Web Datamining

1 - Information Retrieval and basic concepts of mining the web

Guibert J. TCHINDE

November 2014 – ENSAI

PagesJaunes

Programme

Programme

- Concepts & Définitions

Programme

- Concepts & Définitions
- Indexation & Crawl

Programme

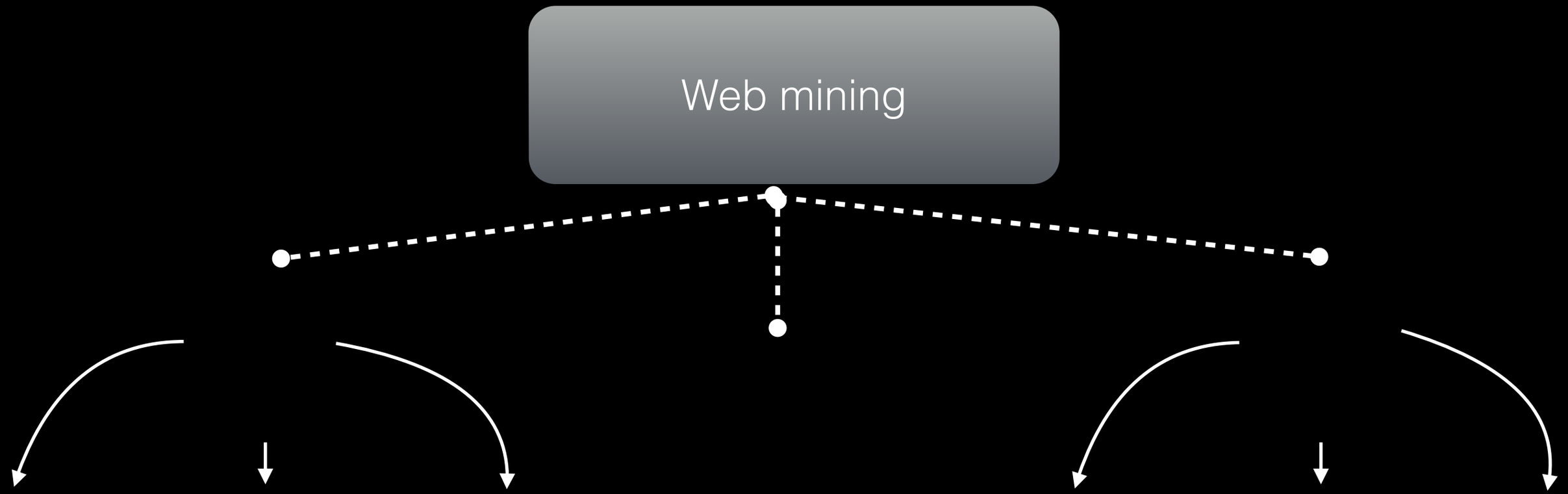
- Concepts & Définitions
- Indexation & Crawl
- Introduction aux techniques de webdatamining

Programme

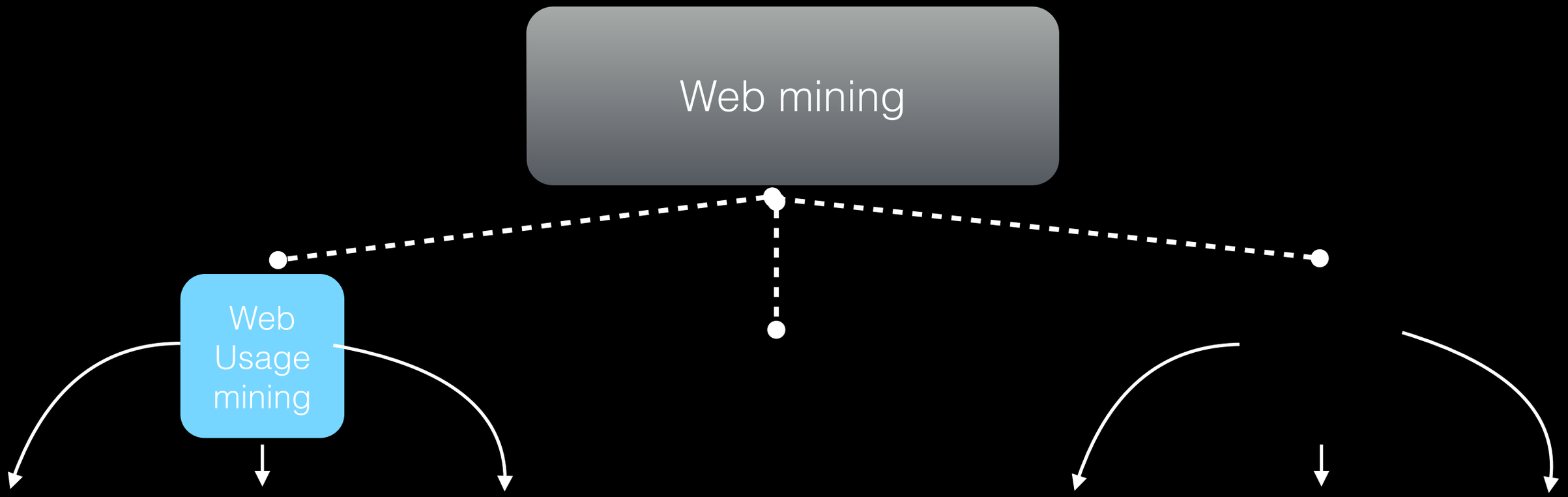
- Concepts & Définitions
- Indexation & Crawl
- Introduction aux techniques de webdatamining
- Applications a la classification des documents

