IIT KHARAGPUR

NPTEL ONLINE
CERTIFICATION COURSES



# E-BUSINESS



**PROF. MAMATA JENAMANI**
DEPARTMENT OF INDUSTRIAL AND SYSTEMS ENGINEERING
IIT KHARAGPUR

Week 9: Lecture1



# DECISION SUPPORT CONCEPTS

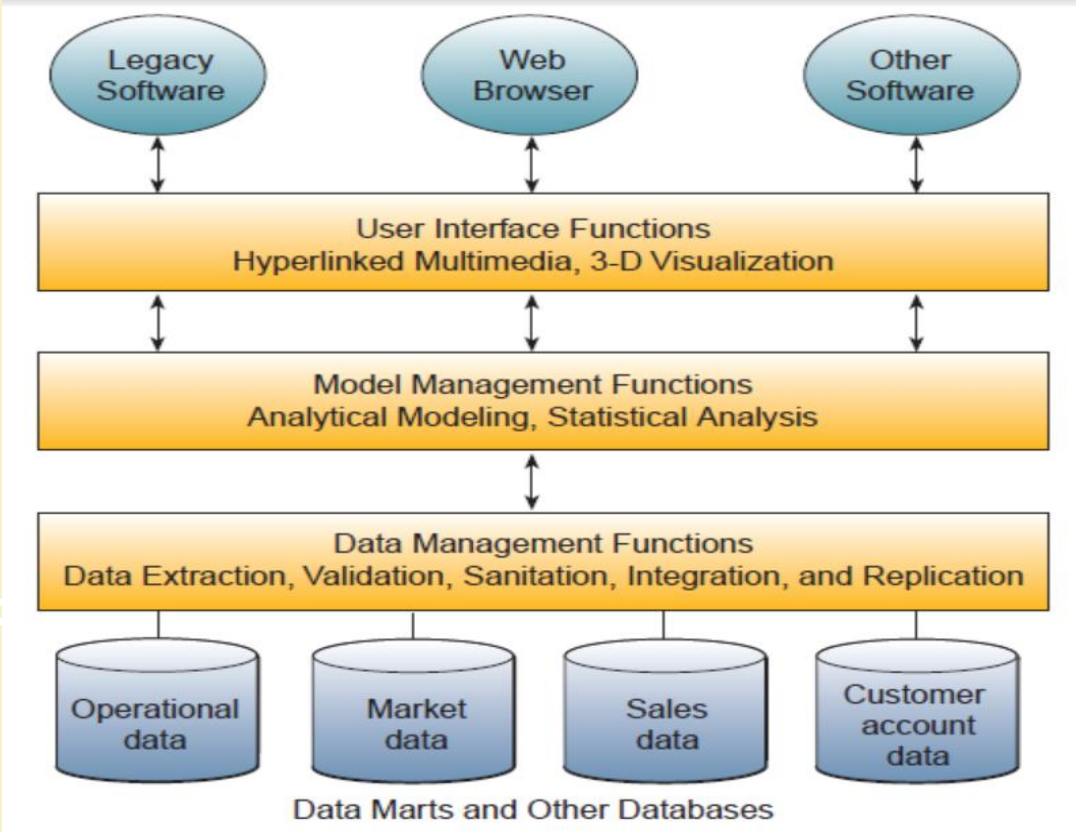
# We are going to learn

- Concepts related to decision support
- Applications

# Decision support systems

- Provide interactive ad hoc support for the decision-making processes of managers and other business professionals.
- Support nonroutine decision making
  - Example: What is impact on production schedule if December sales doubled?
- Model driven DSS,
- Data driven DSS
- Serve middle management
- Examples: product pricing, profitability forecasting, and risk analysis systems.

# Components of a typical DSS



# Analytical Models

- Optimization
- Simulation
- Decision analysis
- Static vs. dynamic models
- Deterministic Vs. Stochastic models

# Statistical Models

- Descriptive statistics
- Outlier analysis
- Univariate predictive models
- Multi variate predictive models

# Data mining Models

- Predictive:
  - Regression
  - Classification
  - Collaborative Filtering
- Descriptive:
  - Clustering / similarity matching
  - Association rules and variants

# Text mining

- Natural Language processing
- Discover patterns

| Term | Time frame | Specific meaning |
|---|---|---|
| Decision support | 1970–1985 | Use of data analysis to support decision making |
| Executive support | 1980–1990 | Focus on data analysis for decisions by senior executives |
| Online analytical processing (OLAP) | 1990–2000 | Software for analyzing multidimensional data tables |
| Business intelligence | 1989–2005 | Tools to support data driven decisions, with emphasis on reporting |
| Analytics | 2005–2010 | Focus on statistical and mathematical analysis for decisions |
| Big data | 2010–present | Focus on very large, unstructured, fast-moving data |

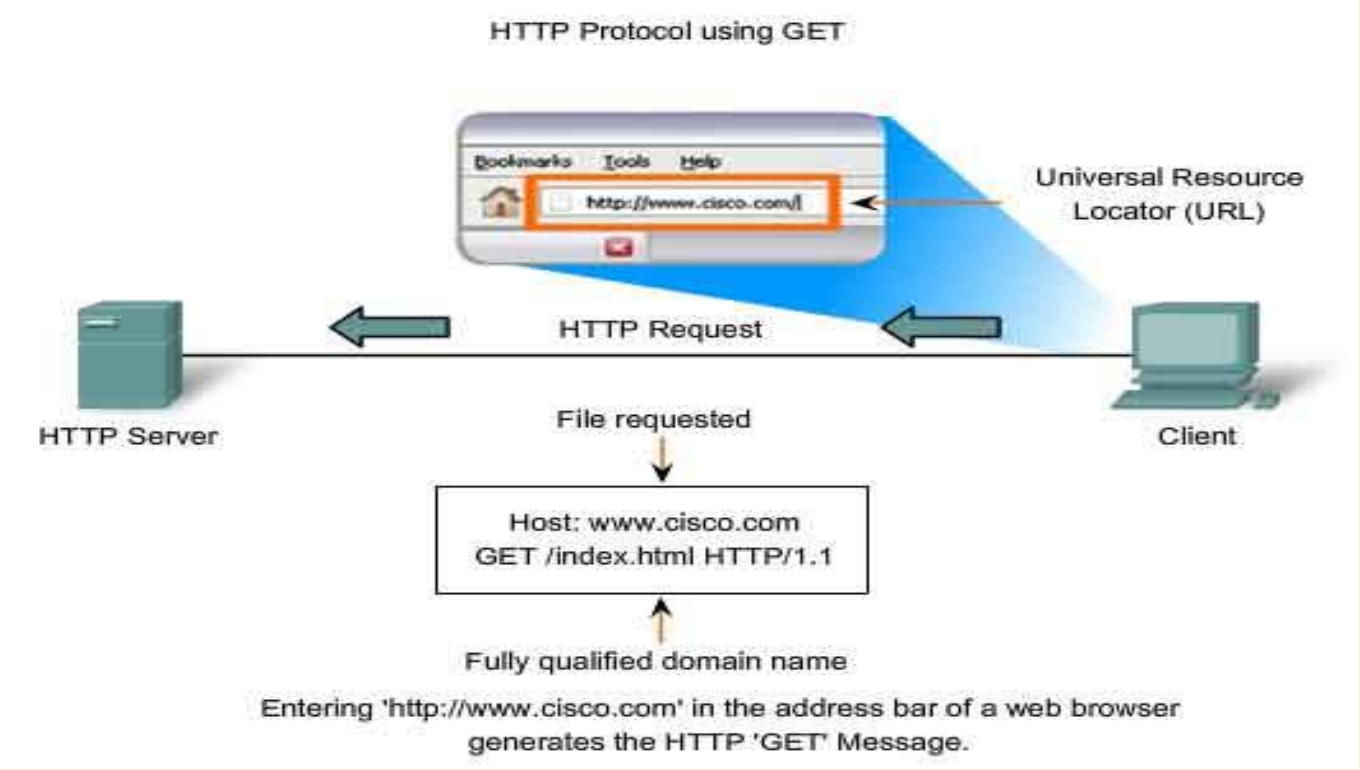
Week 9: Lecture2

# UNDERSTANDING THE WEB LOG

# We are going to learn

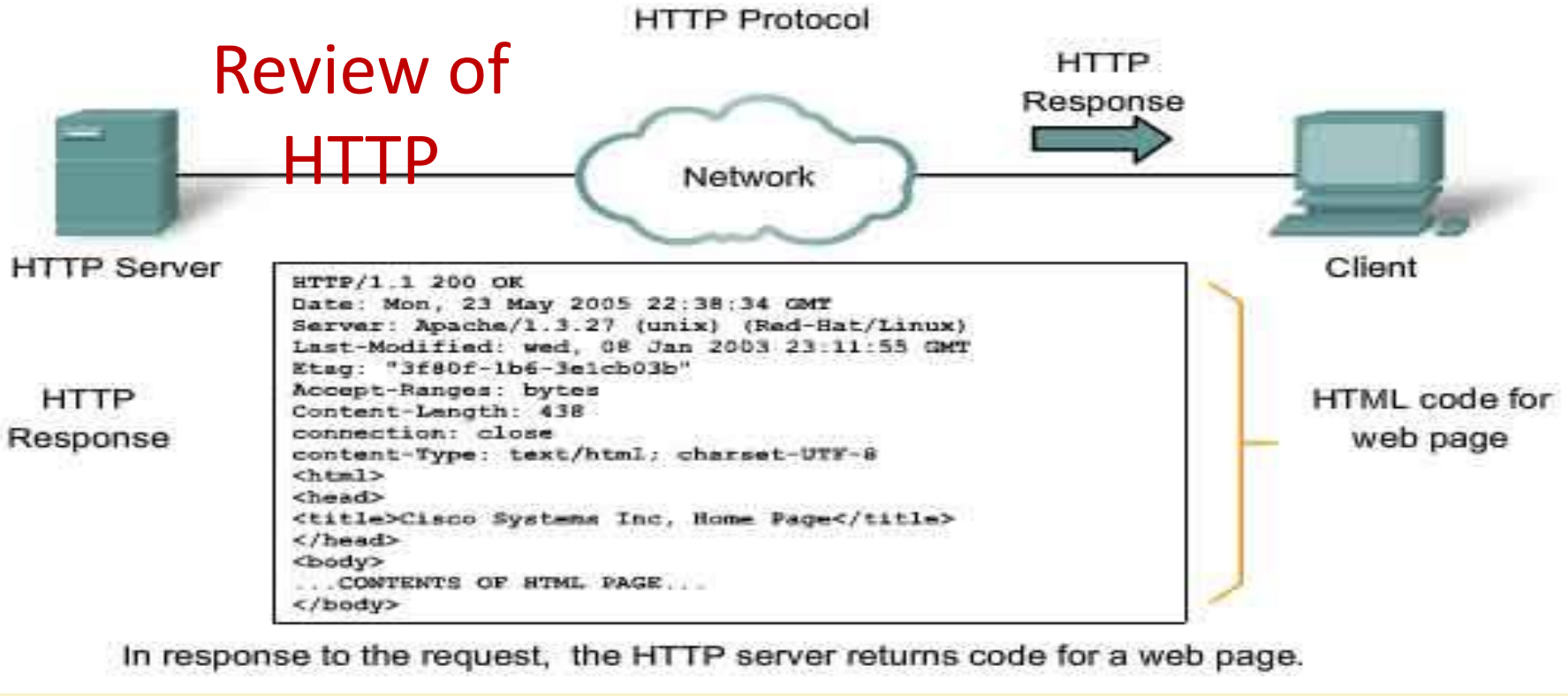
- How web logs are generated
- Structure of the access log
- Pre-processing
- Session identification

