Webminimg

una breve introducción y la perspectiva Java

M. D. López De Luise

IEEE ARGENTINA

UP | Universidad de Palermo
Facultad de Ingeniería

Próximas charlas

Análisis y auditoría de performance en internet	(A. Popovski)	22/09/05 (19:00hs)
Filtro semántico e IR	(N. Di Tada)	06/10/05 (19:00hs)
NN con Java	(D. López De Luise)	20/10/05 (19:00hs)
Java y el XML	(D. López De Luise)	27/10/05 (19:00hs)
La IA en Java	(D. López De Luise)	17/11/05 (19:00hs)
Cognitive Memory	(S. Piedrahita)	21/11/05 (19:00hs)
Webbrowsing con Java	(D. López De Luise)	24/11/05 (19:00hs)

informes:

upgrade@palermo.edu ***sec.argentina@ieee.org***

Objetivo

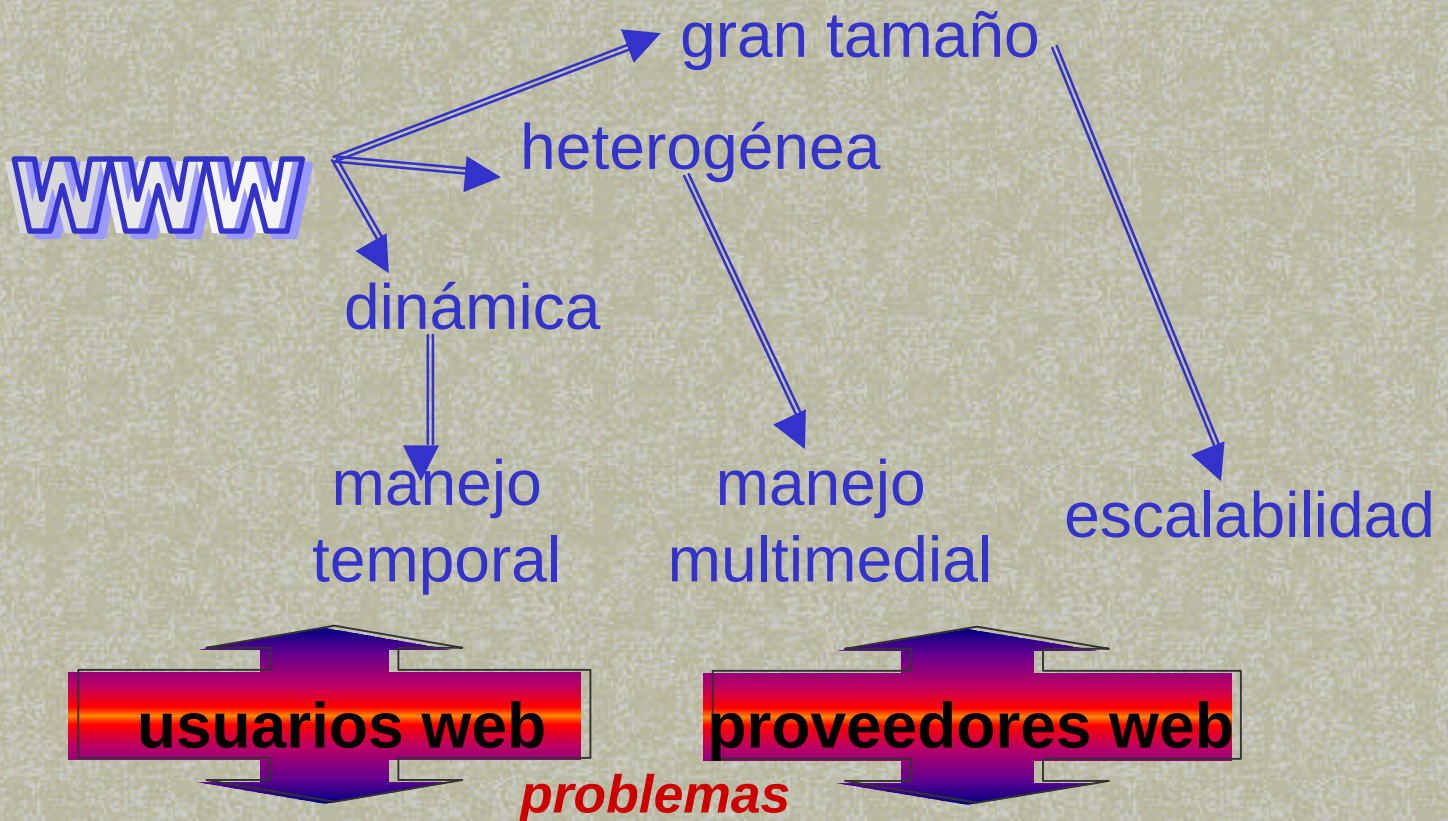
Presentar los conceptos fundamentales de Webmining y su aplicación desde la perspectiva del lenguaje Java.

Temario

- Introducción al webmining.
- Problemas a resolver del WWW.
- WM como una solución posible.
- Incumbencias y tipos de WM.
- Diferencias con IR, IE y ML.
- Relación con Agentes.
- Por qué WM y Java.
- Java en WM: VWV y C-BIRD.
- Java y metadatos.

El Webmining

WWW sus problemas



problemas

1. Hallar información relevante
2. Aprender acerca de consumidores o usuarios individuales
3. Personalización de la información
4. Crear información a partir de la almacenada

1 Hallar información relevante

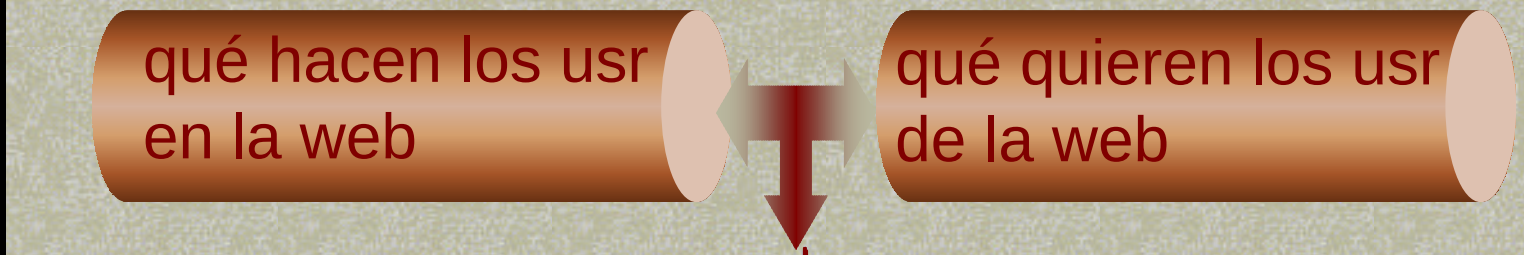
buscadores

- escasa precisión (*precision*) → inf. poco relevante
- escaso alcance (*recall*) → imposible indexar todo

