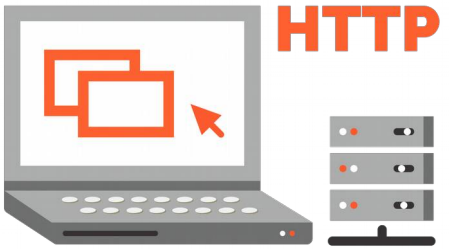
Captura y obtención de datos
Fuentes de datos

Fuentes de datos

- Cuatro fuentes principales



  - Datos de solicitudes HTTP directas
  - Datos a nivel de red y generados por el servidor, asociados con las solicitudes HTTP
  - Datos a nivel de aplicación enviados con la solicitud HTTP
  - Datos externos

Solicitudes HTTP directas

- ¿Qué es una solicitud HTTP?
  - Mensaje que el cliente web envía al servidor web solicitando algún recurso
    - Página web, imagen, …
  - Se utiliza tradicionalmente como medida del tráfico
  - Está descrita por un conjunto de dimensiones
    - Página, visitante, tecnología, ..

Solicitudes HTTP directas

- ¿Qué es una solicitud HTTP?

Request command to get "www.spsu.edu/itdegrees"

HTTP request headers

```
GET /itdegrees/ HTTP/1.1
Host: www.spsu.edu
Connection: keep-alive
User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.4 (KHTML, like Gecko)
Chrome/22.0.1229.94 Safari/537.4
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Referer: http://spsu.edu/it/
Accept-Encoding: gzip,deflate,sdch
Accept-Language: en-US,en;q=0.8
Accept-Charset: ISO-8859-1,utf-8;q=0.7,*;q=0.3
Cookie: __lc.visitor_id.1429232=S1344265841.a17c48bcd2; __utma=76048983.1312207238.131
9314128.1350055043.1350265445.180; __utmb=76048983.4.10.1350265445; __utmc=76048983; _
_utmz=76048983.1349710511.176.52.utmcsr=google|utmccn=(organic)|utmcmd=organic|utmctr=
(not%20provided)
```

Buscar la información de una solicitud http

Zheng, G. & Peltsverger S. (2015) *Web Analytics Overview*, In book: *Encyclopedia of Information Science and Technology*, Third Edition, Publisher: IGI Global, Editors: Mehdi Khosrow-Pour

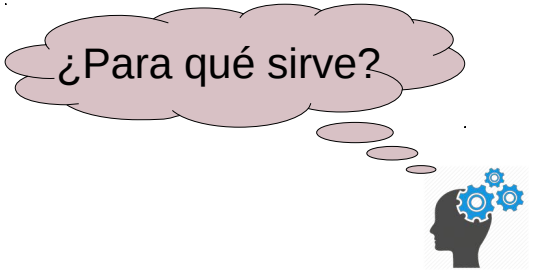
Solicitudes HTTP directas

- Formato

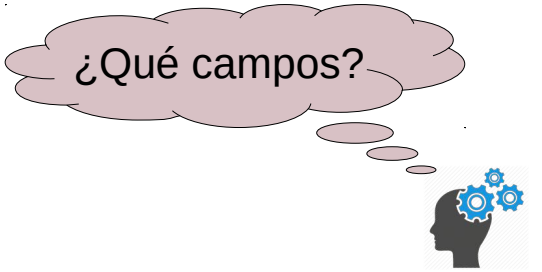
Request command to get "www.spsu.edu/itdegrees"

```
GET /itdegrees/ HTTP/1.1
```

  - Comando
    - Información URI (unified resource identifier)
      - Domino del host o IP y una ruta
      - Permite contabilizar el número de visitas



  - Cabecera
    - Pares de elementos (campo, valor)
    - Cada par proporciona una información