PagesJaunes

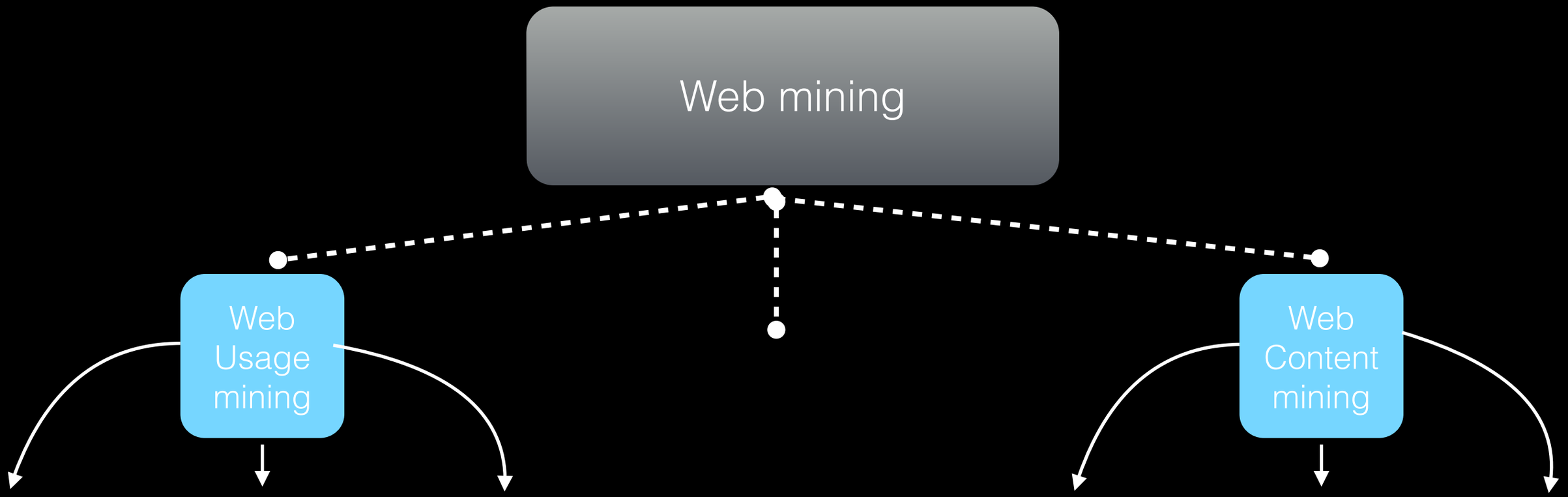
Pour rappel



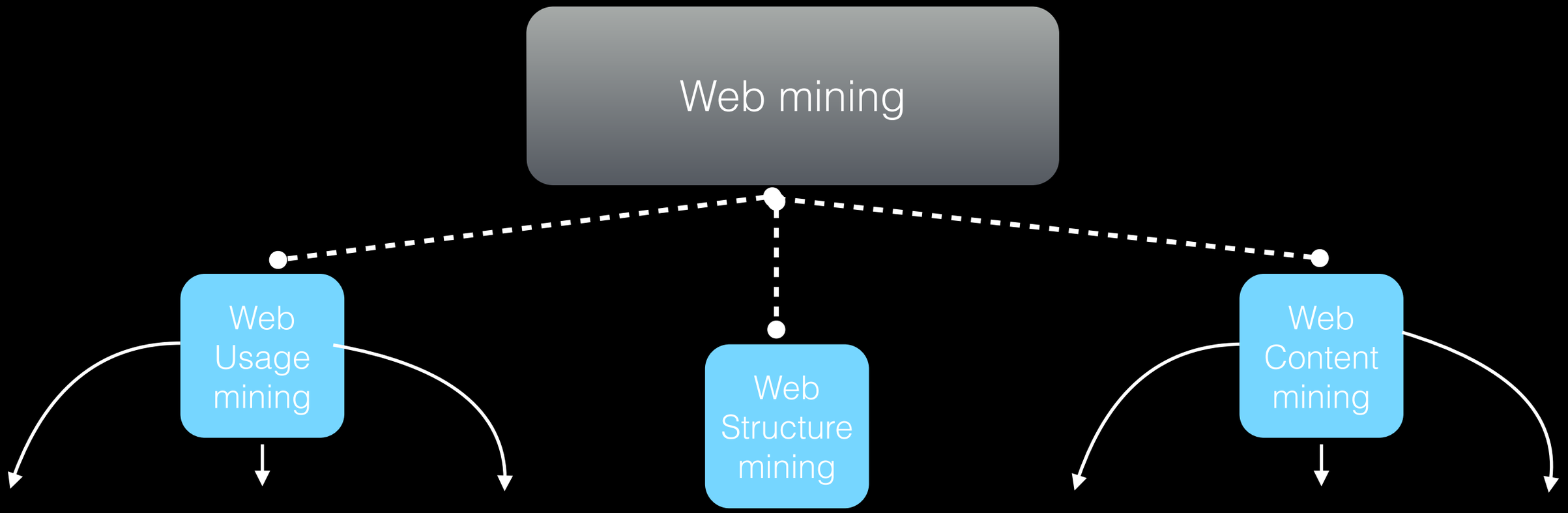
Pour rappel



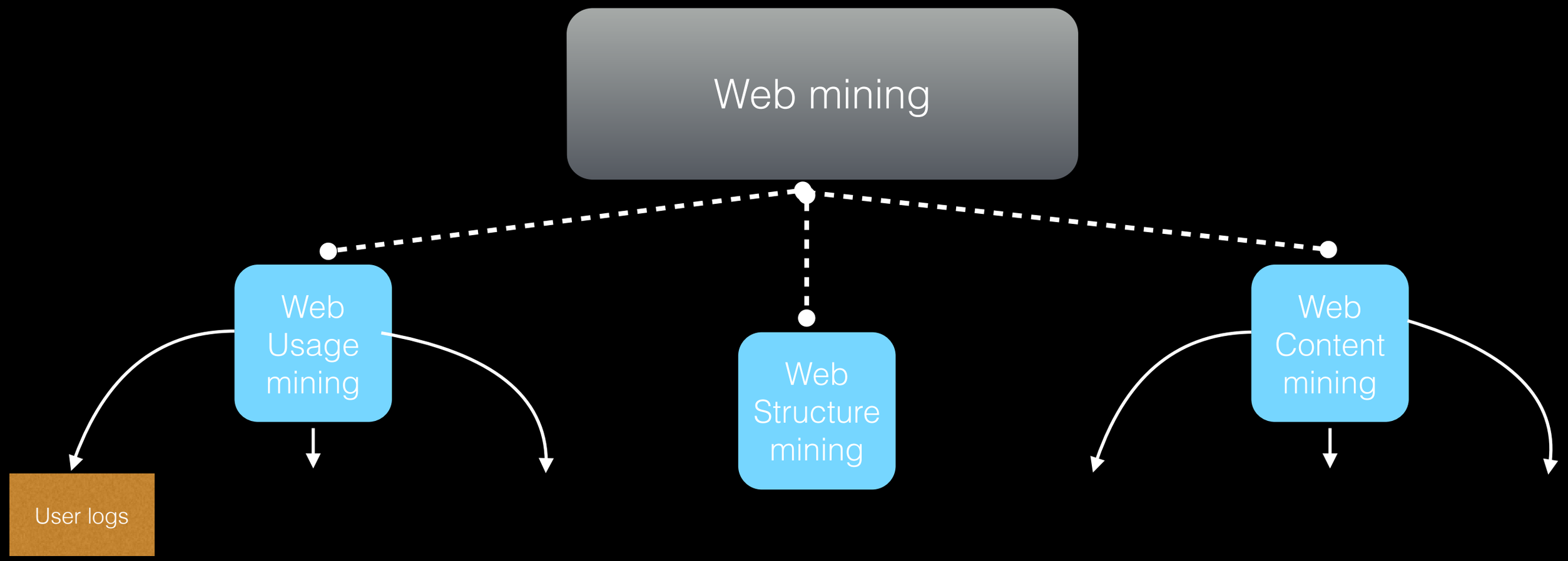
Pour rappel



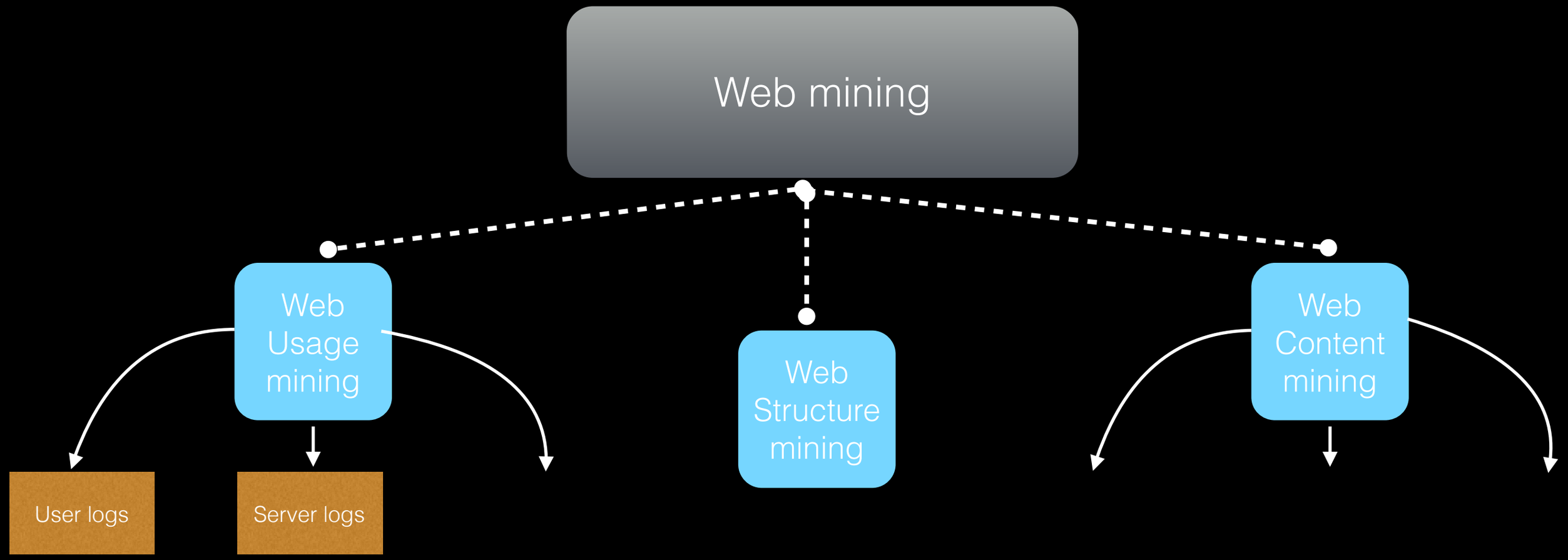
Pour rappel


Web mining
Web Usage mining
Web Structure mining