query triggered problem

retrieval oriented problem

2 Aprender acerca de consumidores o usuarios individuales



determinar qué inf. debe adaptarse a un grupo

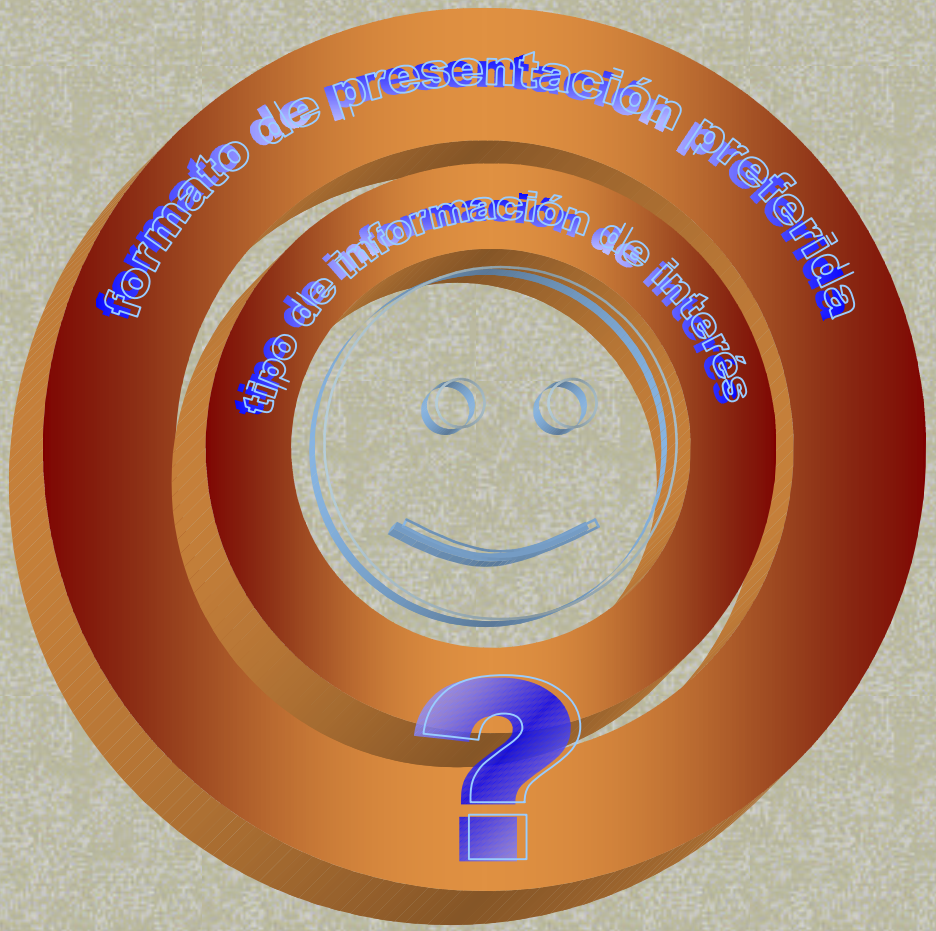
determinar qué inf. debe adaptarse a un usr

diseño eficiente de los websites

administración eficiente de los websites

uso de la inf. del usr para marketing

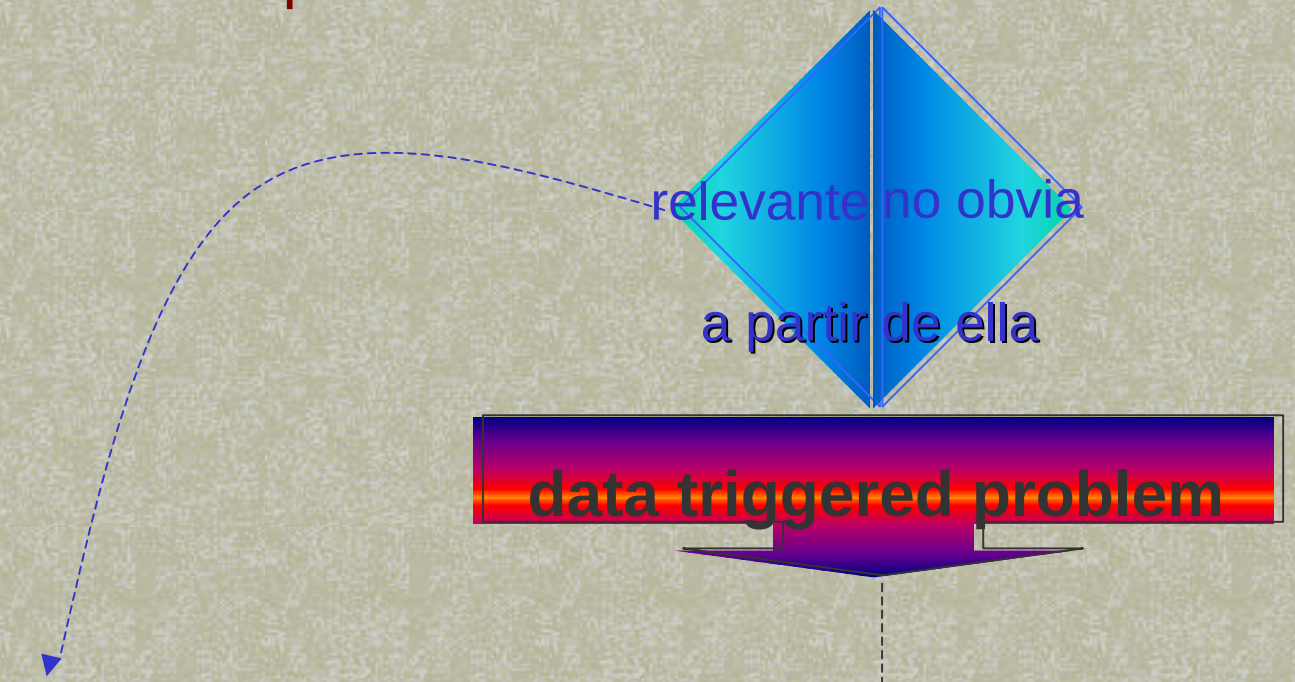
3 Personalización de la información



4 Crear información a partir de la almacenada

dada una colección de datos en la web

el problema es sacar información



M. D. López De Luise

El Webmining:

- como una de las soluciones
- incumbencias
- tipos
- diferencias con R, E, ML
- relación con Agentes

El Webmining:

como una solución

El Webmining: como una de las soluciones

DB

IR

WM

NLP

El Webmining:

incumbencias

El Webmining: incumbencias

definición

aplicación de técnicas de **DM**

técnicas que permiten descrubrir inf. → *desconocida*, *no evidente*, *relevante*

para → ***descubrir inf***, ***extraer inf.***

desde → ***docs. Web***, ***servs. Web***

BD, **newsgroup**, **newsletter**, **doc.**, **html** → **off line**, **on line**