# Review of HTTP



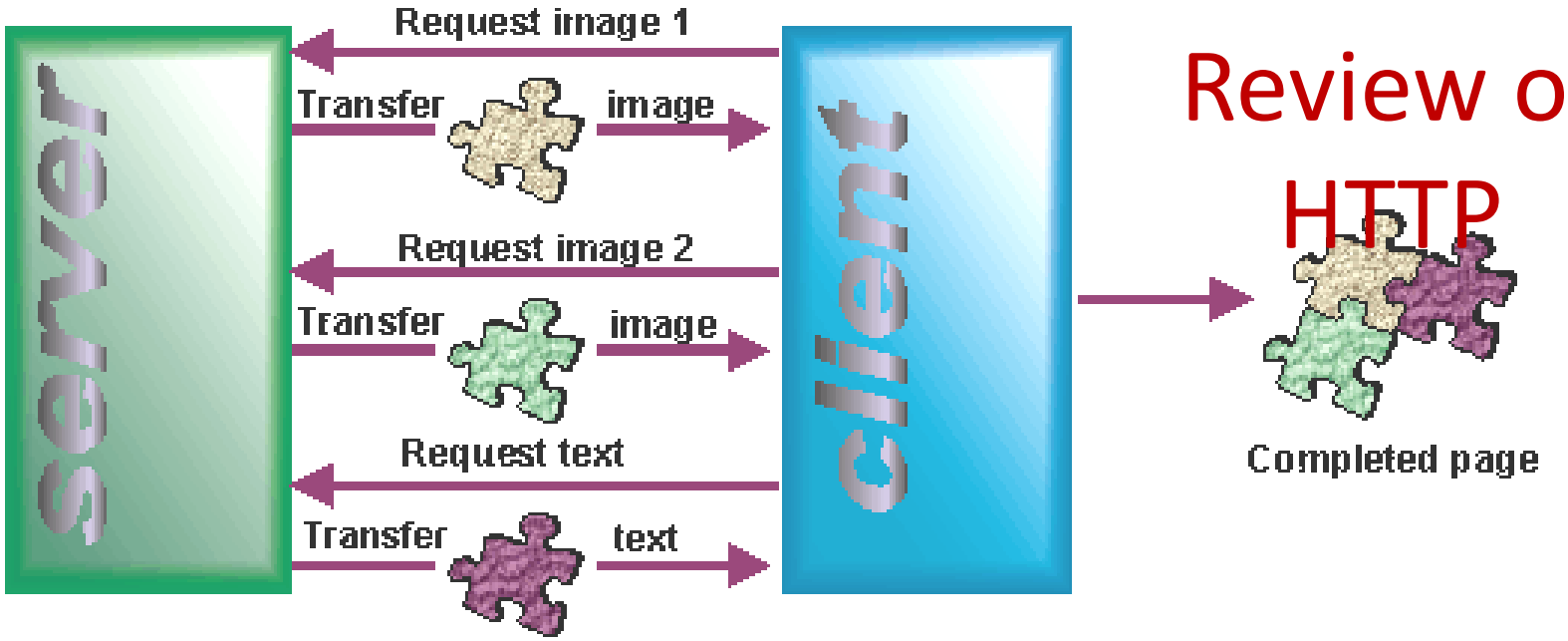

http://www.highteck.net/EN/Application/Application_Layer_Functionality_and_Protocols.html

# Review of HTTP



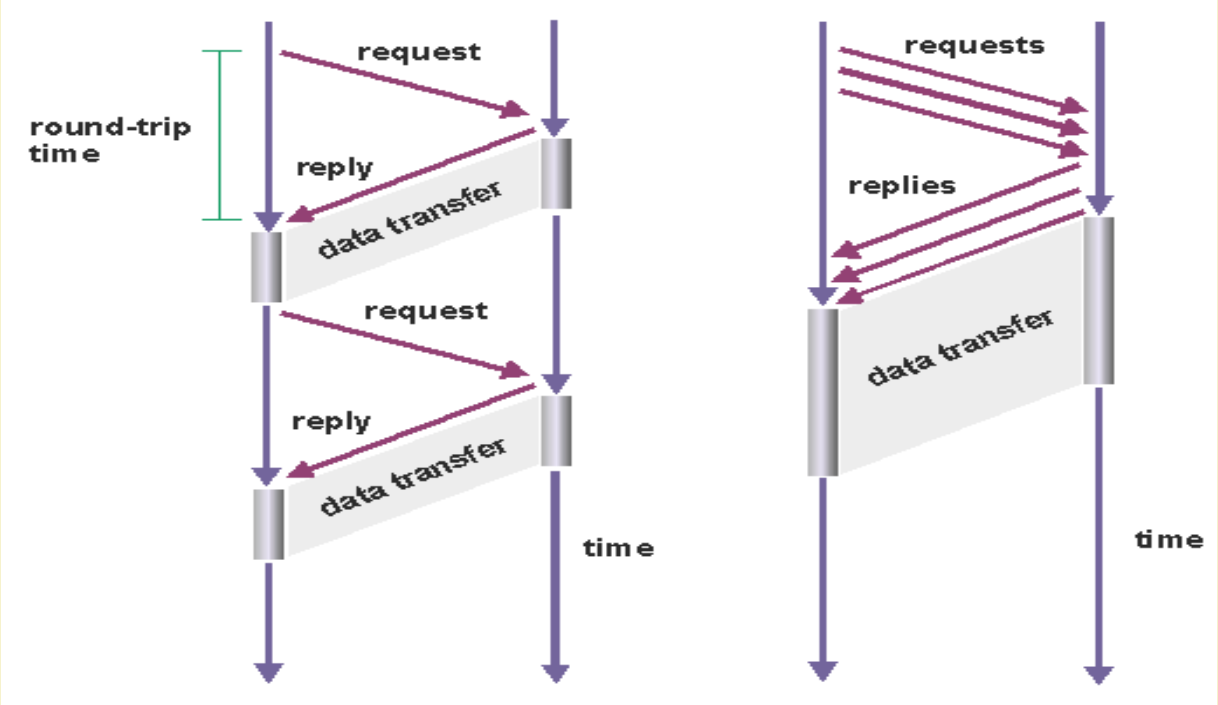

http://www.highteck.net/EN/Application/Application_Layer_Functionality_and_Protocols.html