Web Content mining

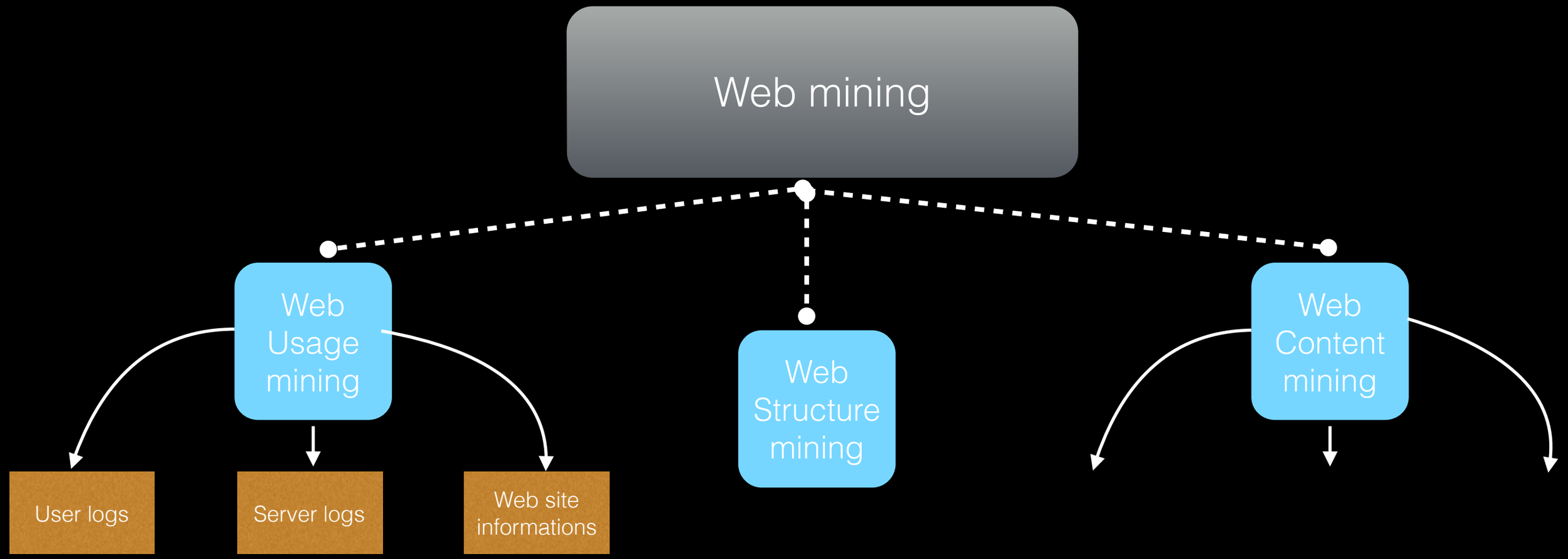
Pour rappel


Web mining
Web Usage mining
Web Structure mining
Web Content mining
User logs

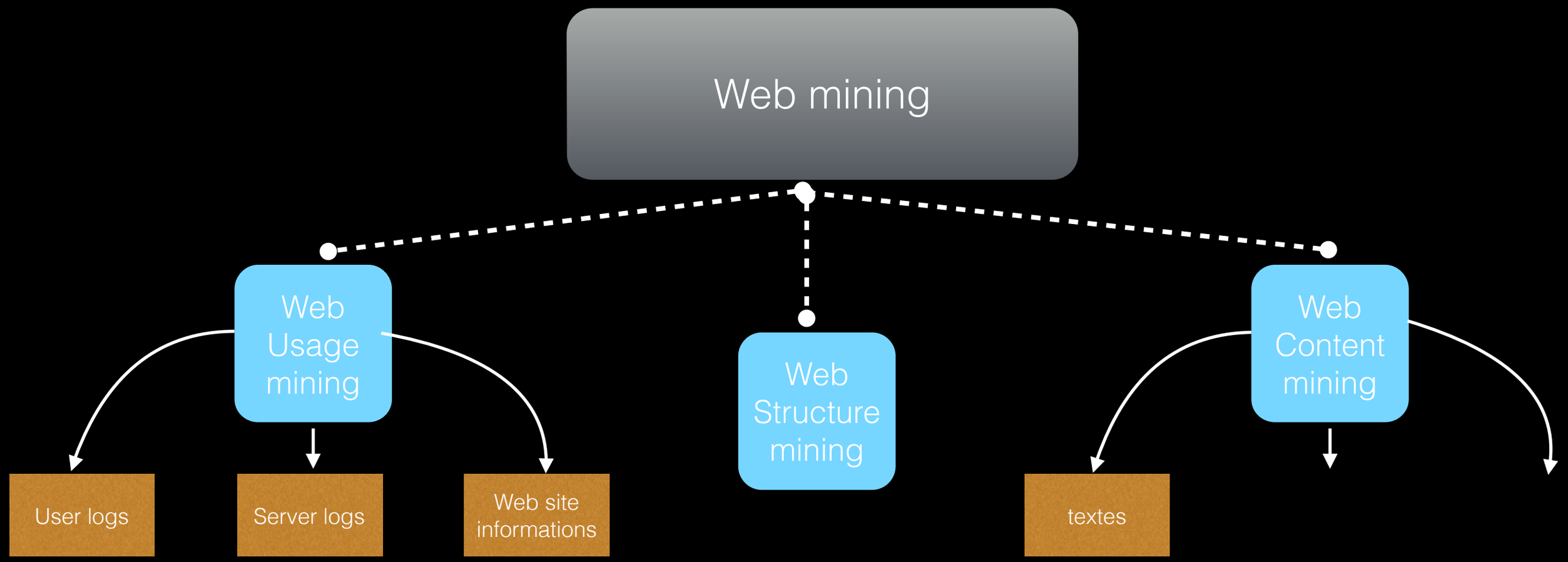
Pour rappel


Web mining
Web Usage mining
Web Structure mining
Web Content mining
User logs
Server logs