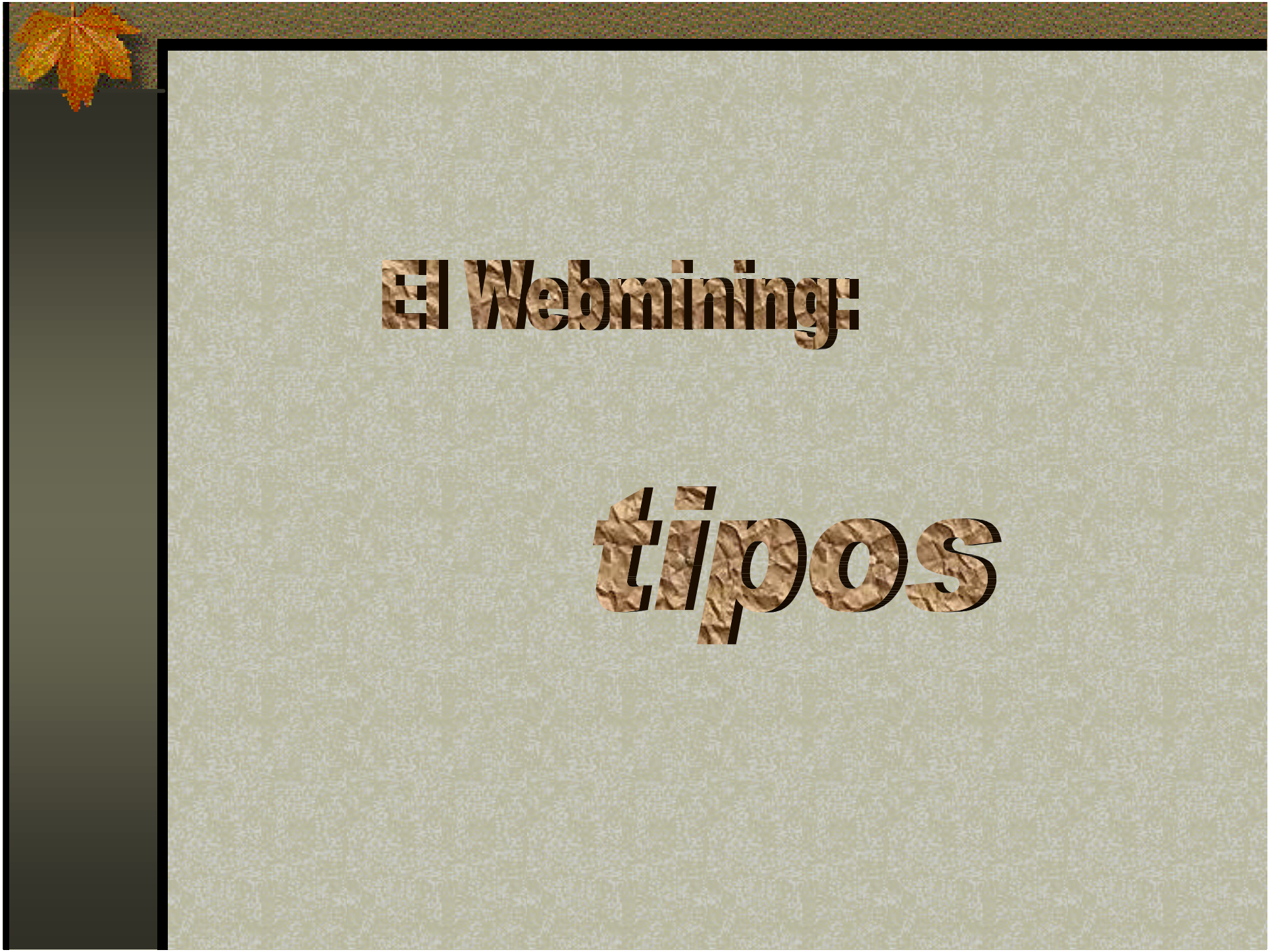
tareas (*Etzioni*)

detección de los recursos potencialmente útiles

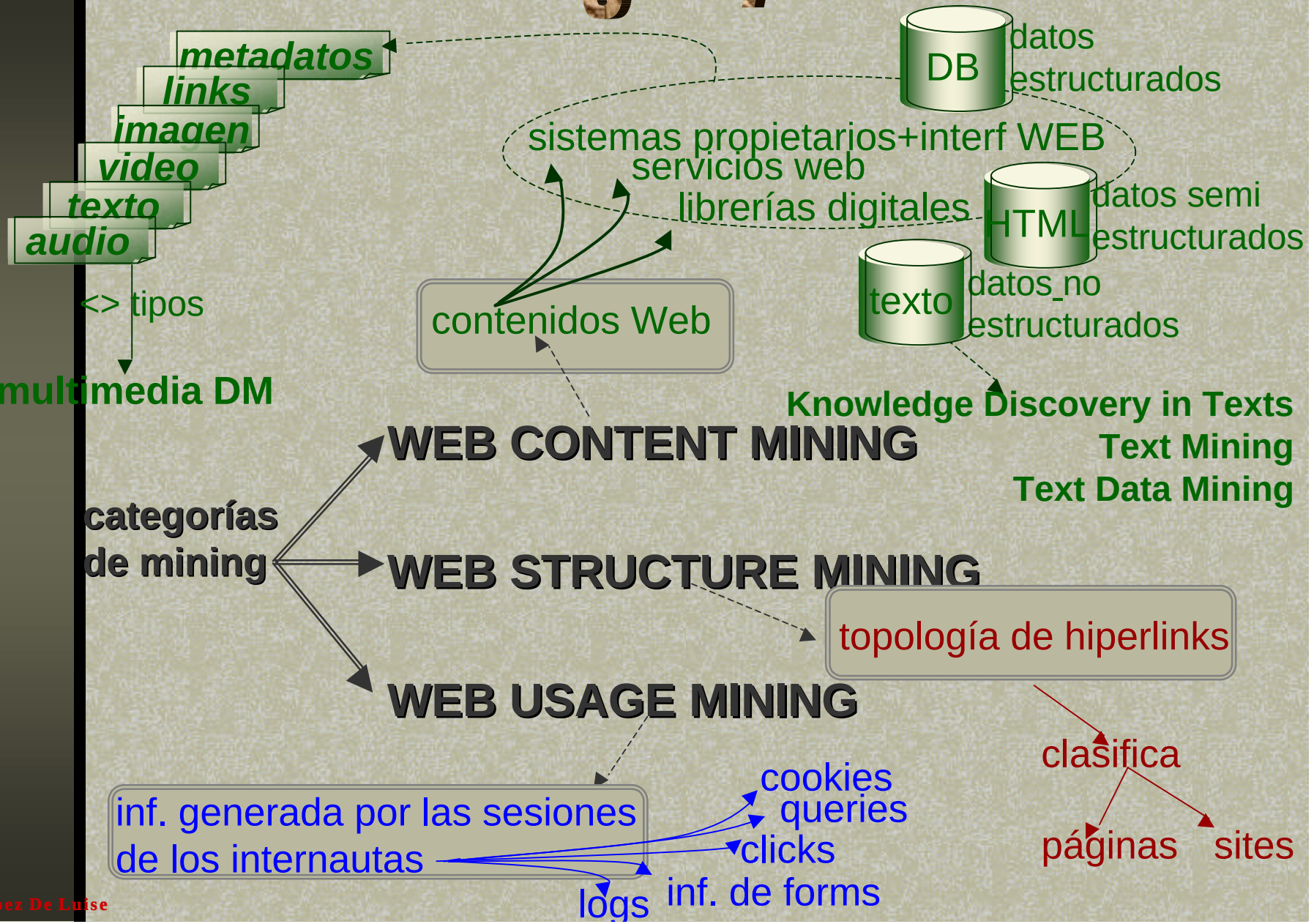
selección automática y preproceso de esos recursos

detección automática de patrones(***en el mismo site o entre sites***)

validación e interpretación de los patrones


El Webmining:
tipos

El Webmining: tipos



El Webmining:

diferencias con IR IE ML

El Webmining: relación con IR

WM Vs **Information Retrieval**

extracción de inf.
no evidente + relevante
a partir de datos

recuperación automática de inf.
procurando
máx. cantidad docs relevantes
+
min. cantidad docs no relevantes

tareas

- clasificación
- categorización
- filtrado de información
- formalización de patrones
- …

tareas

- clasificación
- categorización
- filtrado de información
- indexación
- interfaces usr

El Webmining: relación con IE

WM Vs **Information Extraction**

extracción de inf.
no evidente + relevante
a partir de datos

transforma un conj de datos
(a veces con IR)
simplificando su interpretación

tareas

detecta patrones relevantes	detecta hechos relevantes
procesos automáticos	procesos automáticos o no
escalable	a veces no escalable

El Webmining: relación con ML.