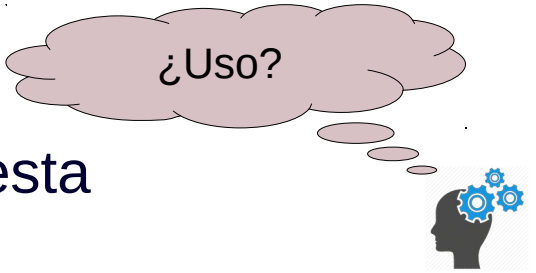
Solicitudes HTTP directas

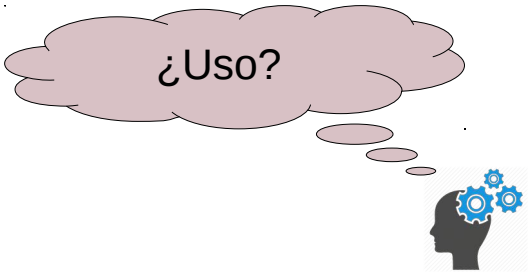
- Campos de la cabecera
  - *User-Agent*
    - Información sobre tipo de navegador y SO
    - Se utiliza para conocer el perfil tecnológico del cliente

  - *Referer*
    - URL visitada anteriormente (desde la que se llega)
    - Se utiliza para
      - Análisis de clicks
      - Métricas tasa de entradas y salidas

Solicitudes HTTP directas

- Campos de la cabecera
  - *Accept-Language*
    - Lista de lenguajes indicados para la respuesta
    - Se basa en los *locate* del SO
    - Se utiliza para conocer el lenguaje del usuario
  - Cookie
    - Información a nivel de aplicación
    - Almacenada en el lado del cliente
      - Acciones del ratón, teclado
    - Se utiliza para calcular algunos eventos