# Review of HTTP



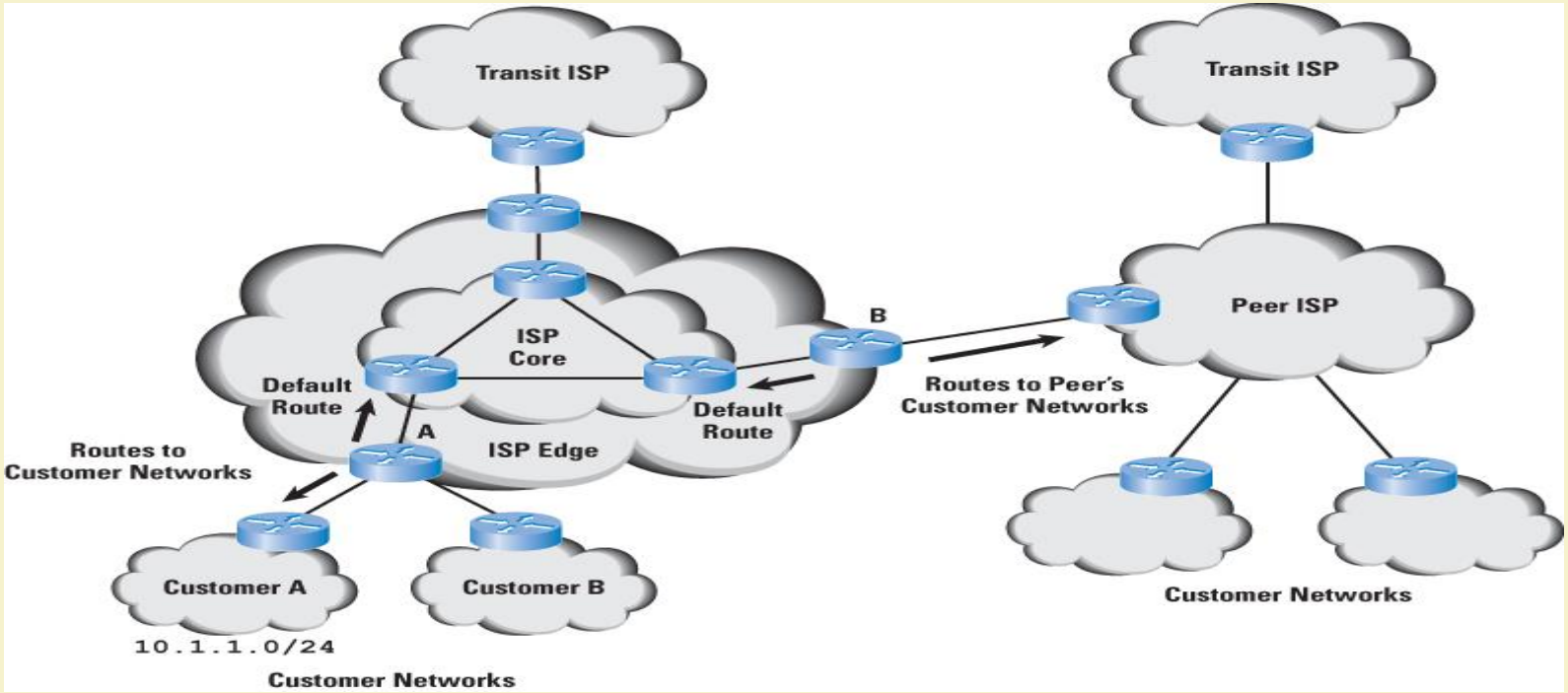

http://www.doc.ic.ac.uk/~nd/surprise_97/journal/vol2/pcg1/

# Review of HTTP



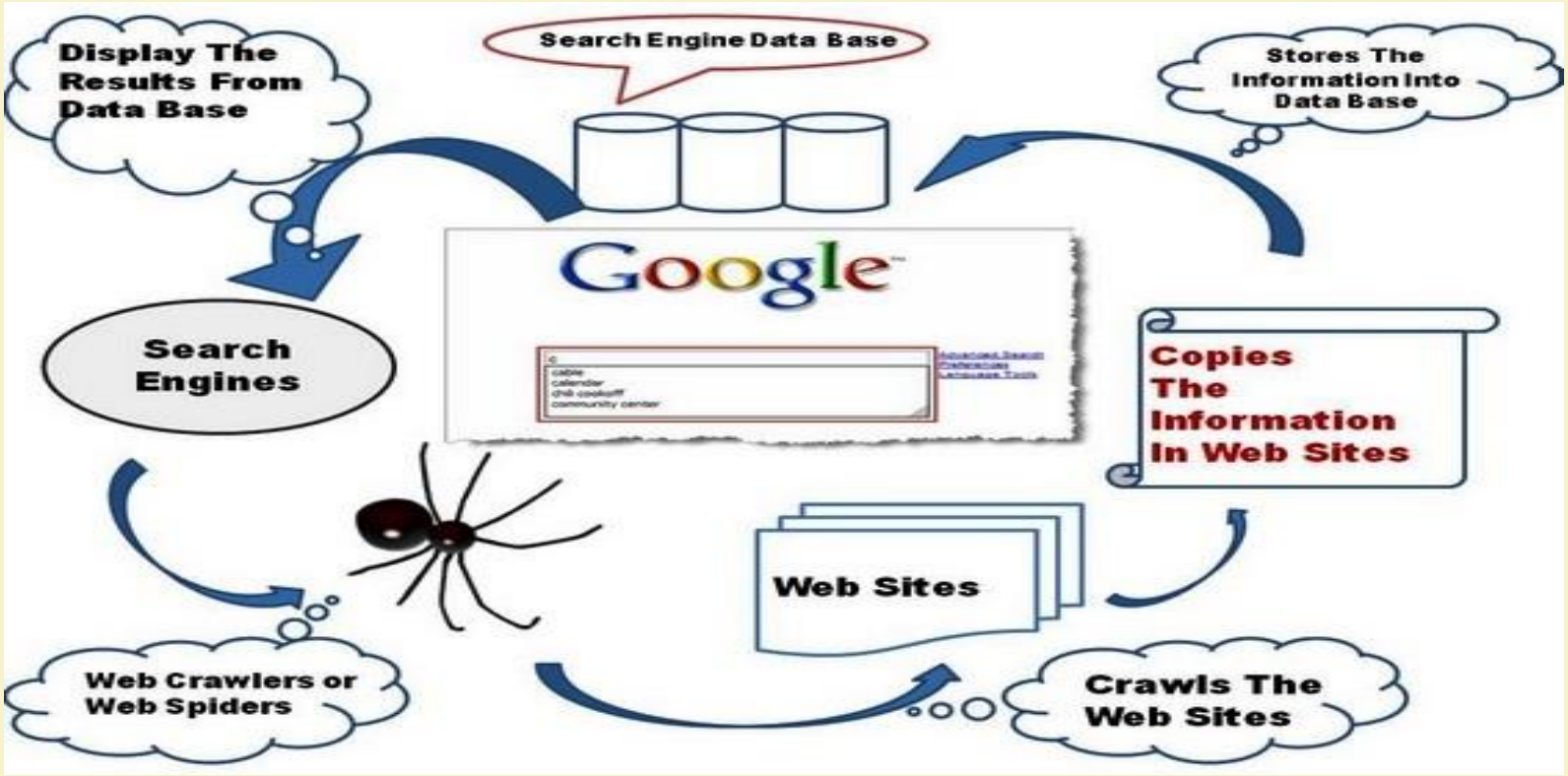

http://www.doc.ic.ac.uk/~nd/surprise_97/journal/vol2/pcg1/

# Review of HTTP




http://www.cisco.com/c/en/us/about/press/internet-protocol-journal/back-issues/table-contents-47/131-aggregation.html

# Review of HTTP




http://seo-advisors.com/searchegnies-web-crawlers/

# Collecting site navigation data

- Server Log Files
  - Access log
    - Common log format
    - Extended (Combined) log format
  - Error log



http://httpd.apache.org/docs/1.3/logs.html

# What's in a typical Web server log …

<ip_addr> - - <date><method><file><protocol><code><bytes><referrer><user_agent>

```
203.30.5.145 - - [01/Jun/1999:03:09:21 -0600] "GET /Calls/OWOM.html HTTP/1.0" 200 3942 "http://www.lycos.com/cgi-bin/pursuit?query=advertising+psychology-&maxhits=20&cat=dir" "Mozilla/4.5 [en] (Win98; I)"

203.30.5.145 - - [01/Jun/1999:03:09:23 -0600] "GET /Calls/Images/earthani.gif HTTP/1.0" 200 10689 "http://www.acr-news.org/Calls/OWOM.html" "Mozilla/4.5 [en] (Win98; I)"

203.30.5.145 - - [01/Jun/1999:03:09:24 -0600] "GET /Calls/Images/line.gif HTTP/1.0" 200 190 "http://www.acr-news.org/Calls/OWOM.html" "Mozilla/4.5 [en] (Win98; I)"

203.252.234.33 - - [01/Jun/1999:03:12:31 -0600] "GET / HTTP/1.0" 200 4980 "" "Mozilla/4.06 [en] (Win95; I)"

203.252.234.33 - - [01/Jun/1999:03:12:35 -0600] "GET /Images/line.gif HTTP/1.0" 200 190 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"

203.252.234.33 - - [01/Jun/1999:03:12:35 -0600] "GET /Images/red.gif HTTP/1.0" 200 104 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"

203.252.234.33 - - [01/Jun/1999:03:12:35 -0600] "GET /Images/earthani.gif HTTP/1.0" 200 10689 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"

203.252.234.33 - - [01/Jun/1999:03:13:11 -0600] "GET /CP.html HTTP/1.0" 200 3218 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"

203.30.5.145 - - [01/Jun/1999:03:13:25 -0600] "GET /Calls/AWAC.html HTTP/1.0" 200 104 "http://www.acr-news.org/Calls/OWOM.html" "Mozilla/4.5 [en] (Win98; I)"
```

| Field | Data | Description |
|---|---|---|
| Date | date | The date that the activity occurred |
| Time | time | The time that the activity occurred |
| Client IP address | c-ip | The IP address of the client that accessed your server |
| User Name | cs-username | The name of the autheticated user who access your server, anonymous users are represented by - |
| Service Name | s-sitename | The Internet service and instance number that was accessed by a client |
| Server Name | s-computername | The name of the server on which the log entry was generated |
| Server IP Address | s-ip | The IP address of the server that accessed your server |
| Server Port | s-port | The port number the client is connected to |
| Method | cs-method | The action the client was trying to perform |
| URI Stem | cs-uri-stem | The resource accessed |
| URI Query | cs-uri-query | The query, if any, the client was trying to perform |
| Protocol Status | sc-status | The status of the action, in HTTP or FTP terms |
| Win32 Status | sc-win32-status | The status of the action, in terms used by Microsoft Windows |
| Bytes Sent | sc-bytes | The number of bytes sent by the server |
| Bytes Received | cs-bytes | The number of bytes received by the server |
| Time Taken | time-taken | The duration of time, in milliseconds, that the action consumed |
| Protocol Version | cs-version | The protocol (HTTP, FTP) version used by the client |
| Host | cs-host | Display the content of the host header |
| User Agent | cs(User Agent) | The browser used on the client |
| Cookie | cs(Cookie) | The content of the cookie sent or received, if any |
| Referrer | cs(Referrer) | The previous site visited by the user. This site provided a link to the current site |

## W3C Extended Log File Format

s = server actions
c = client actions

cs = client-to-server actions
sc = server-to-client actions

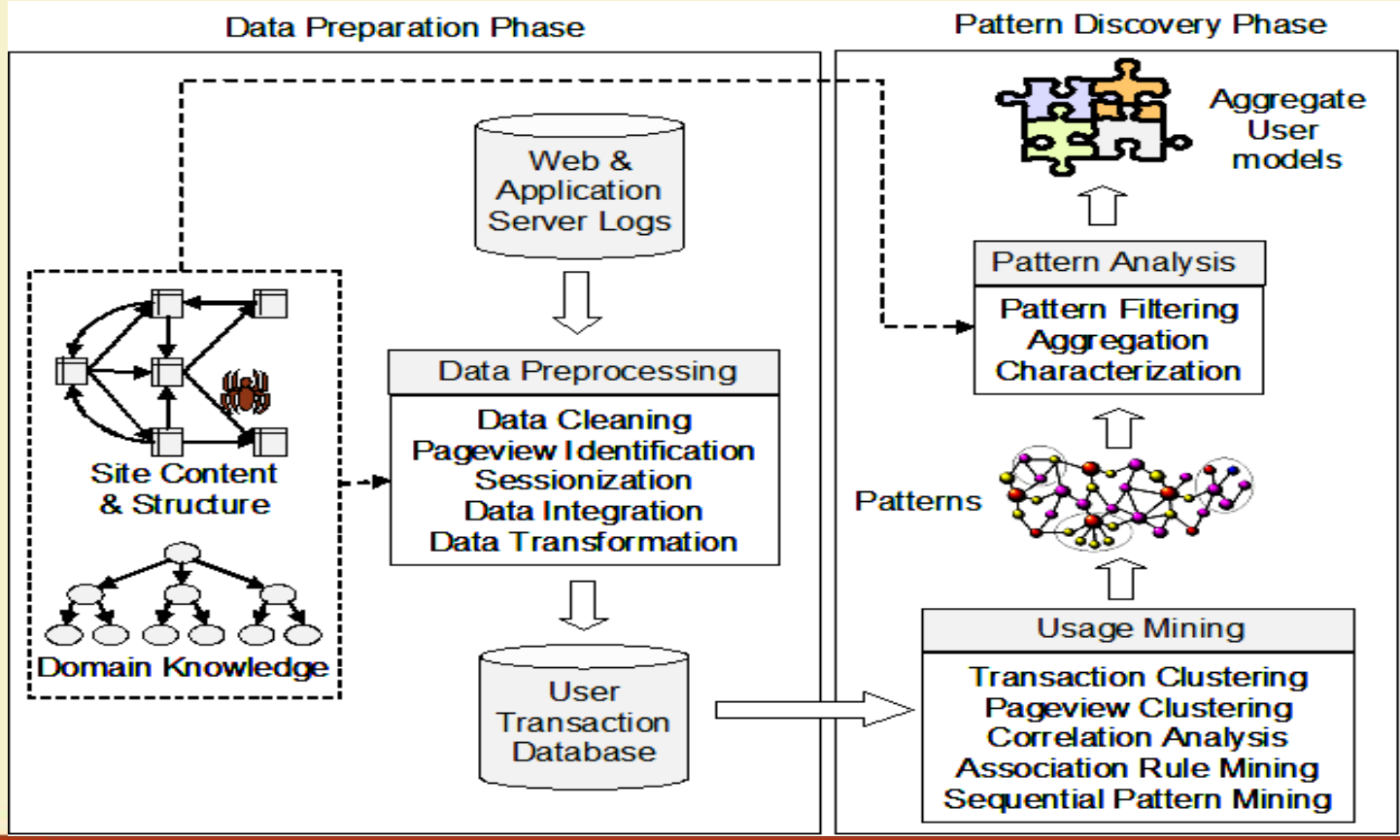
# Extended Log: Few Important Fields with a closer look