WM Vs Machine Learning

extracción de inf.
no evidente + relevante
a partir de datos

usa aprendizaje automático
con distintos fines
(ej.optimización, filtrados, etc.)

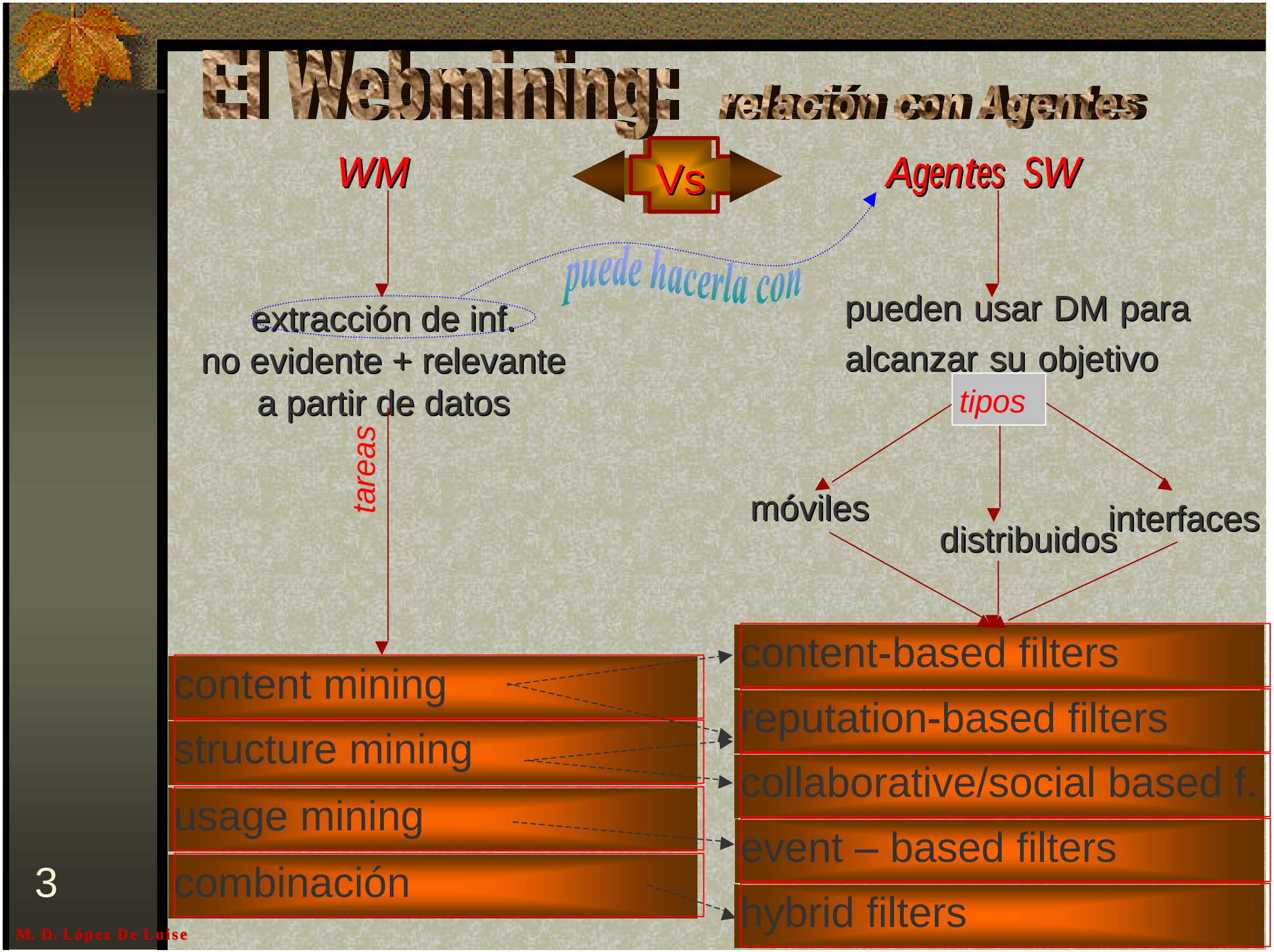
tareas

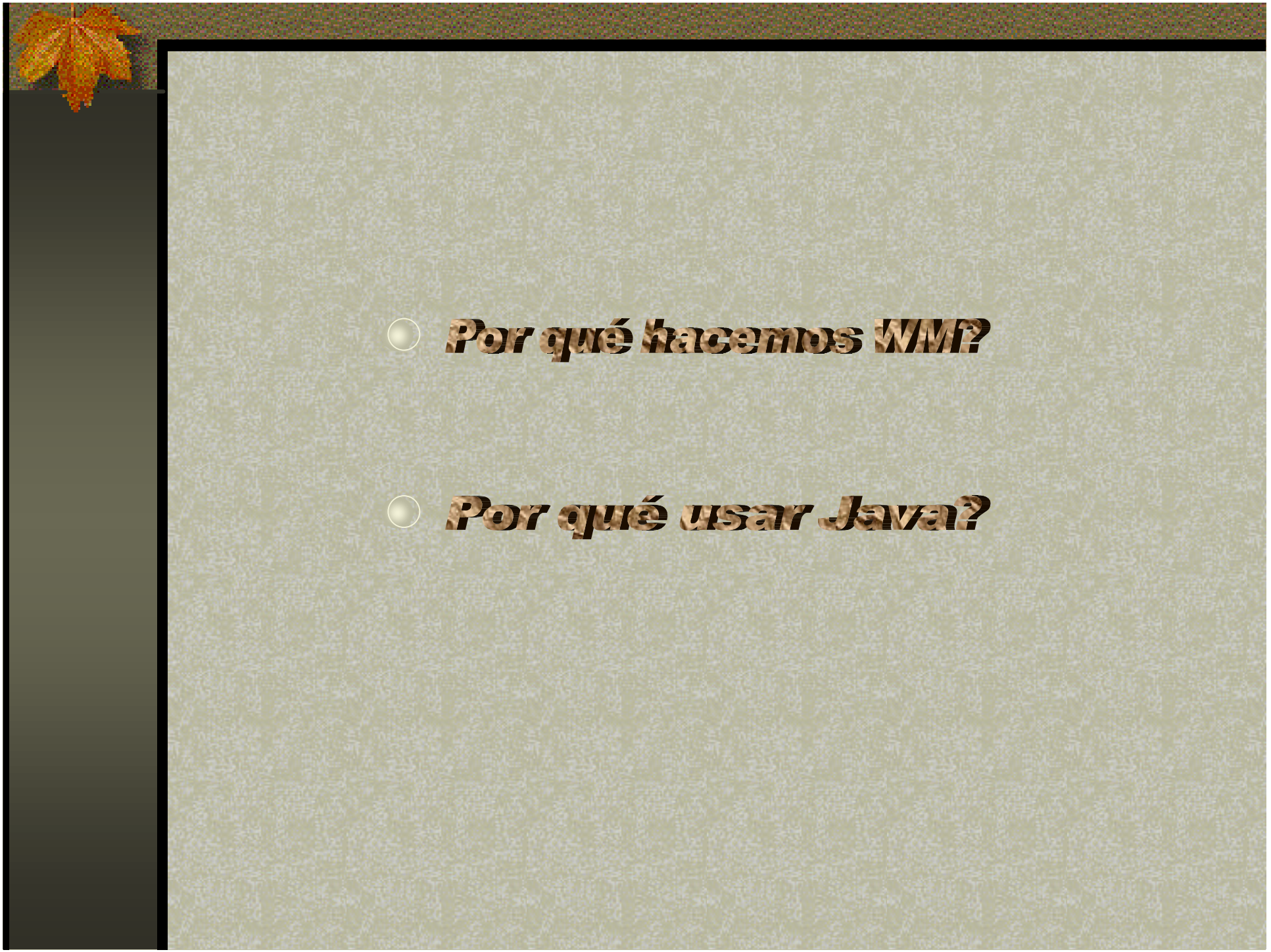
optimización del proceso
de clasificación en mining
de textos