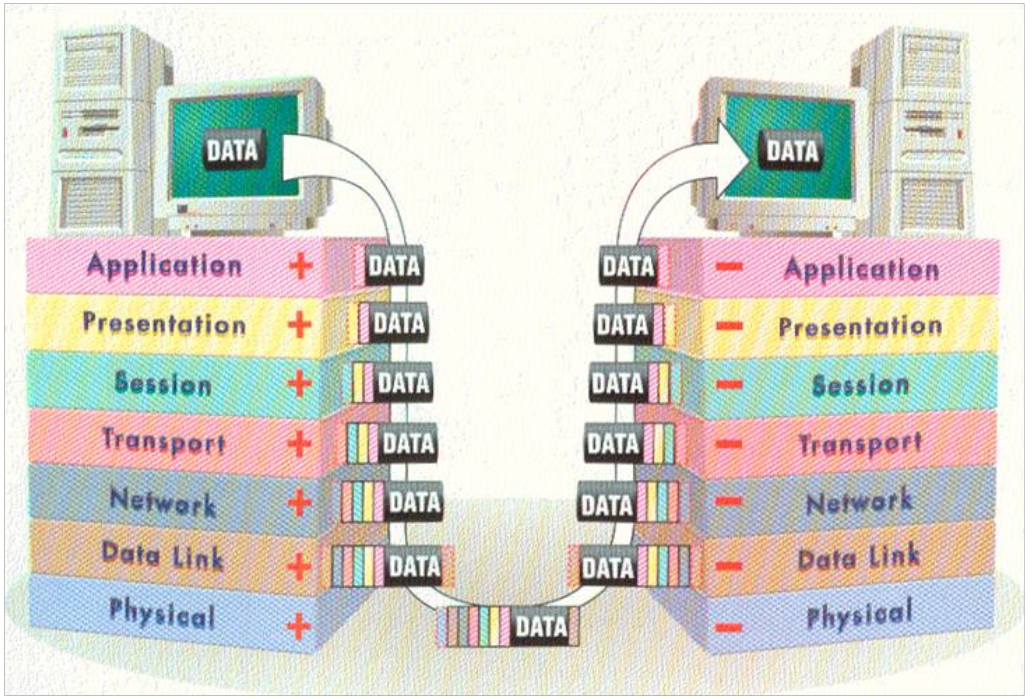




Identifica los campos anteriores en una solicitud http

Datos a nivel de red y generados por el servidor

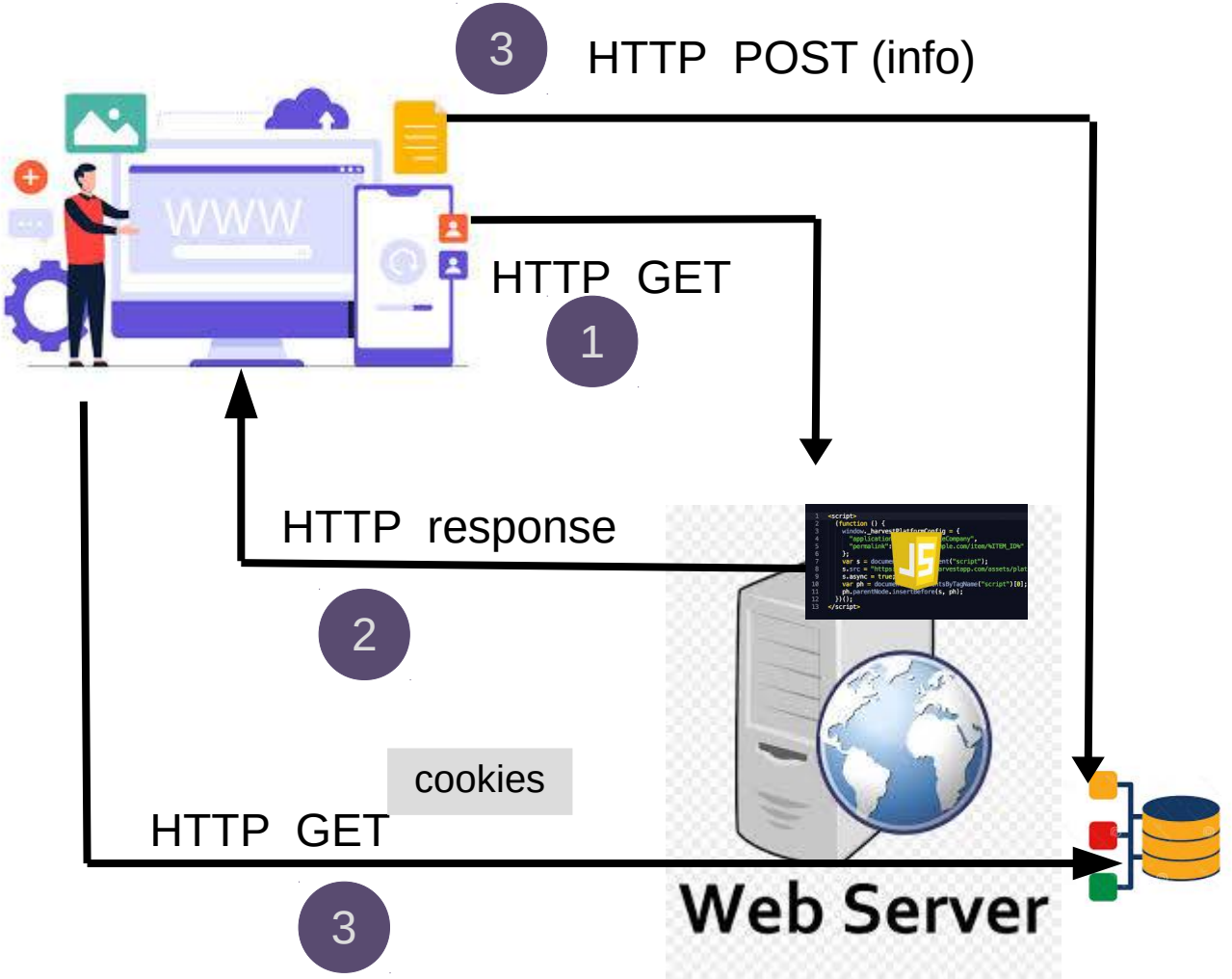
HTTP GET



Datos a nivel de red y generados por el servidor

- Datos a nivel de red
  - No forman parte de la solicitud HTTP
  - Necesarios para que la transmisión tenga éxito
  - Dirección IP del solicitante
    - Se necesita junto al puerto para devolver la respuesta
    - Se manda a nivel TCP/IP
- Datos generados por el servidor
  - Se usan para referencias internas y se almacenan en ficheros de log.
    - Tamaño fichero, tiempo procesamiento, IP servidor, ...