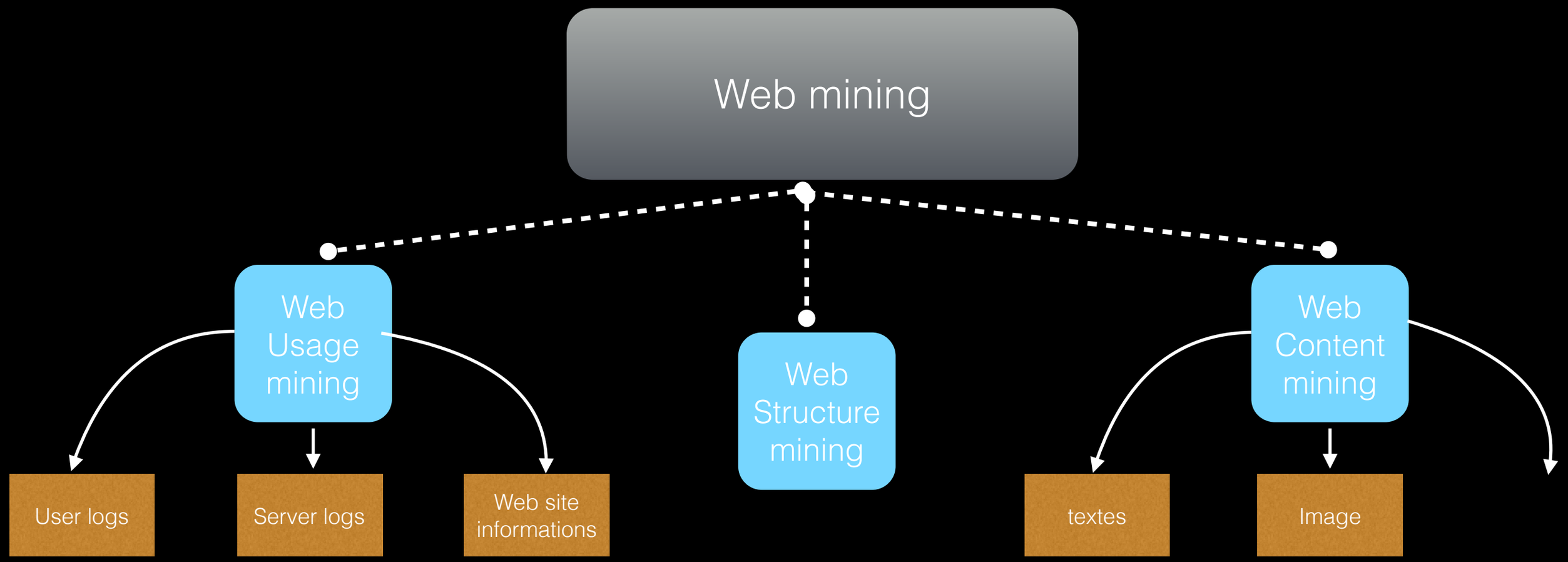
Pour rappel


Web mining
Web Usage mining
Web Structure mining
Web Content mining
User logs
Server logs
Web site informations

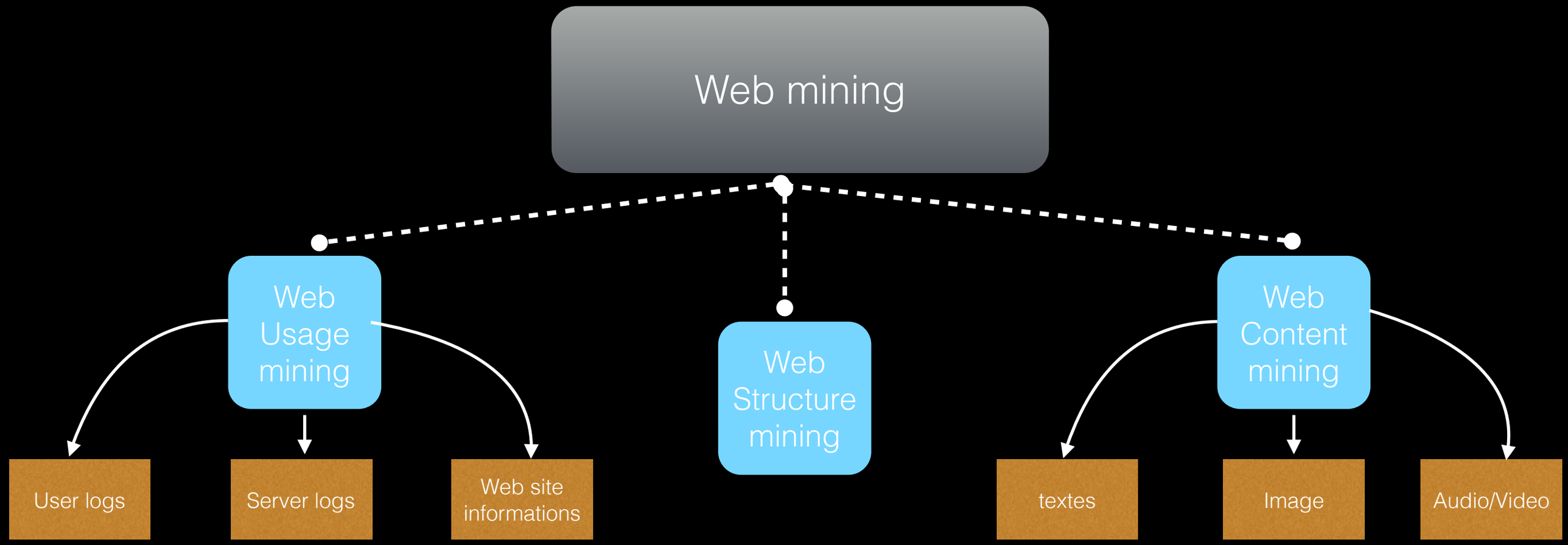
Pour rappel


Web mining
Web Usage mining
Web Structure mining
Web Content mining
User logs
Server logs
Web site informations
textes

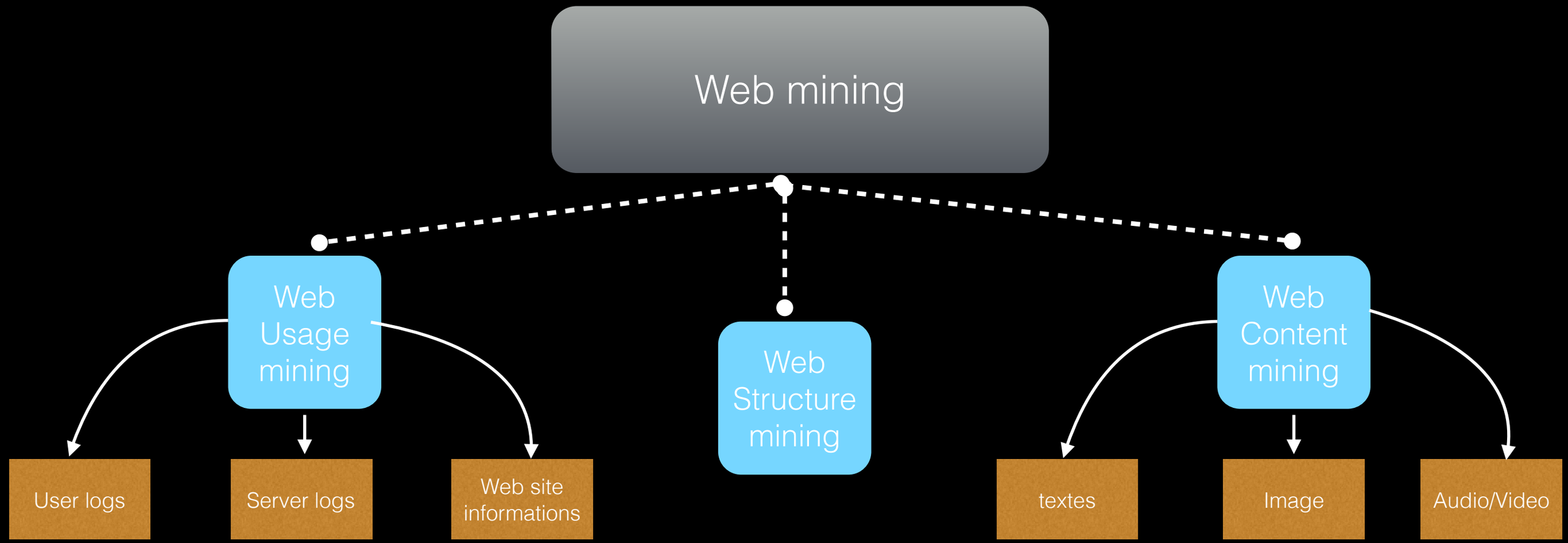
Pour rappel


Web mining
Web Usage mining
Web Structure mining
Web Content mining
User logs
Server logs
Web site informations
textes
Image