El Webmining:

relación con Agentes

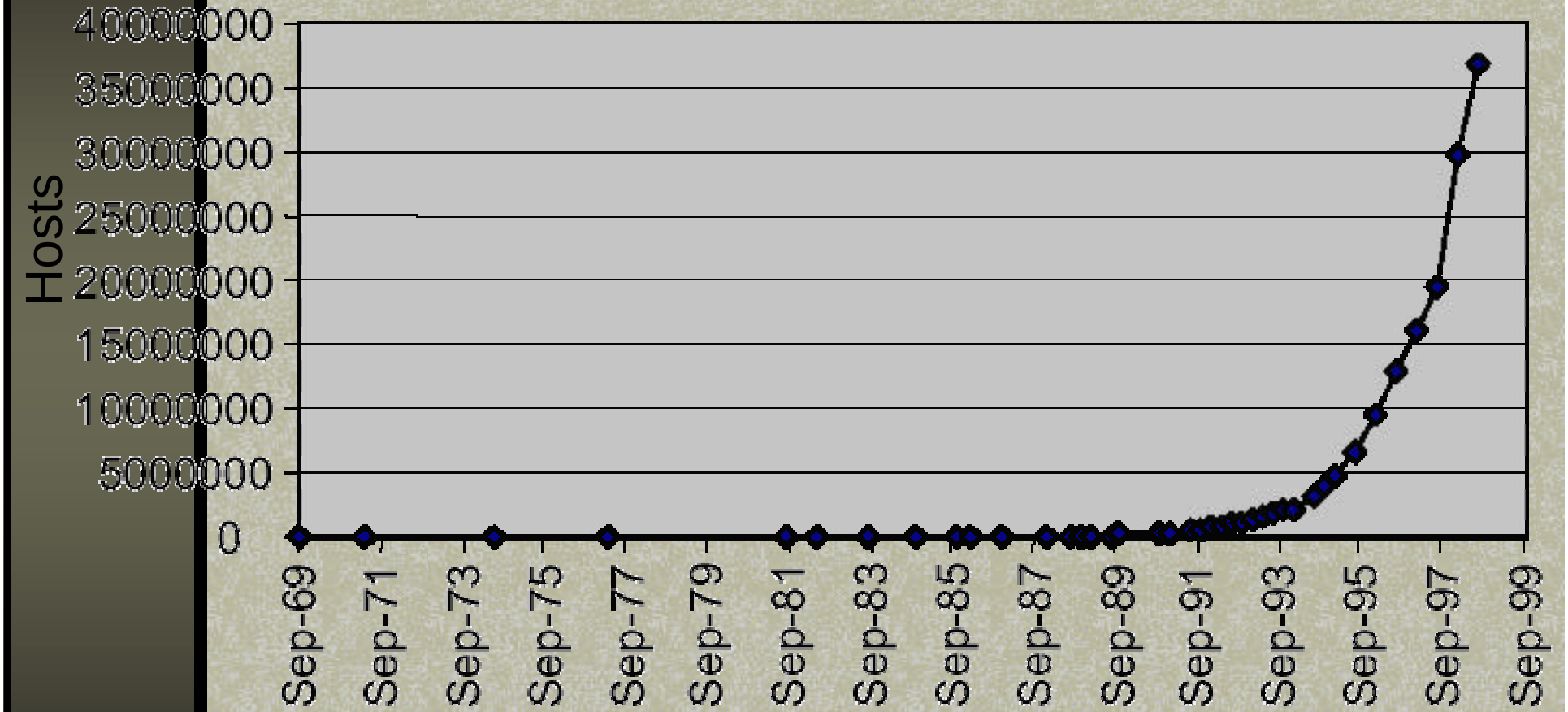
El Webmining: relación con Agentes




Por qué hacemos WM?
Por qué usar Java?

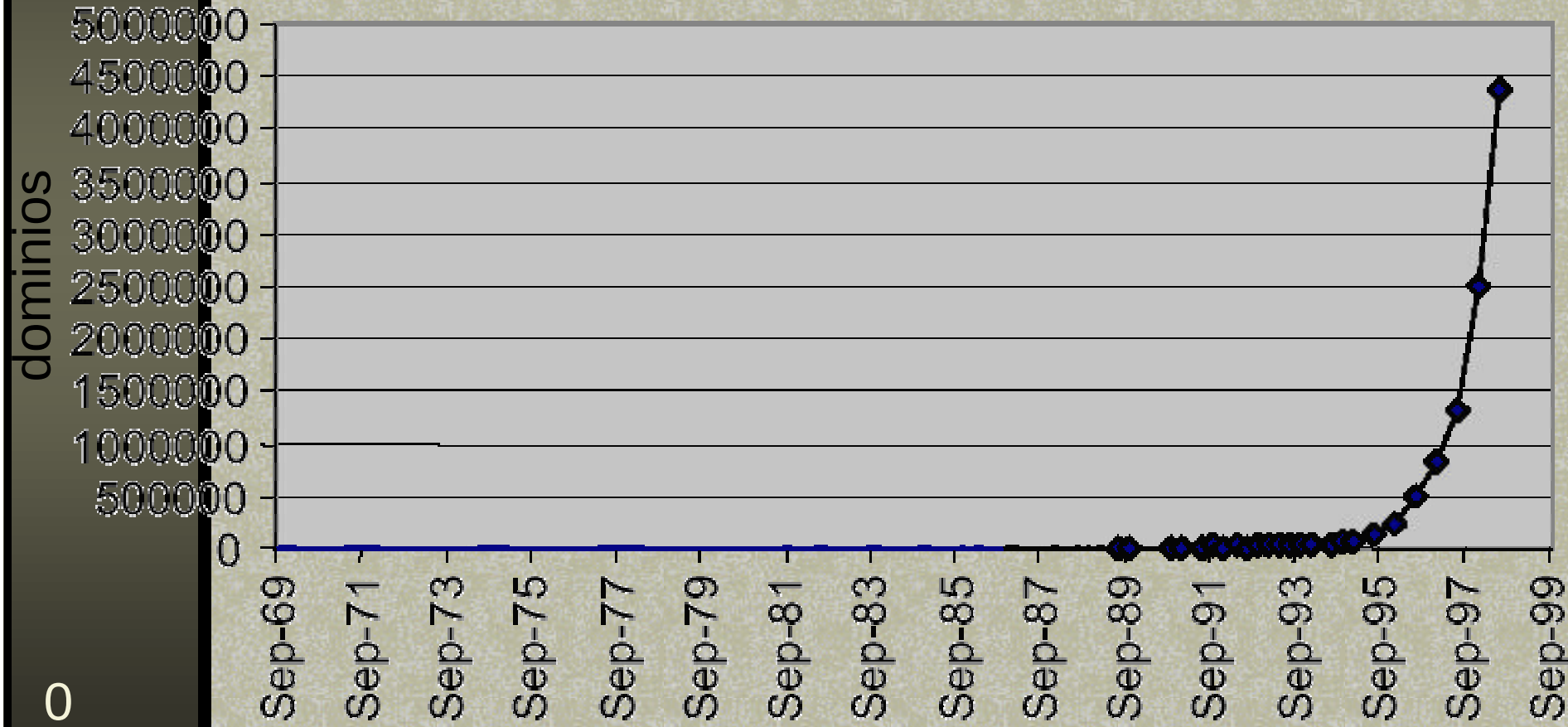
Por qué WMI?