- IP Address (ISP Provided)
  - 144.16.192.247
- User name
  - determined by HTTP authentication
- Time
- Method/URL/Protocol:
  - Method of transaction such as GET or POST
  - URL
  - Version of the HTTP Protocol used by the server
- Status
  - HTTP status code
- Size
  - Total number of bytes transferred by the server to the client
- Referrer
  - The name of the URL from which the request originated
- Agent
  - Name and version of the browser making the request



**127.0.0.1** - **frank [10/Oct/2000:13:55:36 -0700]**
**"GET /apache_pb.gif HTTP/1.0" 200 2326 "http://www.example.com/start.html"**
"Mozilla/4.08 [en] (Win98; I ;Nav)"

# Web usage mining process



Week 9: Lecture3

# UNDERSTANDING THE WEB LOG-II

# We are going to learn

- How web logs are generated
- Structure of the access log
- Preprocessing
- Session identification

# Problems using the access log

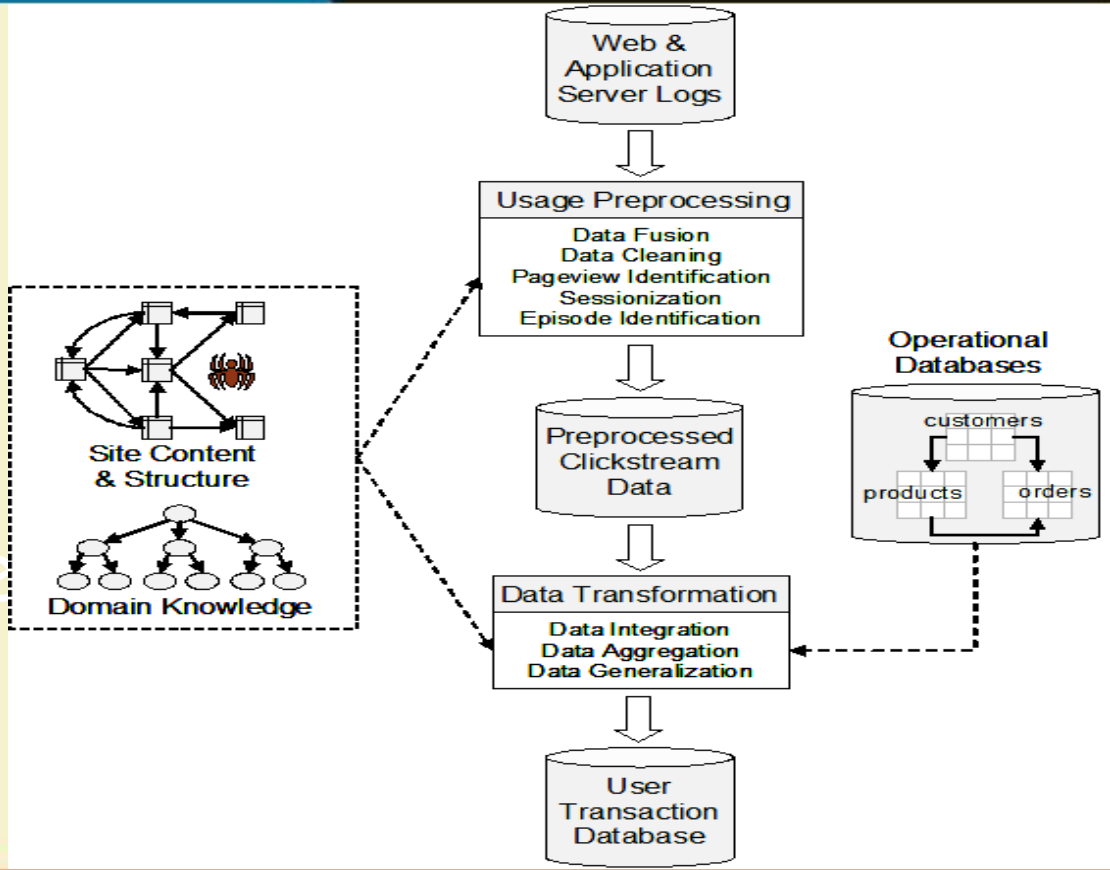
- Caching
- Dynamic Address Allocation by the ISP
- Stateless nature of the HTTP protocol
- Crawler Activity
- Tedious preprocessing and cleaning steps
  - Other Approaches for session tracking
    - Cookies
    - URL rewriting
    - Hidden form fields

# Filtering Merged Access Log Files

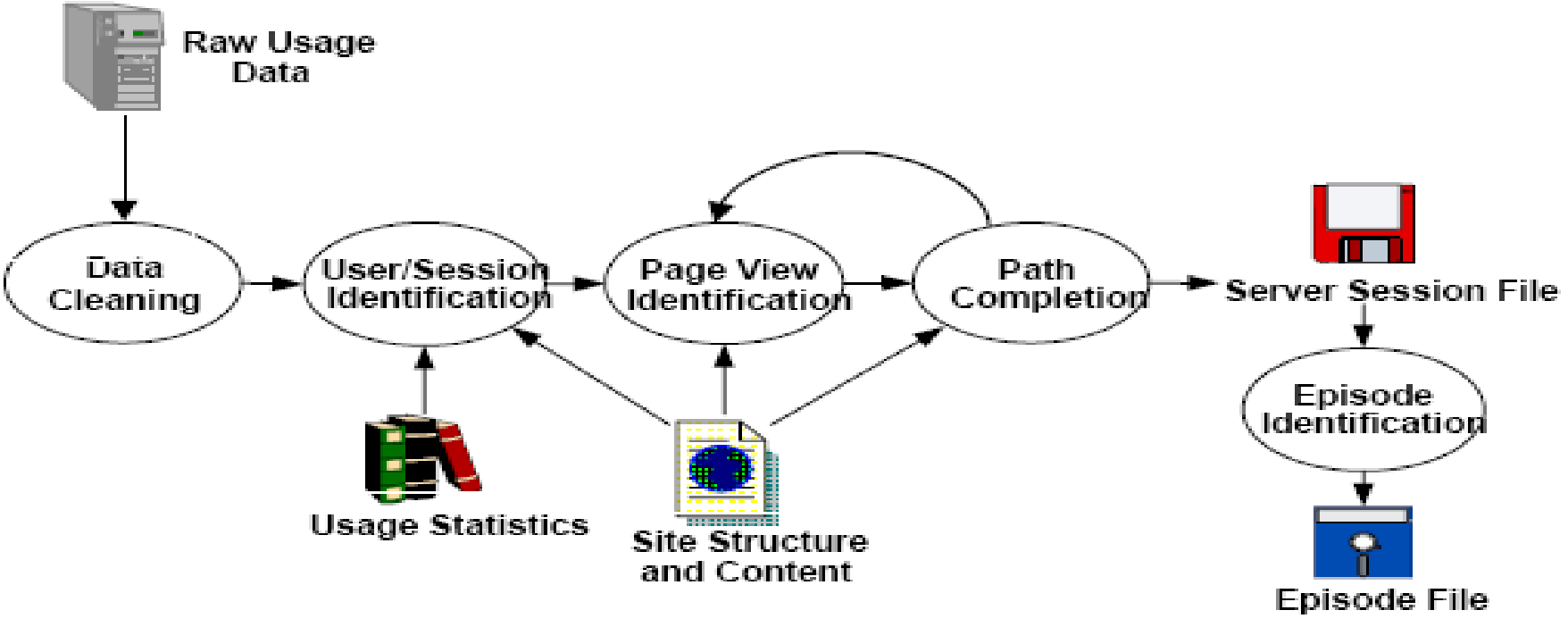
- Filtering the entries for embedded requests
  - Image, video and audio files
  - HTML files within a frame
- Filtering robot entries
  - Not human like trials
    - Searching all the links in an HTML document
    - Requests only for the text documents
  - Analyzing *user agent* fields
    - Tracing popular well-behaved robots
      - Robot.txt
- Using a table of web pages
- Filtering consumes 80% of the effort in log analysis

Detecting Robots: http://www.cs.princeton.edu/~kyoungso/papers/robot-usenix.pdf, http://ieeexplore.ieee.org/iel5/7101/19134/00884534.pdf, http://caltechlib.library.caltech.edu/73/01/Report-2004-NOV.pdf
Popular Robots: http://www.pgts.com.au/pgtsj/pgtsj0502d.html

# Data preparation



# Pre-processing of web usage data



# Data Preprocessing (1)

Data cleaning

- remove irrelevant references and fields in server logs
- remove references due to spider navigation
- remove erroneous references
- add missing references due to caching (done after sessionization)

Data integration

- synchronize data from multiple server logs
- Integrate semantics, e.g.,
  - meta-data (e.g., content labels)
  - e-commerce and application server data
- integrate demographic / registration data

# Data Preprocessing (2)

## Data Transformation

- user identification
- sessionization / episode identification
- pageview identification
  - a pageview is a set of page files and associated objects that contribute to a single display in a Web Browser

## Data Reduction

- sampling and dimensionality reduction (ignoring certain pageviews / items)
- Identifying User Transactions (i.e., sets or sequences of pageviews possibly with associated weights)