Datos a nivel de aplicación

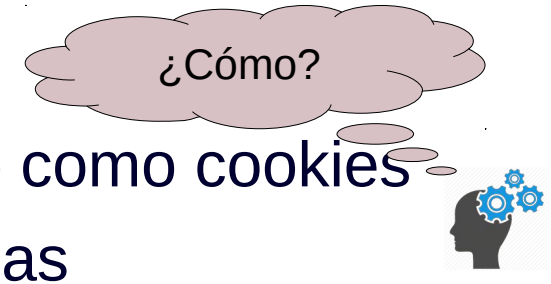


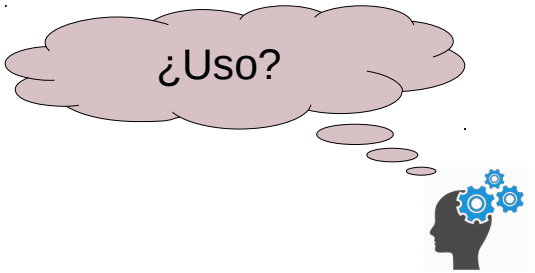
Datos a nivel de aplicación

- ¿Cuáles son?
  - Datos generados y procesados por programas a nivel de aplicación
    - JavaScript, PHP, ASP.Net
  - Capturados por registros internos, no por servicios públicos de análisis web

Datos a nivel de aplicación

- Ejemplos
  - Datos de sesiones
    - Identifican la interacción de un usuario con un sitio web
      - Peticiones relacionadas a una unidad de contenido concreto
    - HTTP no tiene estado → no información de sesión
    - Se gestionan a nivel de aplicación
    - Se envían como parámetros de la URL o como cookies
    - Importantes para calcular algunas métricas
      - Número de visitas, duración de la visita, número de páginas por visita.





Datos a nivel de aplicación

- Ejemplos
  - Datos de referencia (*referral data*)
    - Diferente del referer de la solicitud HTTP
      - Suele ser una URL
    - Representa diferentes fuentes que conducen al recurso web y suele ser un valor codificado
    - Uso
      - Analizar el nivel de tráfico de las diferentes fuentes
      - Medir la efectividad del canal en el seguimiento de publicidad

Datos a nivel de aplicación

- Ejemplos
  - Datos de acciones del usuario
    - Todas las acciones del teclado y ratón
      - Términos de búsqueda, coordenadas y movimientos del ratón
    - Acciones específicas de las aplicación
      - Votar, Reproducir un vídeo o un audio, marcar como favorito
  - Datos del lado del cliente/navegador
    - Información sobre el estado del ordenador
      - Resolución y profundidad de colores