a.Crecimiento físico de Internet



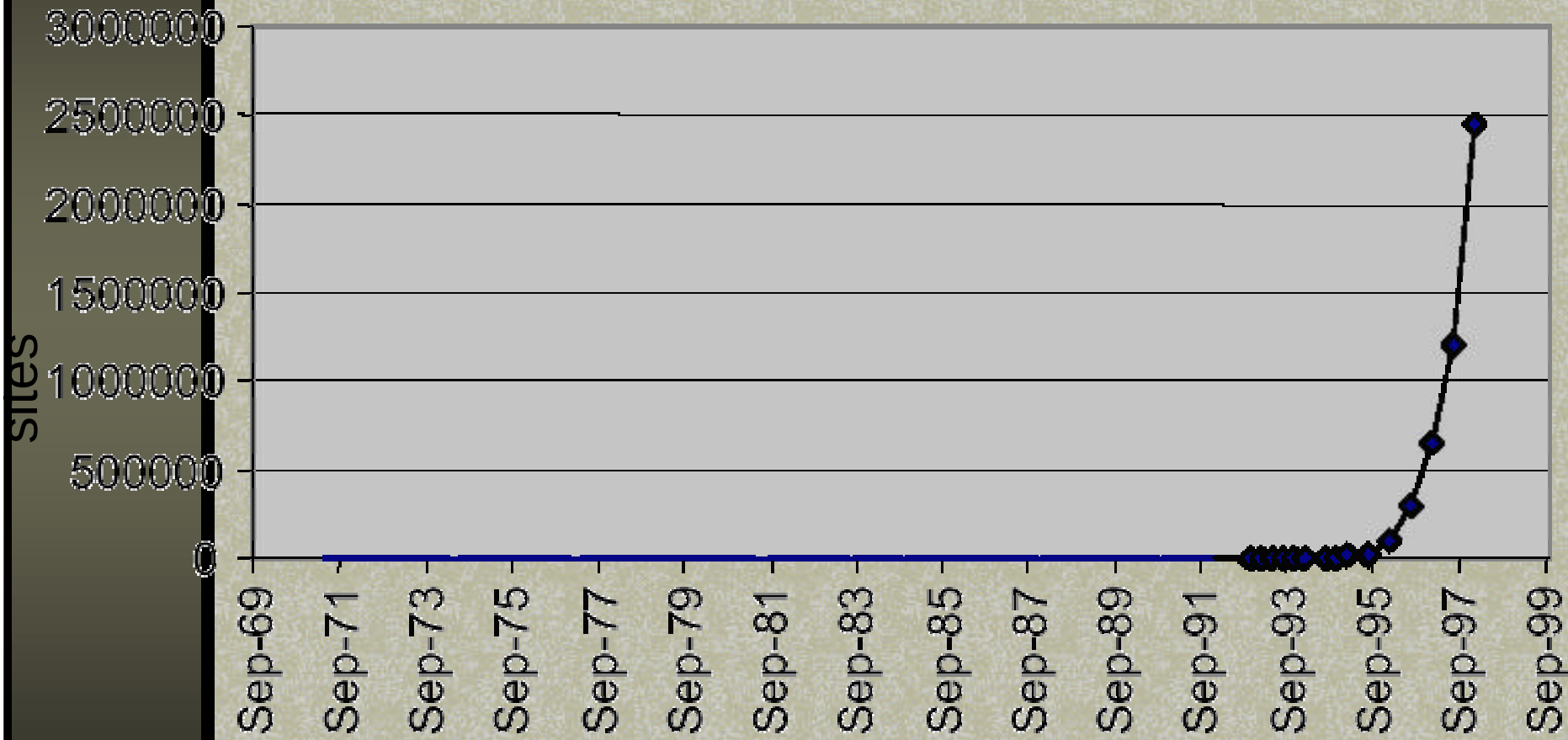
Por qué WM?

b.Crecimiento de cantidad de dominios



Por qué WMI?

c.Crecimiento en cantidad de sites



M. D. López De Luise

Por qué Java?

a.tecnología reciente

1969
inicia proyecto
Arpanet

1972
demostración
Arpanet

1973
primeras conex.
internacionales

1974
TCP/IP

1979
Usenet

1981
Bitnet
CSnet

1982
Arpanet
migra a TCP/IP

1983
Arpanet se divide
en Arpanet y
Milnet

1985
ftp sobre TCP

1986
NNTP
NSF-NET

1988
IRC

1990
fin de
Arpanet

1990
Archie

1991
Gopher
WAIS
NetFind

1992
Veronica
MBONE

1993
MOSAIC
Crawlers
W3C
Aliweb

1994
Harvest
E-Commerce
Yahoo
MLDB+WebQL
UCSTRI

1995
Java

1996
Internet phone
Altavista
WebSQL
WebLog

1997
Wireless Internet access
WebOQL

1998
IBM Clever
Google

1999
internet2
RSVP

M. D. López De Luise

Por qué Java?

b.tecnología flexible

- Diseñado para resolver problemas comunes de los programadores *profesionales* y mejorar su productividad
- Solución dentro del propio *espacio* de solución
  =>una línea de código puede hacer muchas cosas
- Soluciones más *sencillas* y leíbles por un entendido
  =>mejor mantenibilidad
- Uso sencillo y flexible de *librerías*
  (mejor programa ??mejor uso de librerías)
  =>el pgmador se desentiende de tareas menores
  (ej. inicialización y cleanup)

Por qué Java?

c.tecnología trasportable

- manejo localizado de *errores*
  => fácil detección del código involucrado en su manejo

- Java fue diseñado con la complejidad necesaria para ser *escalable*.