# Why sessionize?

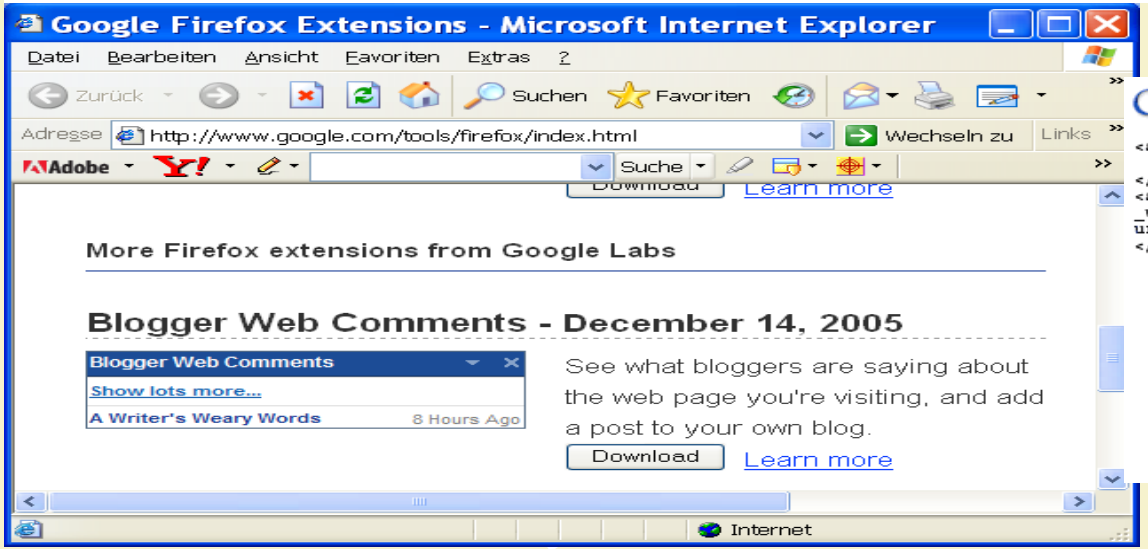
- Quality of the patterns discovered depends on the quality of the data on which mining is applied.
- In Web usage analysis, these data are the sessions of the site visitors: the activities performed by a user from the moment she enters the site until the moment she leaves it.
- Difficult to obtain reliable usage data due to proxy servers and anonymizers, dynamic IP addresses, missing references due to caching, and the inability of servers to distinguish among different visits.
- Cookies and embedded session IDs produce the most faithful approximation of users and their visits, but are not used in every site, and not accepted by every user.
- Therefore, *heuristics* are needed that can sessionize the available access data.

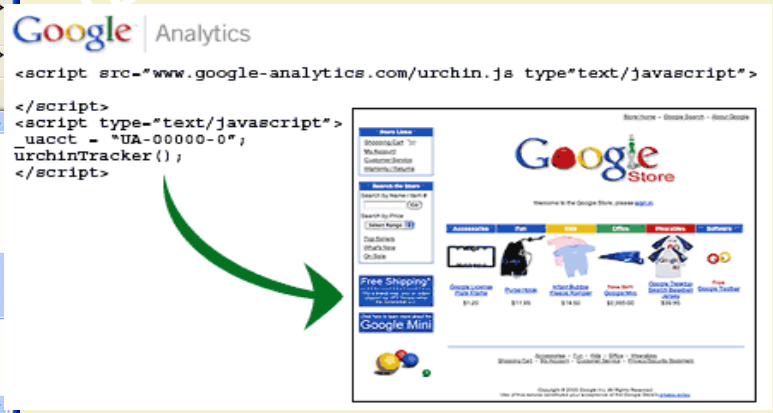
## Mechanisms for User Identification

| Method | Description | Privacy Concerns | Advantages | Disadvantages |
|---|---|---|---|---|
| IP Address + Agent | Assume each unique IP address/Agent pair is a unique user | Low | Always available. No additional technology required. | Not guaranteed to be unique. Defeated by rotating IPs. |
| Embedded Session Ids | Use dynamically generated pages to associate ID with every hyperlink | Low to medium | Always available. Independent of IP addresses. | Cannot capture repeat visitors. Additional overhead for dynamic pages. |
| Registration | User explicitly logs in to the site. | Medium | Can track individuals not just browsers | Many users won't register. Not available before registration. |
| Cookie | Save ID on the client machine. | Medium to high | Can track repeat visits from same browser. | Can be turned off by users. |
| Software Agents | Program loaded into browser and sends back usage data. | High | Accurate usage data for a single site. | Likely to be rejected by users. |

Examples: page tags (use javascript), some browser plugins

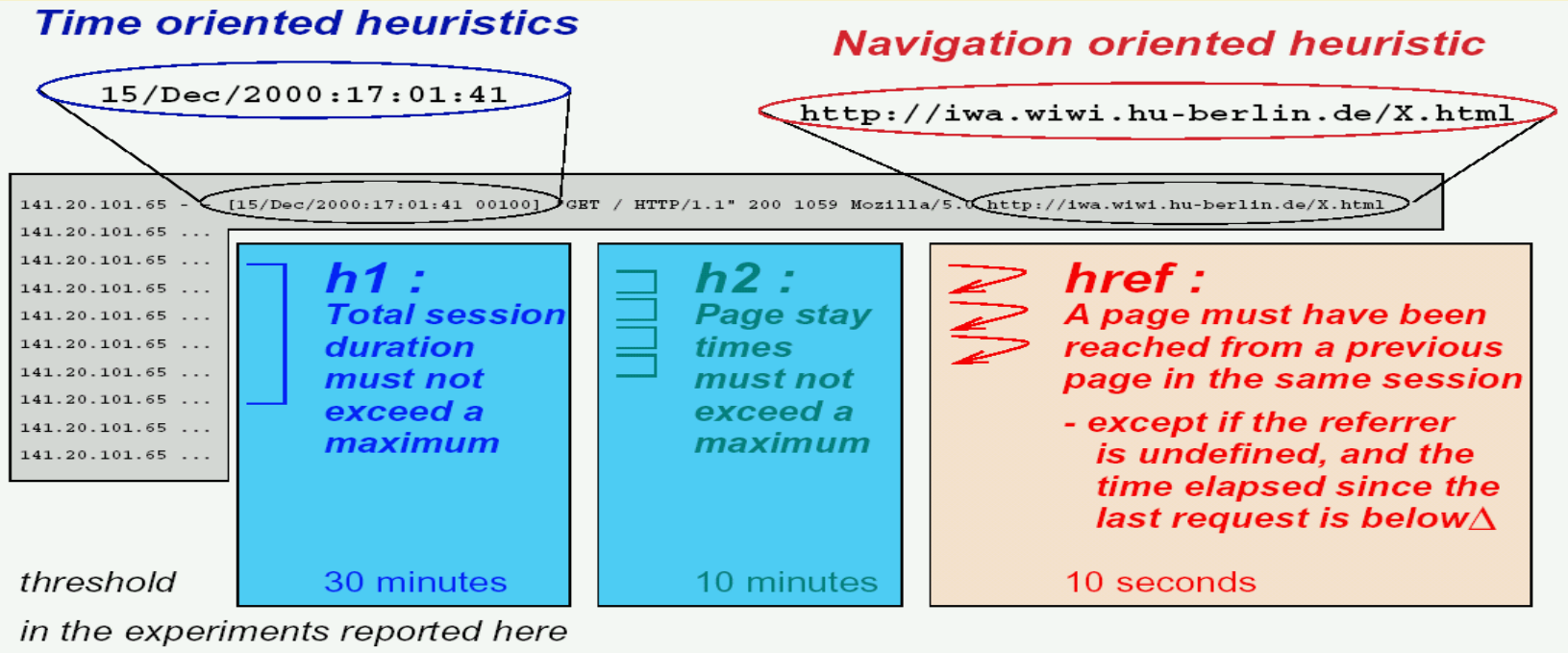
# Examples of “software agents“


Page tagging with Javascript: see also
http://www.bruceclay.com/analytics/disadvantages.htm

# Sessionization strategies: Sessionization heuristics




These heuristics are quite accurate! (see Spiliopoulou et al., 2003)

# Path Completion

- Refers to the problem of inferring missing user references due to caching.
- Effective path completion requires extensive knowledge of the link structure within the site
- Referrer information in server logs can also be used in disambiguating the inferred paths.
- Problem gets much more complicated in frame-based sites.

# Why integrate semantics?

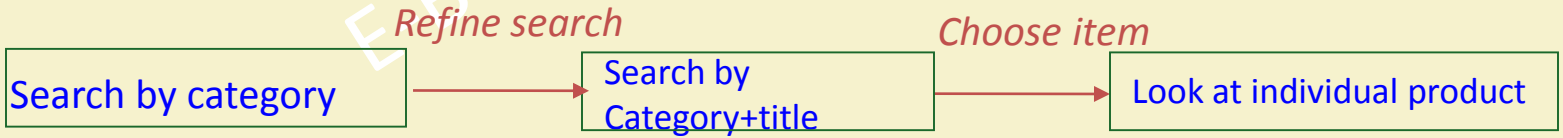
Basic idea: associate each requested page with one or more domain concepts, to better understand the process of navigation / Web usage

*Example: a shopping site*

From ...

```
p3ee24304.dip.t-dialin.net - - [19/Mar/2002:12:03:51 +0100]
     "GET /search.html?l=ostsee%20strand&syn=023785&ord=asc HTTP/1.0" 200 1759
p3ee24304.dip.t-dialin.net - - [19/Mar/2002:12:05:06 +0100]
     "GET /search.html?l=ostsee%20strand&p=low&syn=023785&ord=desc HTTP/1.0" 200 8450
p3ee24304.dip.t-dialin.net - - [19/Mar/2002:12:06:41 +0100]
     "GET /mlesen.html?Item=3456&syn=023785 HTTP/1.0" 200 3478
```

To ...