Pour rappel

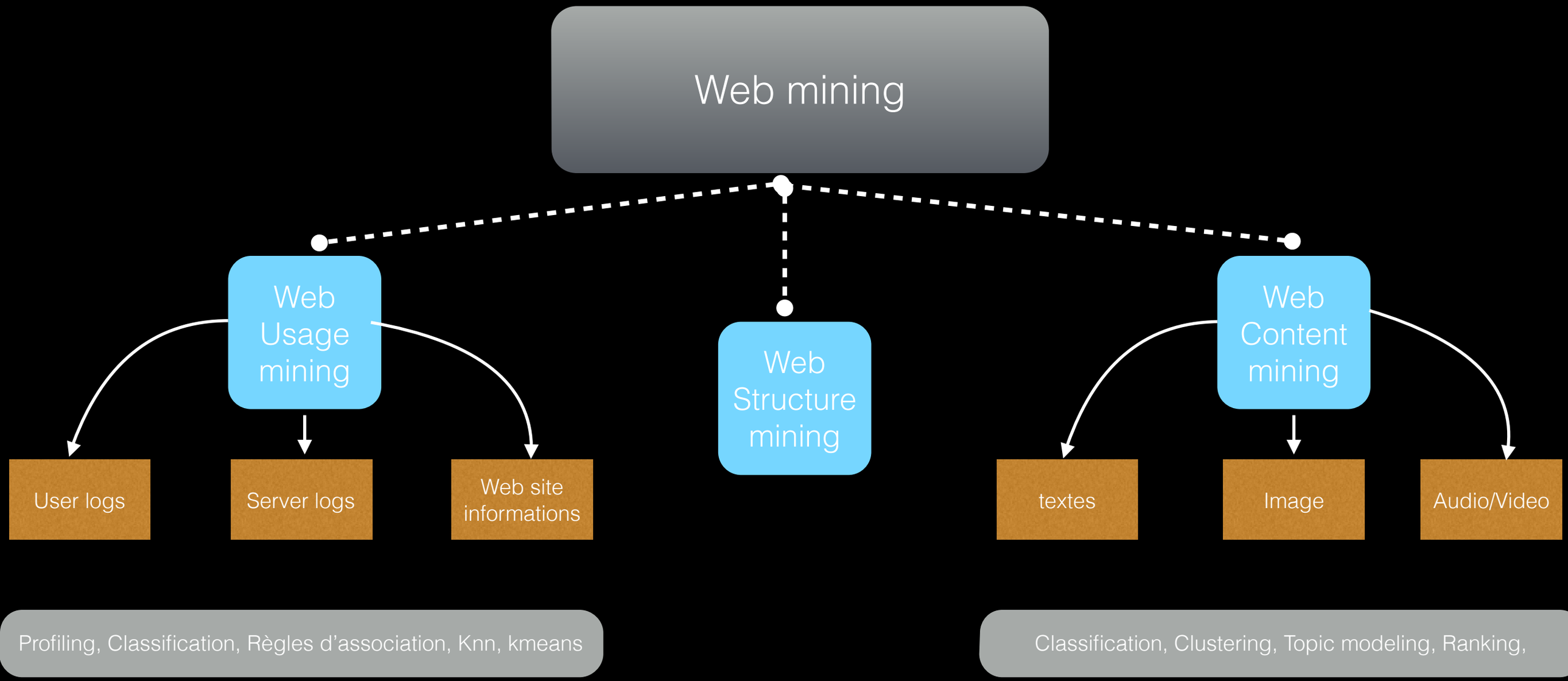

Web mining
Web Usage mining
Web Structure mining
Web Content mining
User logs
Server logs
Web site informations
textes
Image
Audio/Video