- Java fue diseñado para vivir dentro y fuera de internet sin grandes *cambios*


WM + JAVA
Algunos casos

JAVA + WMF:

- el caso C-BIRD
- el caso VWV
- el uso de metadatos

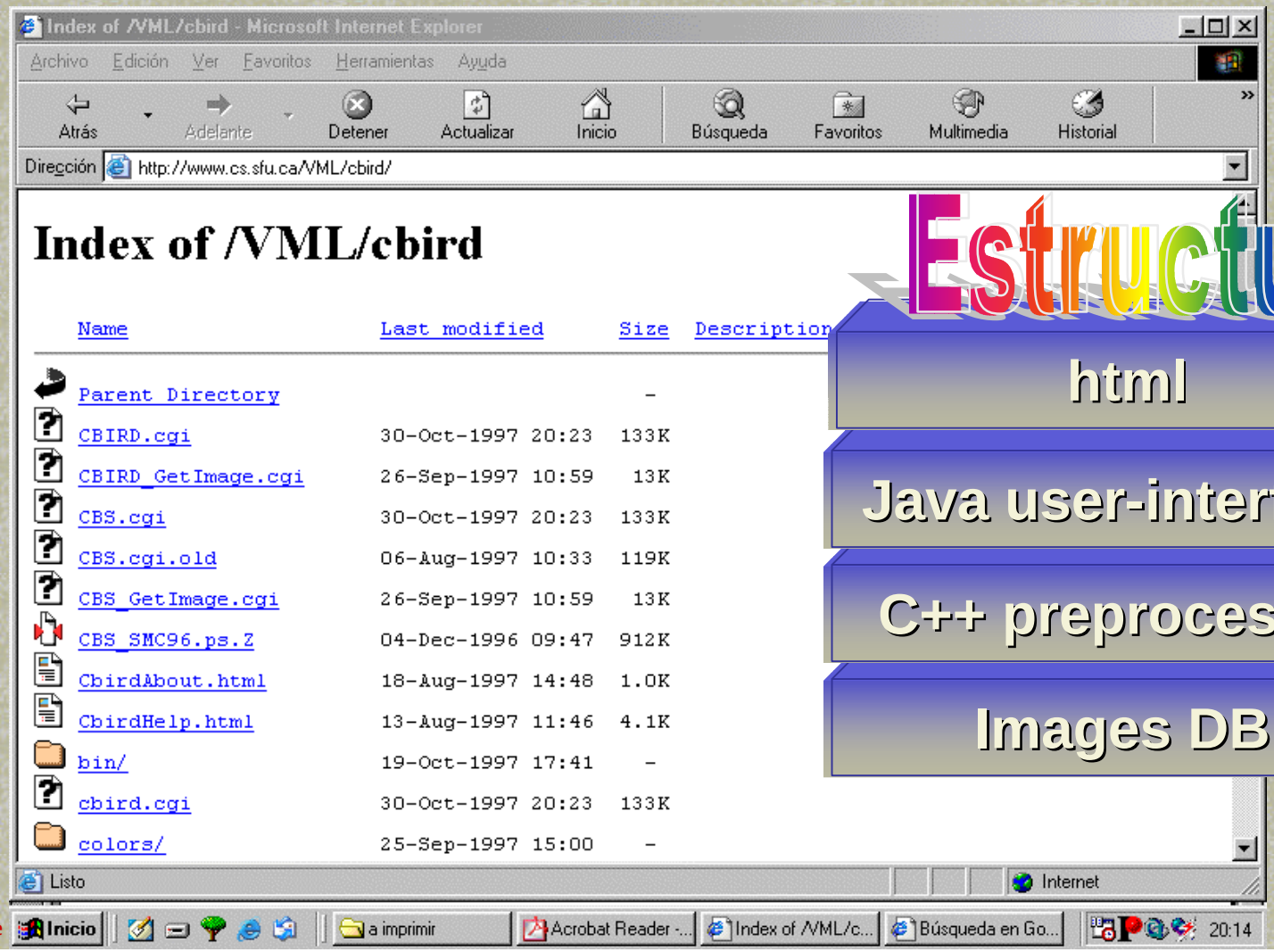
JAVA + WM:

C-BIRD

JAVA + WML: el caso C-BIRD

http://jupiter.cs.sfu.ca/cbird/java/ (IE 4.0 or Netscape 4.0).