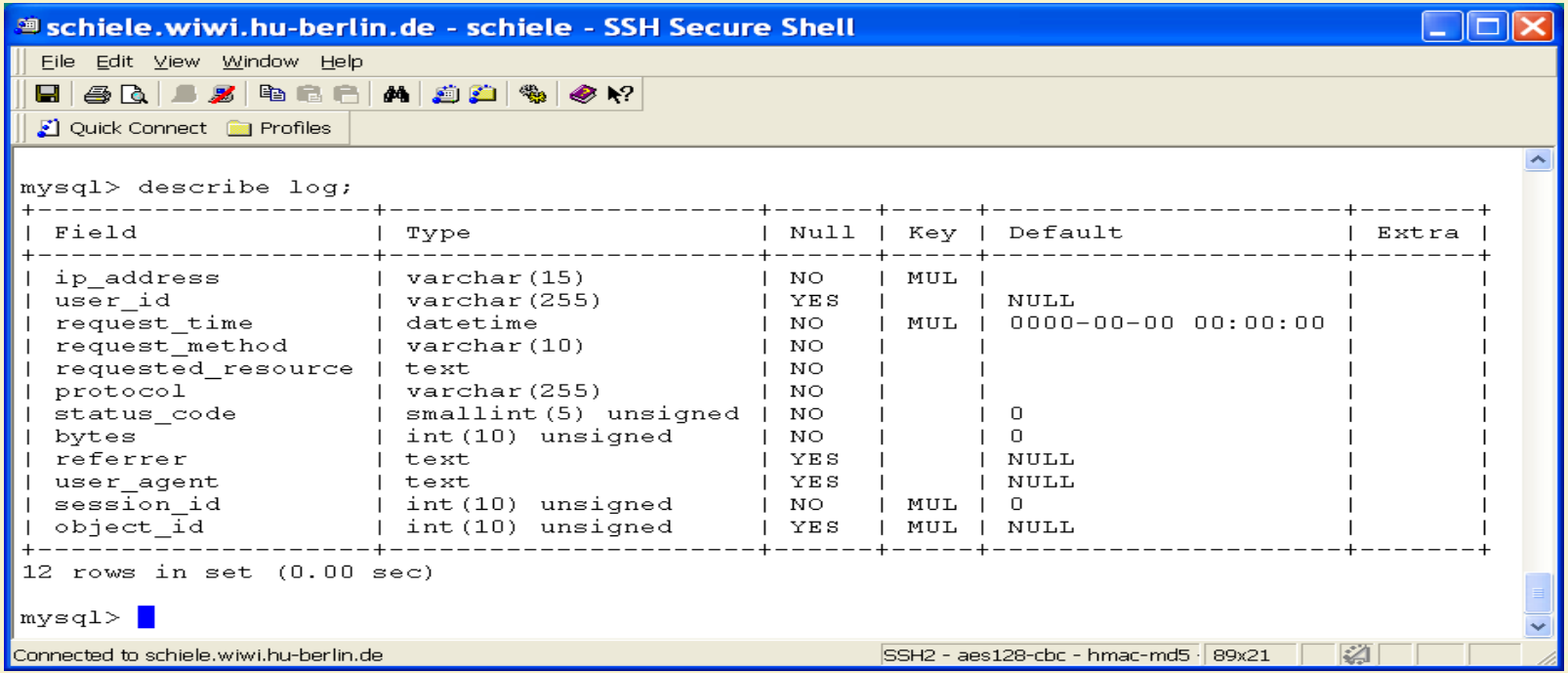
# From URLs to topics / concepts: Basics of semantic session modelling

- 1 request ➔ 1 concept or n concepts
- Concepts can concern content or service
- Concepts can be part of an ontology (simple case: concept hierarchy)
- Session = set / sequence / tree / graph of requests

➔ also possible: n requests ➔ 1 concept

# Resulting format: if the request is the instance

## Usually flat file (format like Web server log) or database



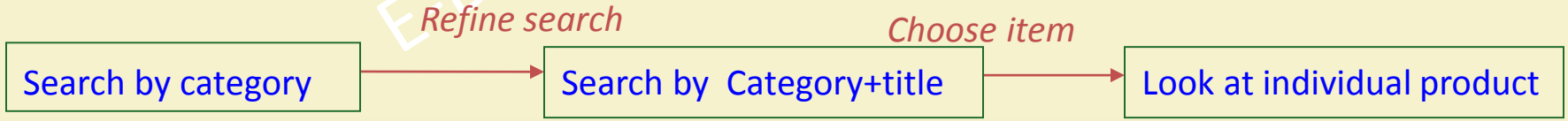
# Resulting format: If a session is the instance

- What features can a session have?
- *Refer again to the example:*



```
p3ee24304.dip.t-dialin.net - - [19/Mar/2002:12:03:51 +0100]
     "GET /search.html?l=ostsee%20strand&syn=023785&ord=asc HTTP/1.0" 200 1759
p3ee24304.dip.t-dialin.net - - [19/Mar/2002:12:05:06 +0100]
     "GET /search.html?l=ostsee%20strand&p=low&syn=023785&ord=desc HTTP/1.0" 200 8450
p3ee24304.dip.t-dialin.net - - [19/Mar/2002:12:06:41 +0100]
     "GET /mlesen.html?Item=3456&syn=023785 HTTP/1.0" 200 3478,
```





E-Business

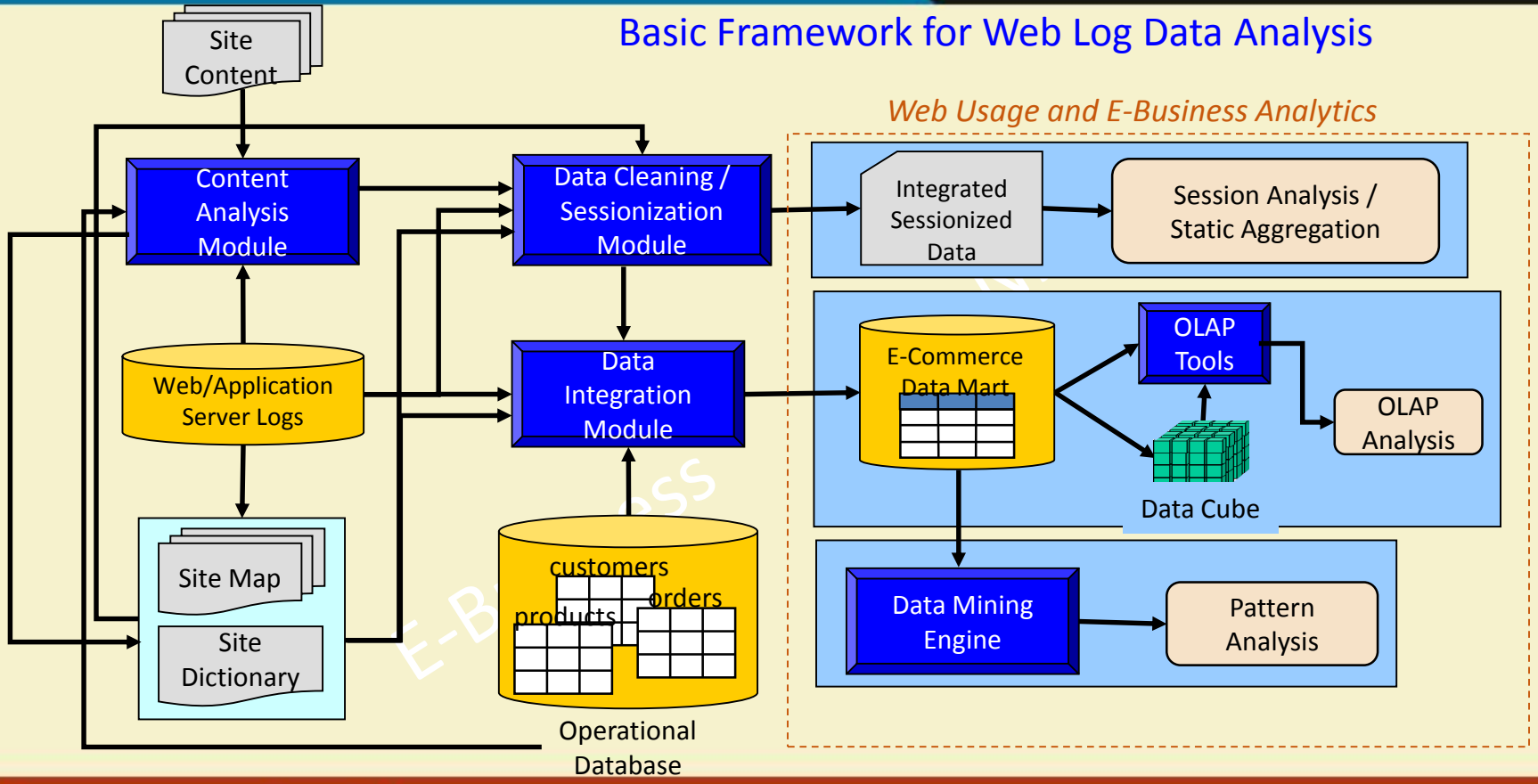
Week 9: Lecture4

# USING THE WEB LOG: WEB USAGE MINING

# We are going to learn

- Framework for analysing the Web Log
- Types of Analysis

# Basic Framework for Web Log Data Analysis



# Web Usage and E-Business Analytics

## Different Levels of Analysis

- Static Aggregation and Statistics
- Session Analysis
- OLAP
- Data Mining

# Static Aggregation (Reports)

Most common form of analysis.

Data aggregated by predetermined units such as days or sessions.

Advantages:

- Gives quick overview of how a site is being used.
- Minimal disk space or processing power required.

Drawbacks:

- No ability to “dig deeper” into the data.

| Page View | Number of Sessions | Average View Count per Session |
|---|---|---|
| Home Page | 50,000 | 1.5 |
| Catalog Ordering | 500 | 1.1 |
| Shopping Cart | 9000 | 2.3 |

# Session Analysis

Simplest form of analysis: examine individual or groups of server sessions and e-commerce data.

Advantages:

- Gain insight into typical customer behaviors.
- Trace specific problems with the site.

Drawbacks:

- LOTS of data.
- Difficult to generalize.

# Online Analytical Processing (OLAP)

Allows changes to aggregation level for multiple dimensions.

Generally associated with a Data Warehouse.