Pour rappel

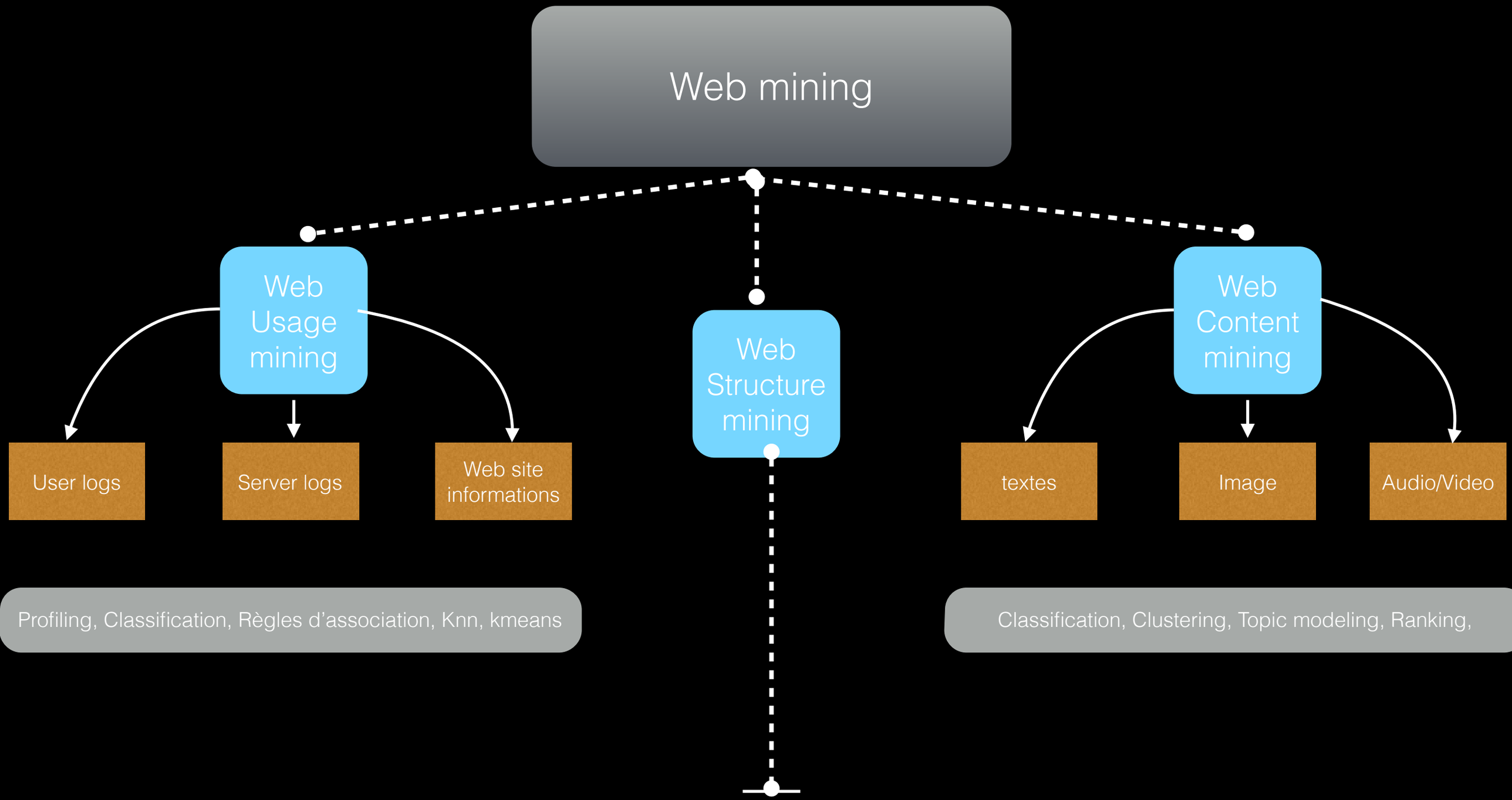


Profiling, Classification, Règles d’association, Knn, kmeans

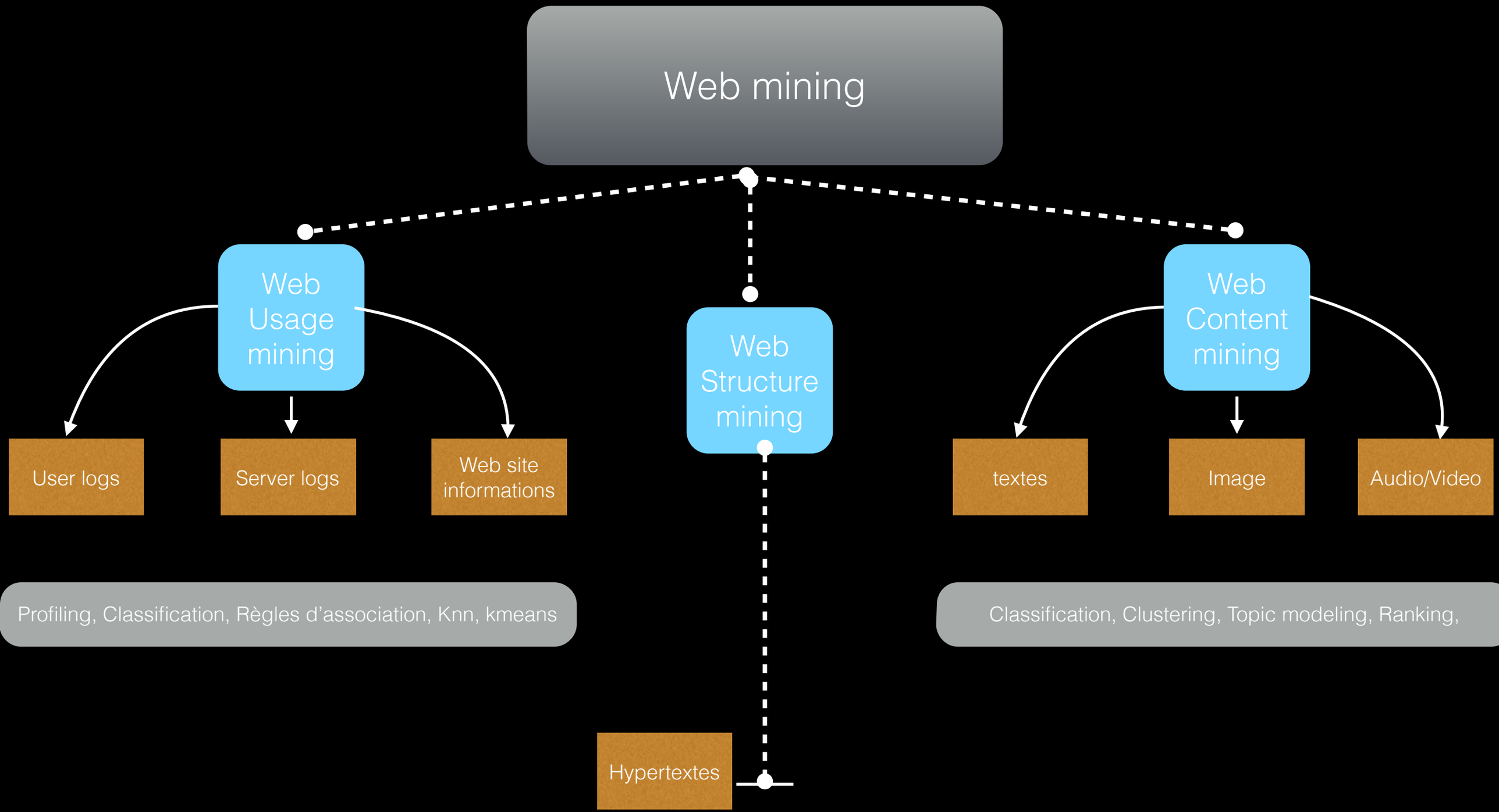
Pour rappel


Web mining
Web Usage mining
Web Structure mining
Web Content mining
User logs
Server logs
Web site informations
textes
Image
Audio/Video
Profiling, Classification, Règles d'association, Knn, kmeans
Classification, Clustering, Topic modeling, Ranking,

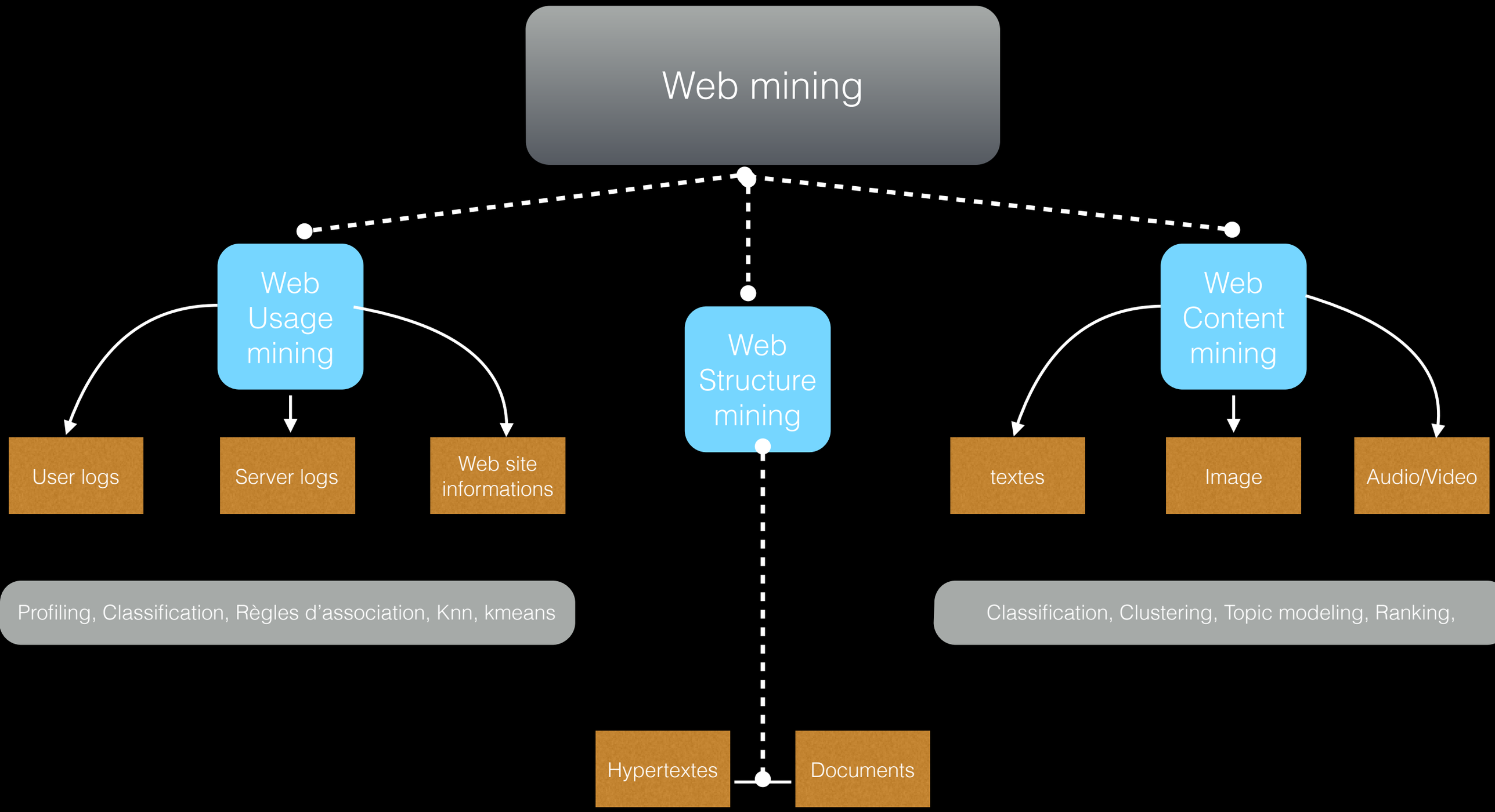
Pour rappel


Web mining
Web Usage mining
Web Structure mining
Web Content mining
User logs
Server logs
Web site informations
textes
Image
Audio/Video
Profiling, Classification, Règles d'association, Knn, kmeans
Classification, Clustering, Topic modeling, Ranking,