Web content mining s/imágenes	C-BIRD	IR

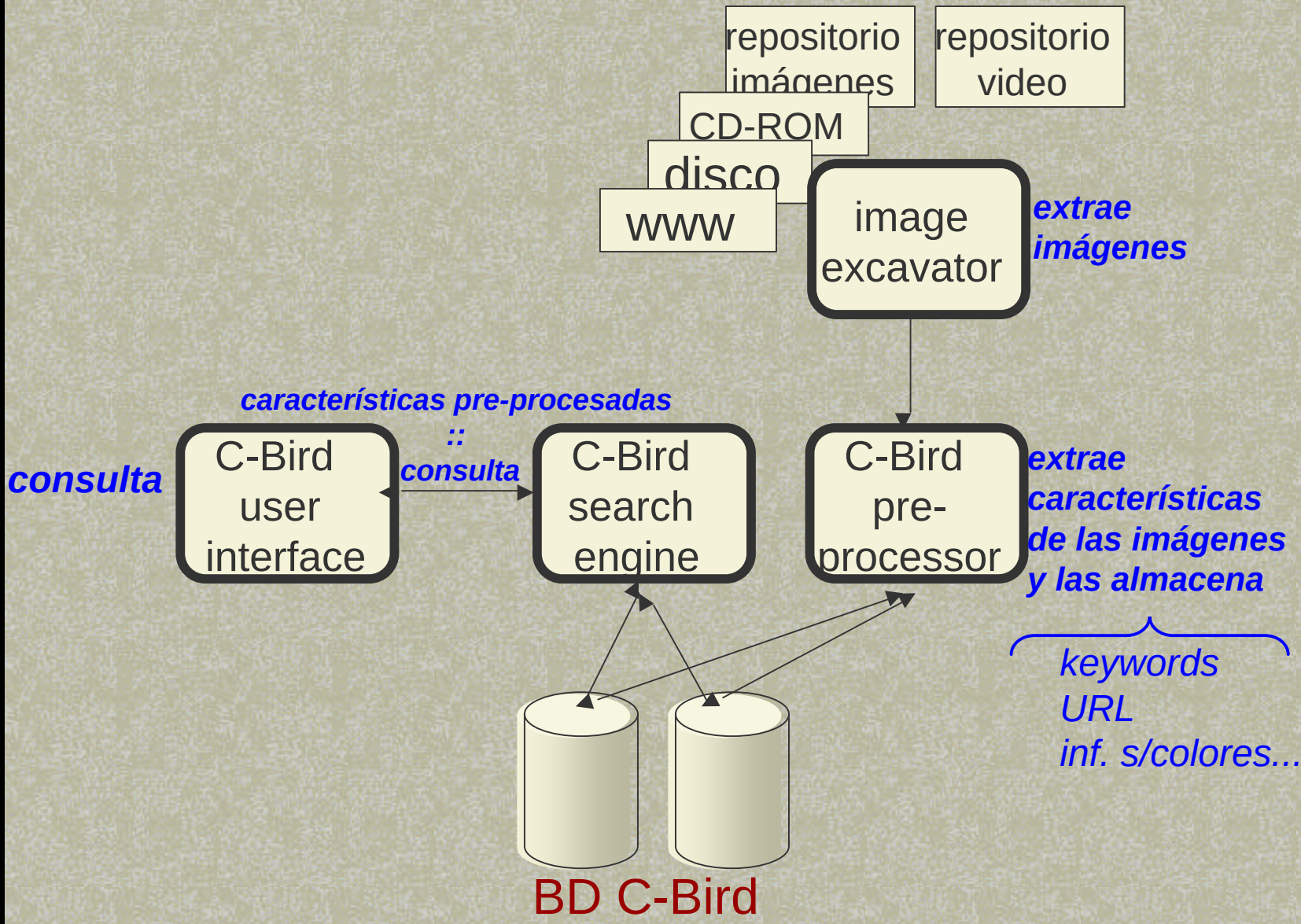


Estructura

- html
- Java user-interface
- C++ preprocessor
- Images DB

M. D. López De Luise

JAVA + WWW: el caso C-BIRD



JAVA + WM: el caso C-BIRD

interface usuario

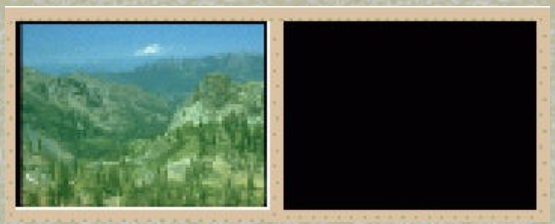
tipos de busqueda

- búsqueda por histogramas de colores
- búsqueda por color layout *(1x1, 2x2, 3x3, 4x4, 8x8)*
- búsqueda por porcentaje (?5 colores)
- búsqueda por ?/? de metadatos
- búsqueda por illumination invariance (cromaticidad) (img normalizada)
- combinaciones y otras

búsqueda por color layout

JAVA + WM: el caso C-BIRD

búsqueda con características de la imagen



búsqueda con características de la imagen y keywords sacadas del contexto

JAVA + WM:

WWW

JAVA + WM: el caso VWV

Information browsing	Virtual Web View	Resource discovering

(Osmar Zaiane et al.)

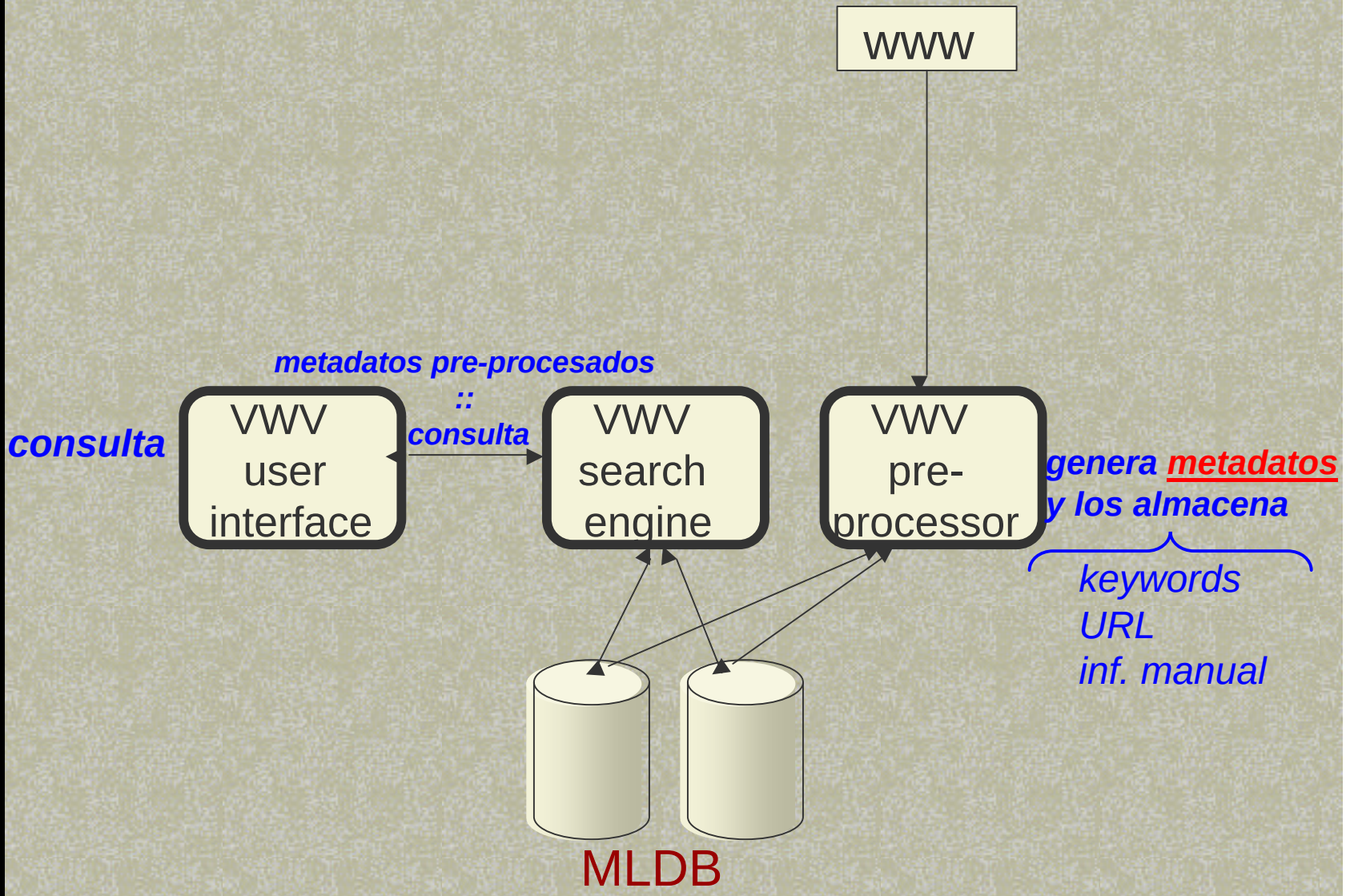
query language

Web Modeling Language
aplicación del XML
notación para especificar
conceptualmente Web sites complejos

def

capas de abstracción sucesivas
datos relativamente estructurados,
manejables y administrable por DBs

JAVA + WM: el caso VWV



JAVA + WM:

METADATOS

JAVA + XML: metadatos en WWW

keywords

metadatos *def: son datos acerca de los datos.*

charla Java y el XML

XML + Java *def: lenguaje flexible del tipo "markup" (basado en tags)*

API XML

JAXP *def: API p/procesar XML*

Interpreta y presenta XML

Indep de cualquier procesador XML

def: verifica si el XML está bien formado

parser XML

SAX DOM

JXerces *basado en*

Document **O**bject **M**odel

representación XML estándar de contenidos y modelos de documentos *(Bindings para JavaScript, Java, C++, etc.)*

Std **A**pplication **X**ML

administrador de docs XML basado en eventos
permite manejo secuencial sin necesidad de cargarlo en MC

M. D. López De Luise

JAVA + WWW:

metadatos en WWW
(WWW como web semántica)

def (Tim Berners-Lee): extensión de la web que permite a humanos y computadoras trabajar en cooperación

Web semántica

- Knowledge Acquisitions
- Knowledge Representations
- Agent Systems
- Multi Agent Systems
- Ontology

charla IA en Java
charla NN en Java

incorpora el paradigna de Soft-Computing

Machine Intelligence JSRs (Java Specification Request) en la JCP (Java Community Process)


FIN