Advantages & Drawbacks

- Very flexible
- Requires significantly more resources than static reporting.

| Page View | Number of Sessions | Average View Count per Session |
|---|---|---|
| Kid's Stuff Products | 2,000 | 5.9 |

| Page View | Number of Sessions | Average View Count per Session |
|---|---|---|
| Kid's Stuff Products | | |
| Electronics | | |
| Educational | 63 | 2.3 |
| Radio-Controlled | 93 | 2.5 |

# Web Log Analytics

- The measurement, collection, analysis and reporting of internet data for purposes of understanding and optimizing web usage
- Tools
  - Webalizer
  - Sawmill
  - WebTrends
  - AWStats
  - WWWStat
  - Apache Logs Viewer
  - Google analytics

## Level of Processing

Static Aggregation and Statistics

Session Analysis

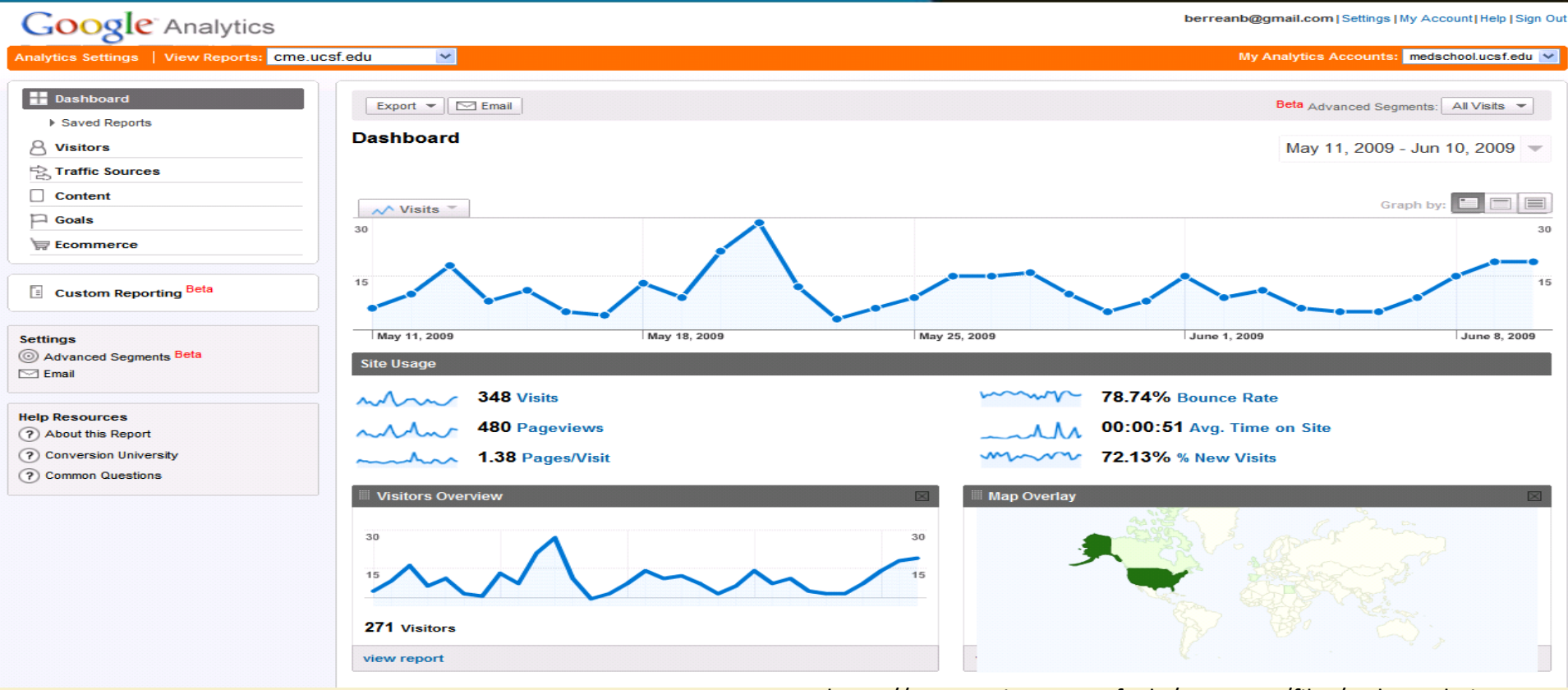
# Few Definitions

- Hits
  - A request for a file from the web server. Available only in log analysis
- Page Views
  - A request for a file whose type is defined as a page
- Visits/Sessions
  - A series of requests from the same uniquely identified client with a set timeout, often 30 minutes. A visit contains one or more page views
- Click Paths
  - the sequence of hyperlinks one or more website visitors follows on a given site

# Page Tagging

```
  <SCRIPT LANGUAGE="JavaScript1.2"
        SRC="/design/redesign/global/js/HM_Loader.js"
        TYPE='text/javascript'></SCRIPT>

<script src="https://www.google-analytics.com/urchin.js" type="text/javascript">
</script>
<script type="text/javascript">
_uacct = "UA-410306-2";
urchinTracker();
</script>
```

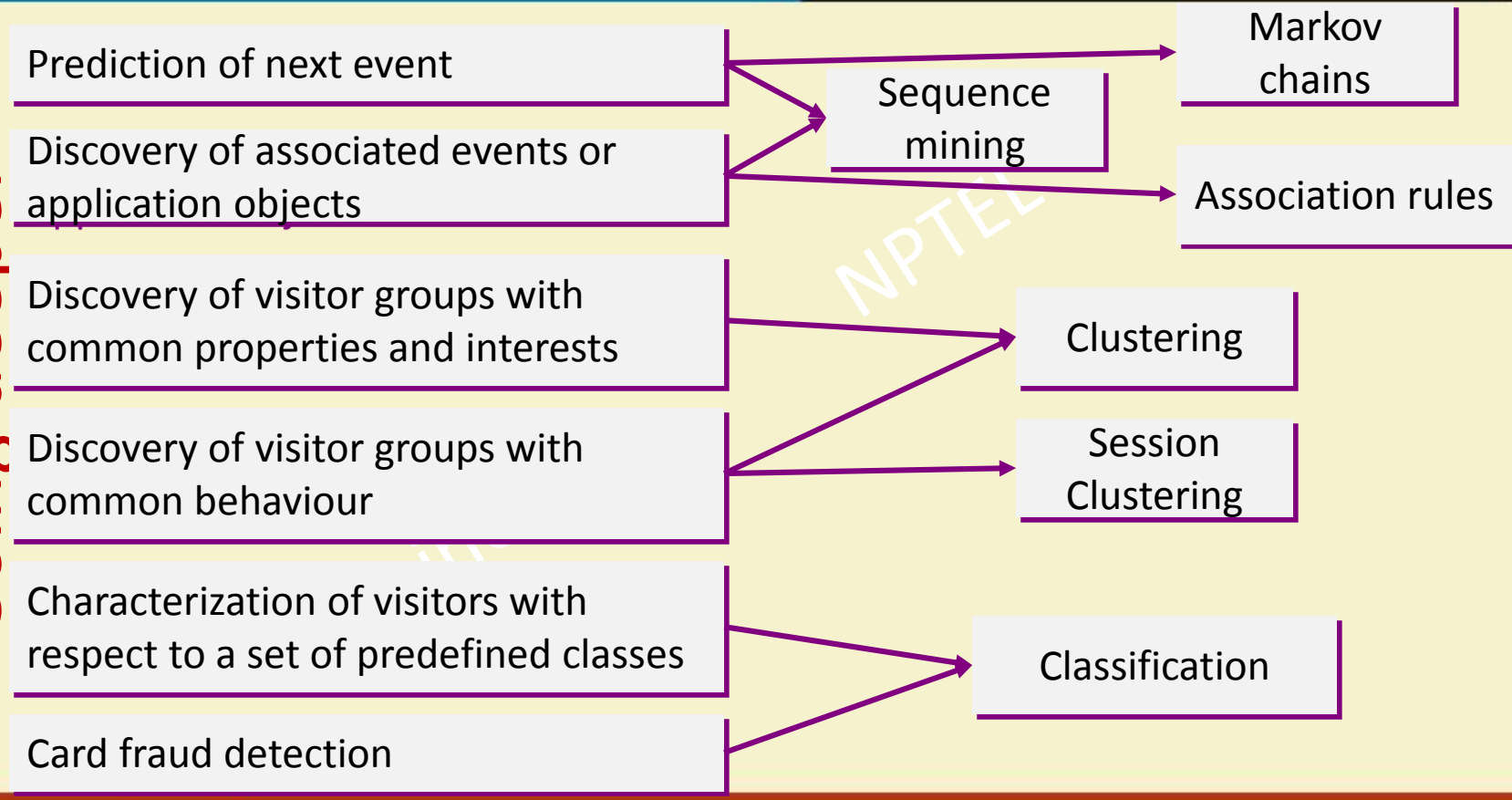



https://communicators.ucsf.edu/resources/files/web_analytics.ppt

# What Numbers Say

- About Navigation
- About Content
- About Users

# Data Mining: Going deeper



# Mining Navigation Patterns

- Each session induces a user trail through the site
- A trail is a sequence of web pages followed by a user during a session, ordered by time of access.
- A pattern in this context is a frequent trail.
- *Co-occurrence* of web pages is important, e.g. *shopping-basket* and *checkout.*
  - Association rule mining
  - Markov chain model.

# Trails inferred from Log data
# (Each session results in a trail)

| ID | Trail |
|---|---|
| 1 | A1 > A2 > A3 |
| 2 | A1 > A2 > A3 |
| 3 | A1 > A2 > A3 > A4 |
| 4 | A5 > A2 > A4 |
| 5 | A5 > A2 > A4 > A6 |
| 6 | A5 > A2 > A3 > A6 |

**Association based Approach**

# Association Rule Mining-The Idea

Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction



Market-Basket transactions

| TID | Items |
|---|---|
| 1 | Bread, Milk |
| 2 | Bread, Diaper, Beer, Eggs |
| 3 | Milk, Diaper, Beer, Coke |
| 4 | Bread, Milk, Diaper, Beer |
| 5 | Bread, Milk, Diaper, Coke |

Example of Association Rules

{Diaper} → {Beer},
{Milk, Bread} → {Eggs,Coke},
{Beer, Bread} → {Milk},

Implication means co-occurrence, not causality!

# Applications

- Pre-fetching and caching web pages
- Web site reorganisation
- Personalisation
- Recommendation of links and products

# Applications

- Calibration of a Web server:
  - Prediction of the next page invocation over a group of concurrent Web users under certain constraints
    - Sequence mining, Markov chains
- Cross-selling of products:
  - Mapping of Web pages/objects to products
  - Discovery of associated products
    - Association rules, Sequence Mining
  - Placement of associated products on the same page

# Applications

Sophisticated cross-selling and up-selling of products:

- Mapping of pages/objects to products of different price groups
- Identification of Customer Groups
  - Clustering, Classification
- Discovery of associated products of the same/different price categories
  - Association rules, Sequence Mining
- Formulation of recommendations to the end-user
  - Suggestions on associated products
  - Suggestions based on the preferences of similar users

# Summary

- Web usage mining has emerged as the essential tool for realizing more personalized, user-friendly and business-optimal Web services.
- The key is to use the user-clickstream data for many mining purposes.
- Traditionally, Web usage mining is used by e-commerce sites to organize their sites and to increase profits.
- It is now also used by search engines to improve search quality and to evaluate search results, etc, and by many other applications.