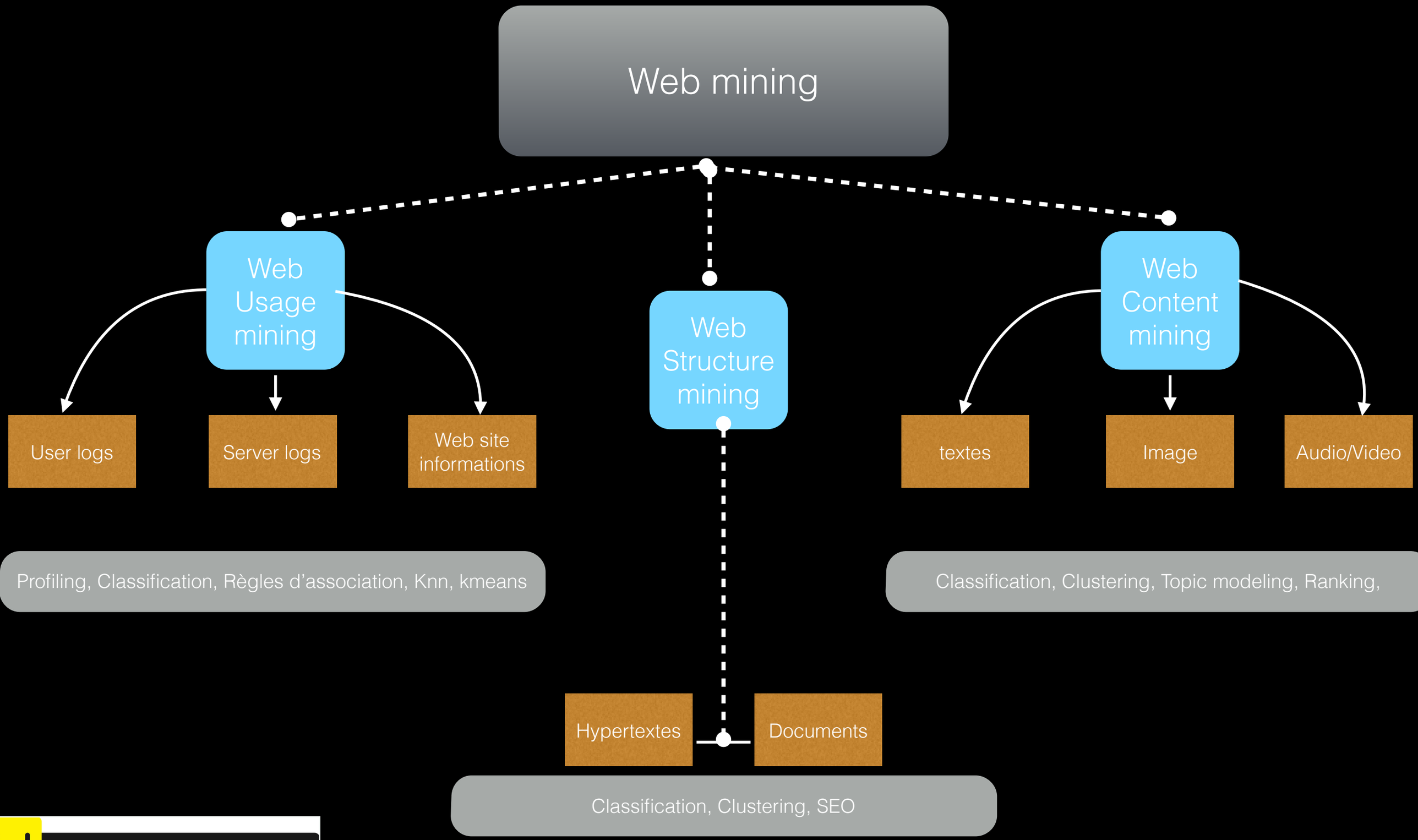
Pour rappel


Web mining
Web Usage mining
Web Structure mining
Web Content mining
User logs
Server logs
Web site informations
Hypertextes
textes
Image
Audio/Video
Profiling, Classification, Règles d'association, Knn, kmeans
Classification, Clustering, Topic modeling, Ranking,

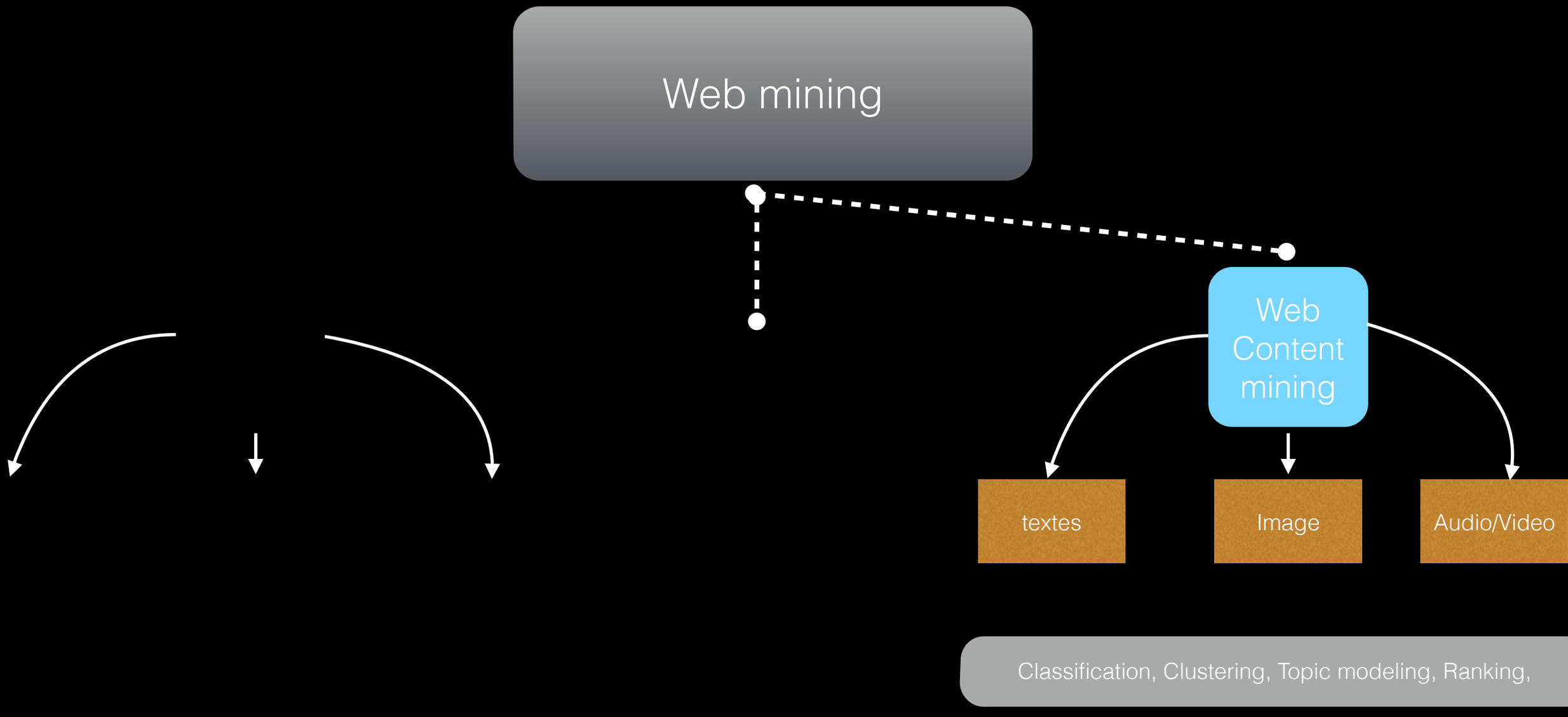
Pour rappel


Web mining
Web Usage mining
Web Structure mining
Web Content mining
User logs
Server logs
Web site informations
textes
Image
Audio/Video
Profiling, Classification, Règles d'association, Knn, kmeans
Classification, Clustering, Topic modeling, Ranking,
Hypertextes
Documents