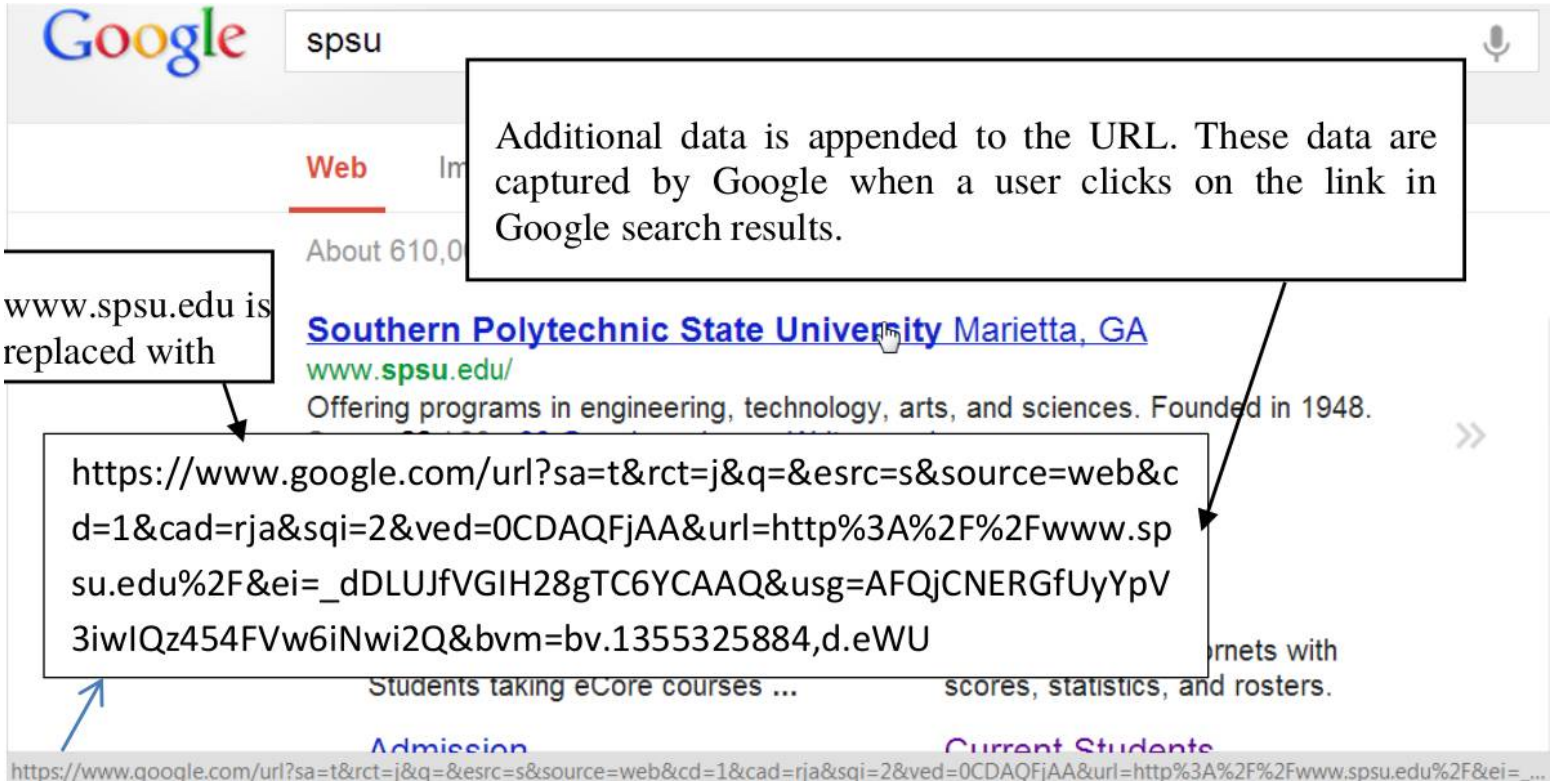
Datos a nivel de aplicación

- ¿Donde los encontramos?
  - Integrados en la solicitud HTTP
    - Añadido a la URL de solicitud como parámetros
      - Los programas del lado del cliente pueden analizar estos parámetros
      - URL creadas por Google para redirigir a los usuarios
  - Como cookies de la cabecera de la solicitud HTTP
  - En el cuerpo de la solicitud HTTP
    - Cuando se utiliza el método POST