Week 9: Lecture 5

# USER BEHAVIOR MODELING FROM WEB LOG

# We are going to learn

- A model of browsing behaviour
- Interpreting the model outcome

# Probabilistic models of browsing behavior

- Useful to build models that describe the browsing behavior of users
- Can generate insight into how users use the website
- Provide mechanism for making predictions
- Can help in pre-fetching and personalization

# Markov models for understanding user behavior

- General approach is to use a finite-state Markov chain
  - Each state can be a specific Web page or a category of Web pages
  - If only interested in the order of visits (and not in time), each new request can be modeled as a transition of states
- Issues
  - Self-transition
  - Time-independence

# Discrete – Time Markov Chains

Many real-world systems contain uncertainty and evolve over time.

Stochastic processes (and Markov chains) are probability models for such systems.

A discrete-time stochastic process is a sequence of random variables $X_0, X_1, X_2, \ldots$ typically denoted by $\{X_n\}$.

# Modeling A Website

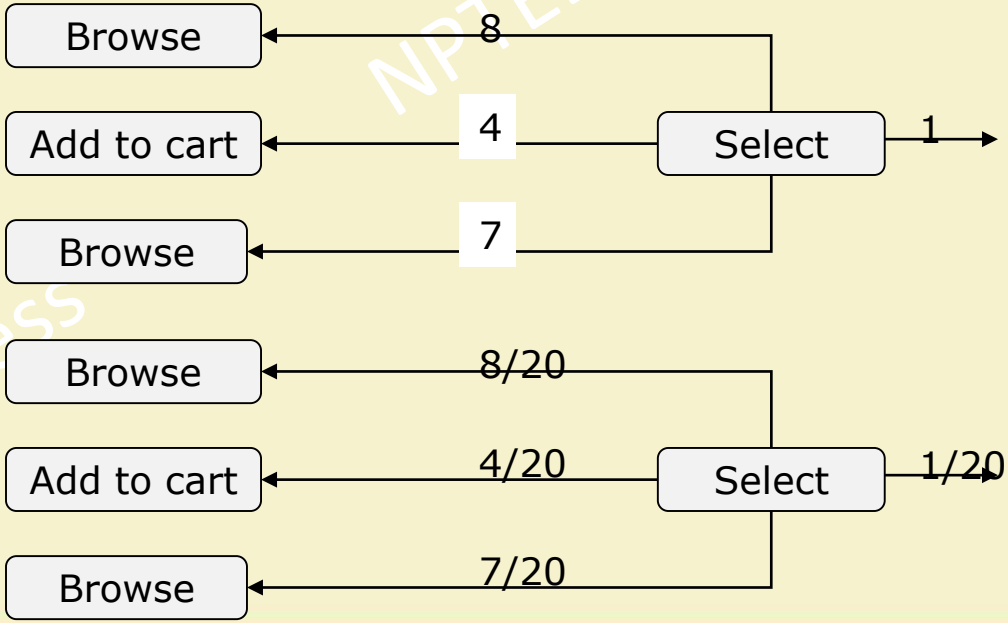
- State: A functional area in the website
  - A page or a group of pages representing the functional area
- Two dummy states
  - *entry* and *exit*
  - Customer is assumed to stay in the *entry* state before entering into the site
  - Customer is assumed to stay in the *exit* state.
- Customer behavior model graph
  - Static part
  - Dynamic part

# Building a User Behavior Graph

- Static Part
  - Determine the set of functions provided by the e-commerce site.
    - States
    - Group of web pages
  - Determine all possible transitions between states
    - From site layout
- Dynamic Part
  - Transition probability matrix
  - Average transition time matrix

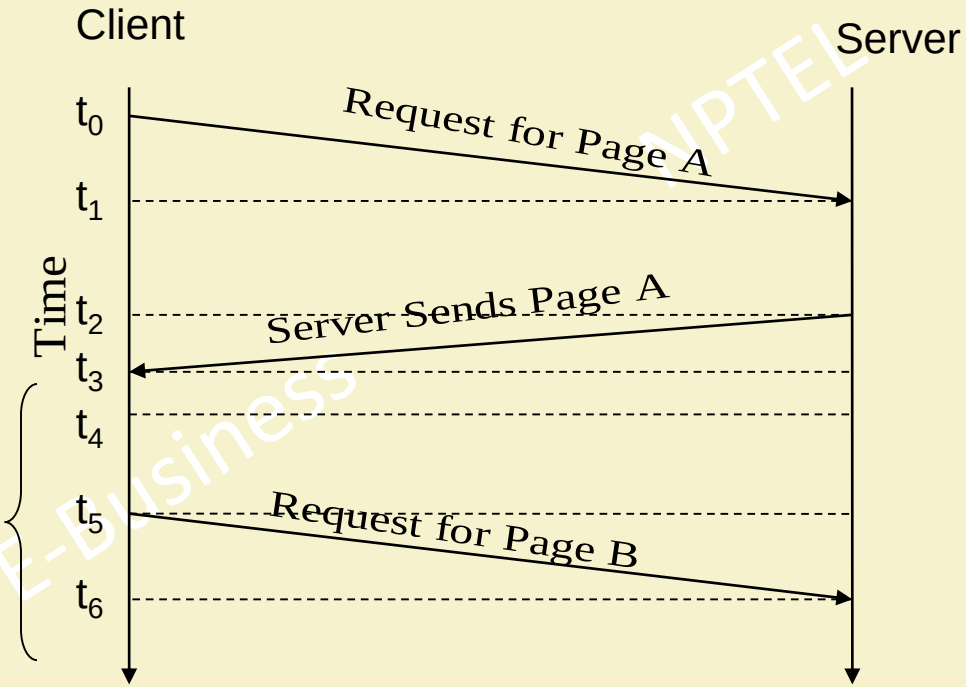
# Determining transition probability matrix

- Count frequency of transitions from one state to another
- Calculate probability



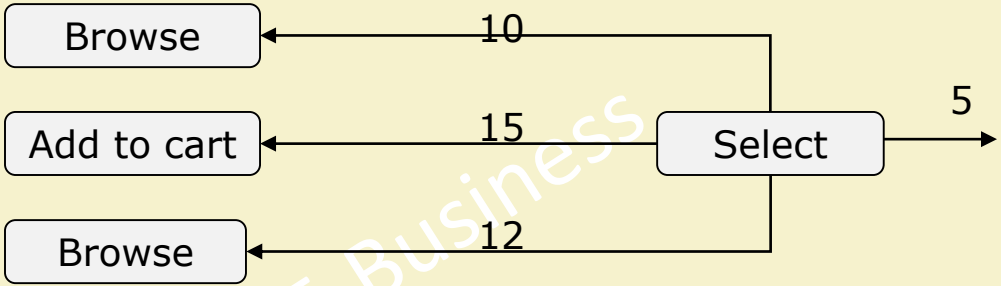
# Customer's think time

Customer's Think time

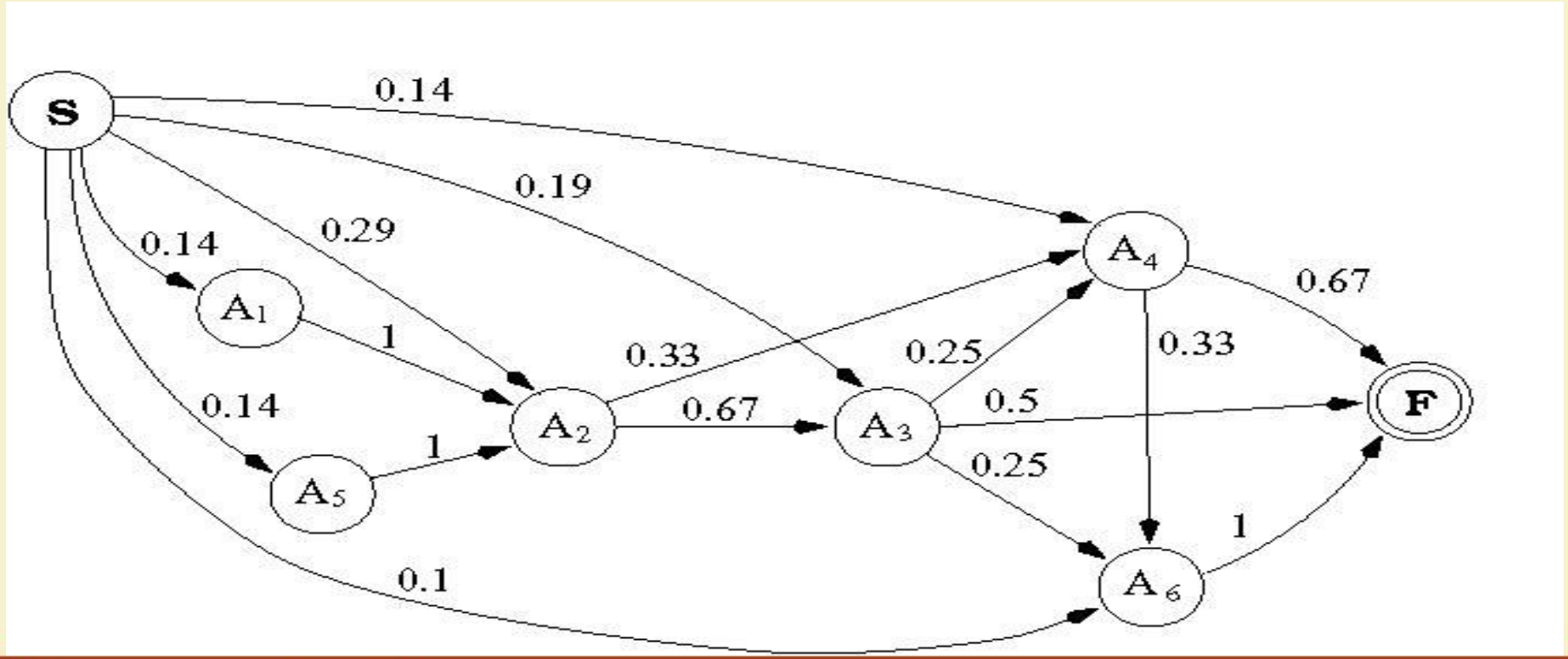


# Finding average think time

- Total think time from all the visits from one state to the other/frequency of visit



# Browsing Behaviour as a Markov Chain



# Properties of the transition probability matrix of a CBMG

- $p_{i1} = 0 \quad 2 \leq i \leq n-1$
  - No transition can be made to the *Entry* state from any state other than the *Exit* state.
- $p_{1n} = 0$
  - No transition can be made from the *Entry* state to the *Exit* state.
- $p_{nj} = 0 \quad 2 \leq j \leq n-1$
  - No transition can be made from the *Exit* state to any state other than the *Entry* state.
- $p_{nn} + p_{n1} = 1$
  - A transition from the *Exit* state to itself or to the *Entry* state.

# End of Week 9