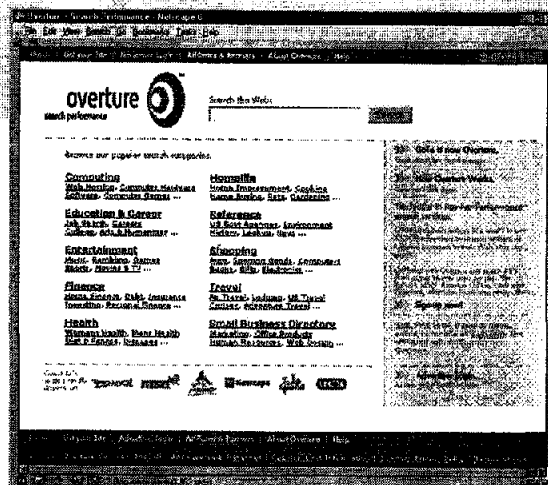


The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



Case Study Calculating Search Engine ROI



The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



Case Study Search Engine ROI

An on-line retailer spends \$12,523.25 with Overture for keyword placement.

Their goal is to acquire new customers who will spend money.

The question is, was it worth it?

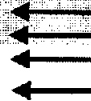
The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



Acquisition De-Headed Sessions

<u>Acquisition</u>	<u>De-Headed Sessions</u>
Advertising Expense	\$12,540
Click-Through	4,236
Buyers	
Initial Sales	\$363
Initial Profit on Sales	\$127
6-Month Follow-up Campaign	
<u>De-Headed Sessions</u>	
Advertising Expense	\$17
Multi-Buyers	3
Sales	\$182
Profit on Sales	\$63
Summary	
<u>De-Headed Sessions</u>	
Total Profit on Sales	\$190
Total Advertising Expense	\$12,540
Lifetime Value	(\$12,349)



The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



Follow-up Campaign De-Headed Sessions

<u>Acquisition</u>	<u>De-Headed Sessions</u>
Advertising Expense	\$12,523
Click-Through	4,236
Buyers	7
Initial Sales	\$363
Initial Profit on Sales	\$127
6-Month Follow-up Campaign	
<u>De-Headed Sessions</u>	
Advertising Expense	\$17
Multi-Buyers	3
Sales	\$182
Profit on Sales	\$63
Summary	
<u>De-Headed Sessions</u>	
Total Profit on Sales	\$190
Total Advertising Expense	\$12,540
Lifetime Value	(\$12,349)



The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



Summary De-Headed Sessions

<u>Acquisition</u>	<u>De-Headed Sessions</u>
Advertising Expense	\$12,523
Click-Through	4,236
Buyers	7
Initial Sales	\$363
Initial Profit on Sales	\$127
<u>6-Month Follow-up Campaign</u>	
Advertising Expense	\$17
Multi-Buyers	3
Sales	\$182
Profit on Sales	\$63
<u>Summary</u>	
Total Profit on Sales	\$190
Total Advertising Expense	\$12,540
Lifetime Value	(\$12,349)



The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



Acquisition Re-Headed Sessions

<u>Acquisition</u>	<u>De-Headed Sessions</u>	<u>Re-Headed Sessions</u>
Advertising Expense	\$12,523	\$12,523
Click-Thru	4,236	4,236
Buyers	7	452
Initial Sales	\$363	\$23,432
Initial Profit on Sales	\$127	\$8,201
<u>6-Month Follow-up Campaign</u>		
Advertising Expense	\$17	\$1,130
Multi-Buyers	3	293
Sales	\$182	\$35,453
Profit on Sales	\$63	\$12,408
<u>Summary</u>		
Total Profit on Sales	\$190	\$20,610
Total Advertising Expense	\$12,540	\$13,653
Lifetime Value	(\$12,349)	\$6,956



The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.

Follow-up Campaign Re-Headed Sessions

<u>Acquisition</u>	<u>De-Headed Sessions</u>	<u>Re-Headed Sessions</u>
Advertising Expense	\$12,523	\$12,523
Click-Through	4,236	4,236
Buyers	7	452
Initial Sales	\$363	\$23,432
Initial Profit on Sales	\$127	\$8,201
6-Month Follow-up Campaign		
Advertising Expense	\$17	\$1,130
Multi-Buyers	3	293
Sales	\$182	\$35,453
Profit on Sales	\$63	\$12,408
Summary		
Total Profit on Sales	\$190	\$20,610
Total Advertising Expense	\$12,540	\$13,653
Lifetime Value	(\$12,349)	\$6,956



The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.

Summary Re-Headed Sessions

<u>Acquisition</u>	<u>De-Headed Sessions</u>	<u>Re-Headed Sessions</u>
Advertising Expense	\$12,523	\$12,523
Click-Through	4,236	4,236
Buyers	7	452
Initial Sales	\$363	\$23,432
Initial Profit on Sales	\$127	\$8,201
6-Month Follow-up Campaign		
Advertising Expense	\$17	\$1,130
Multi-Buyers	3	293
Sales	\$182	\$35,453
Profit on Sales	\$63	\$12,408
Summary		
Total Profit on Sales	\$190	\$20,610
Total Advertising Expense	\$12,540	\$13,653
Lifetime Value	(\$12,349)	\$6,956



The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



Was it Worth It?

Summary	De-Headed Sessions	Re-Headed Sessions
Value per Visitor	(\$7.87)	\$1.64
Value per Buyer	(\$1,764)	\$15.39
Value per Multi-Buyer	(\$4,116)	\$23.74

Yes!

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



SAS/WebHound and Re-Heading

- Currently in Beta. Will be included in next release.
- Adds time to processing.
- Highly tuned for efficiency.
- Lots of parameters to control session building
 - MAXIMUM_SECONDS_BETWEEN_NOIDS
 - NUMBER_NOIDS_TO_REHEAD
 - MINUTES_ALLOWED_FOR_UNIQUE

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



Conclusion

Costly site redesigns or retrofits are not necessary to successfully mine click stream data.

Web logs are a goldmine if you ...

- Understand session de-heading and re-heading
- Understand how your site actually works

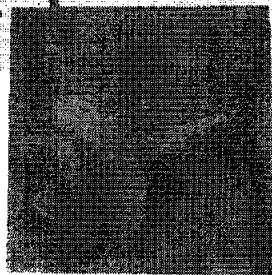
**If you are going to overhaul,
Don't forget the analysts!**

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.



SUGI27



**SAS® Users Group International
Orlando, Florida • April 14–17, 2002**

The Power to Know.

Copyright © 2001, SAS Institute Inc. All rights reserved.