Identifica la URL creada por Google en un resultado de búsqueda

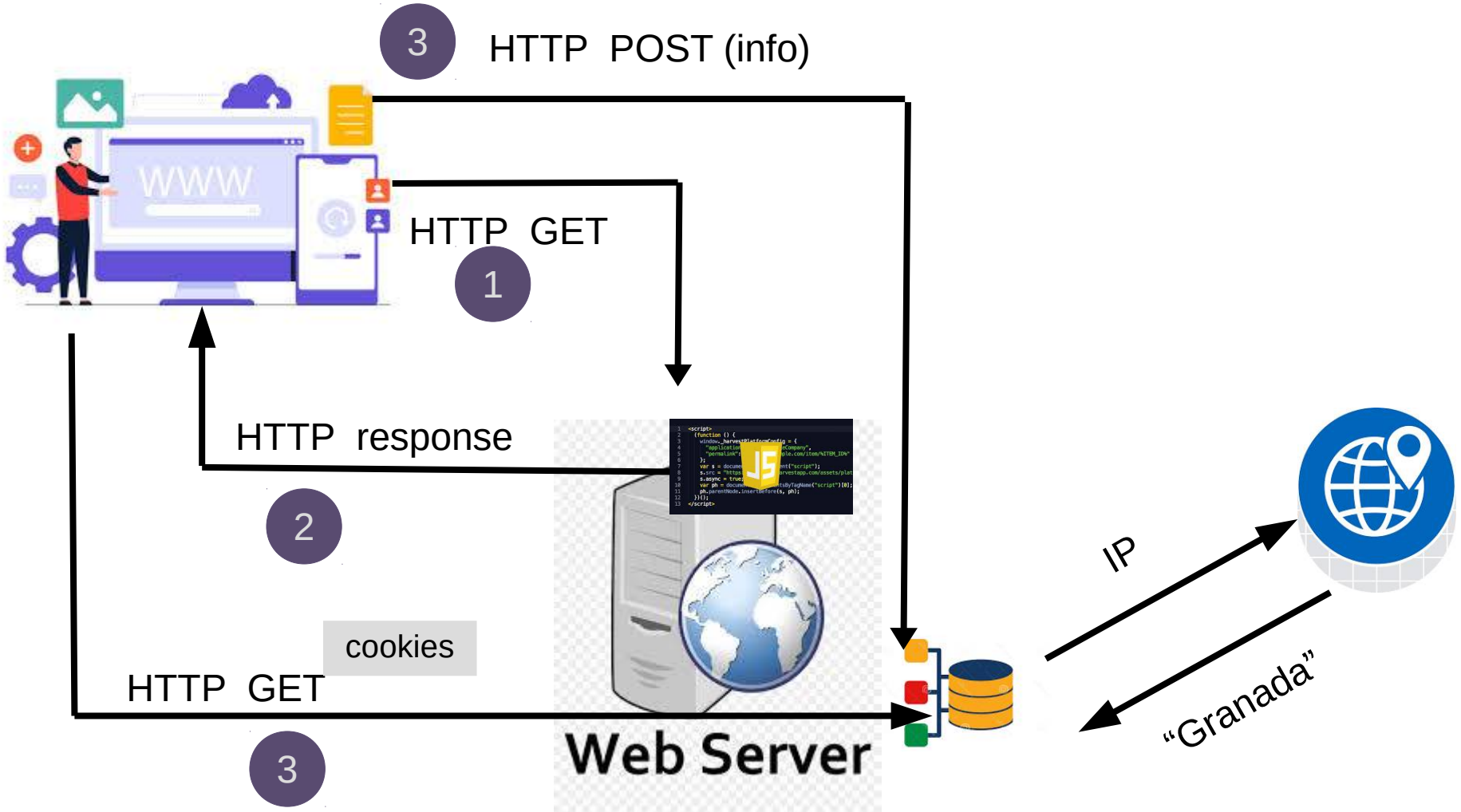
Datos a nivel de aplicación

- ¿Donde los encontramos?



Zheng, G. & Peltsverger S. (2015) *Web Analytics Overview*, In book: *Encyclopedia of Information Science and Technology*, Third Edition, Publisher: IGI Global, Editors: Mehdi Khosrow-Pour

Datos Externos



Datos Externos

- Datos externos a las solicitudes HTTP
  - Se combinan con los del sitio web
    - Mejorar los datos de comportamiento e interpretar el uso de la web
  - Los proporcionan bases de datos o servicios a terceros
    - GeoLite de MaxMind (http://www.maxmind.com),
    - IPInfoDB (http://ipinfodb.com),
    - GeoBytes (http://www.geobytes.com)
  - Ejemplos
    - Regiones geográficas o Proveedores servicios
    - Información del usuario proceso de registro
      - Información de identidad en un sesión
    - Términos de búsqueda

Resumen

- Los datos que se utilizan para analizar el comportamiento y el uso de la web provienen de:
  - Solicitudes HTTP
    - De la petición directa
    - Asociados a la misma (datos a nivel de red)
    - Enviados con la misma (datos a nivel de aplicación)
  - Fuentes externas
    - Combinarlos con los anteriores para mejorarlos