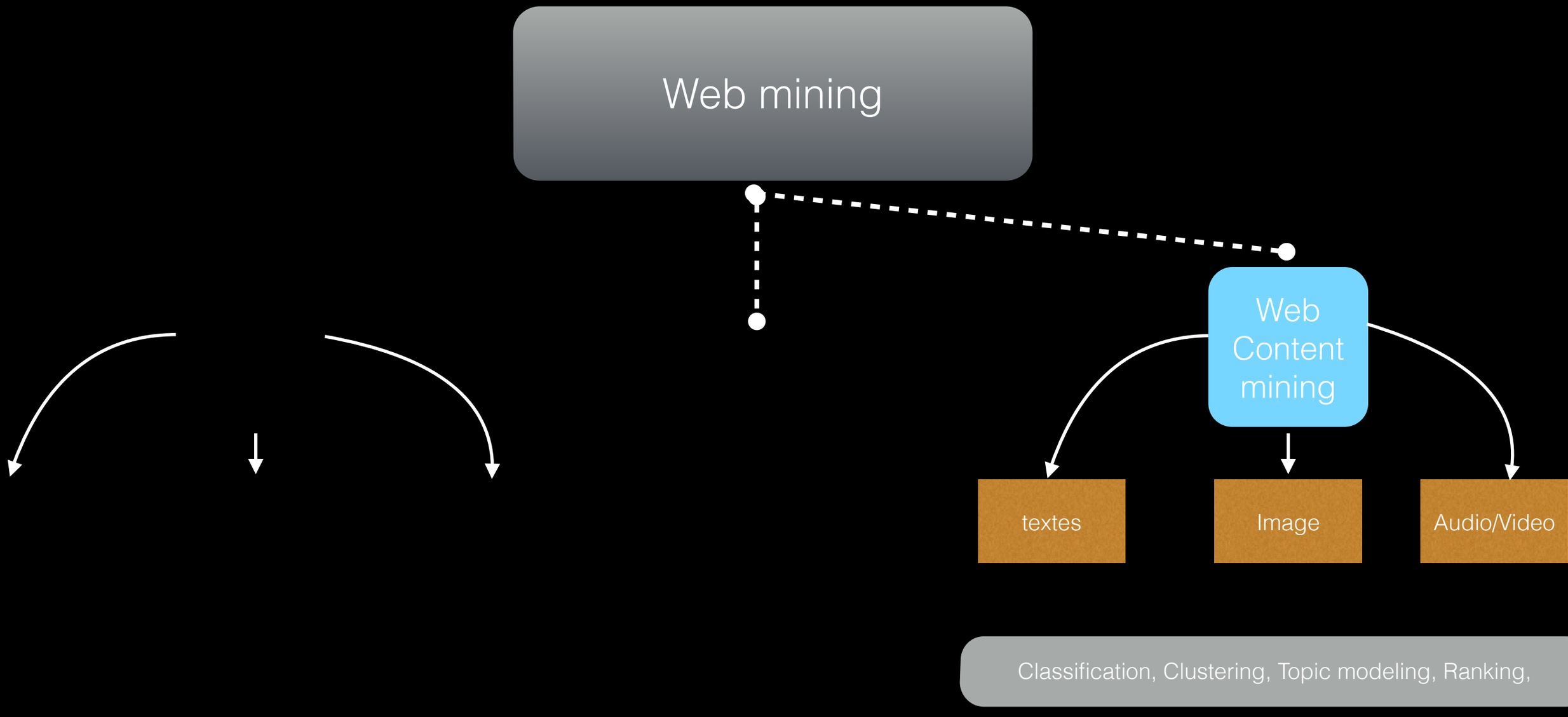
Pour rappel


Web mining
Web Usage mining
Web Structure mining
Web Content mining
User logs
Server logs
Web site informations
textes
Image
Audio/Video
Profiling, Classification, Règles d'association, Knn, kmeans
Classification, Clustering, Topic modeling, Ranking,
Hypertextes
Documents
Classification, Clustering, SEO

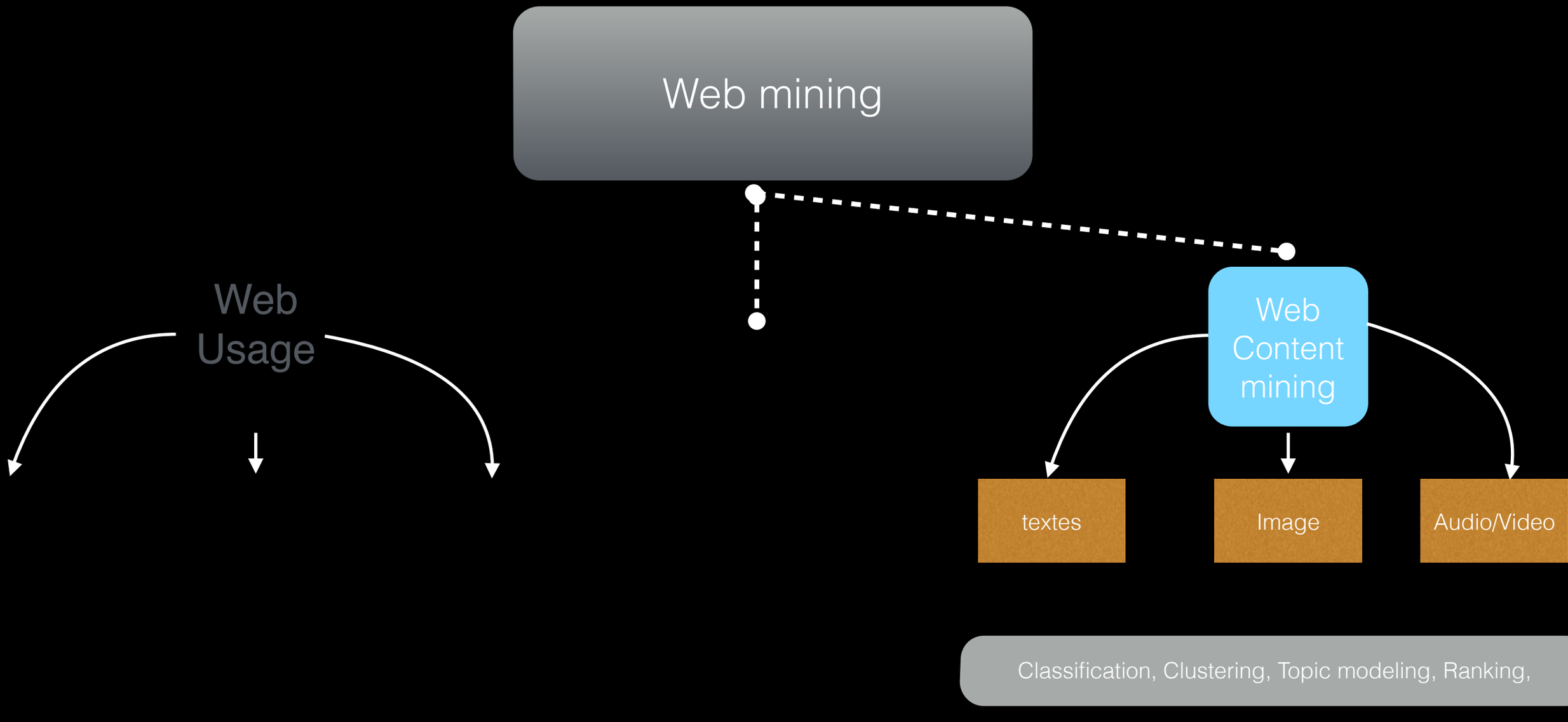
Pour rappel


Web mining
Web Content mining
textes
Image
Audio/Video
Classification, Clustering, Topic modeling, Ranking,

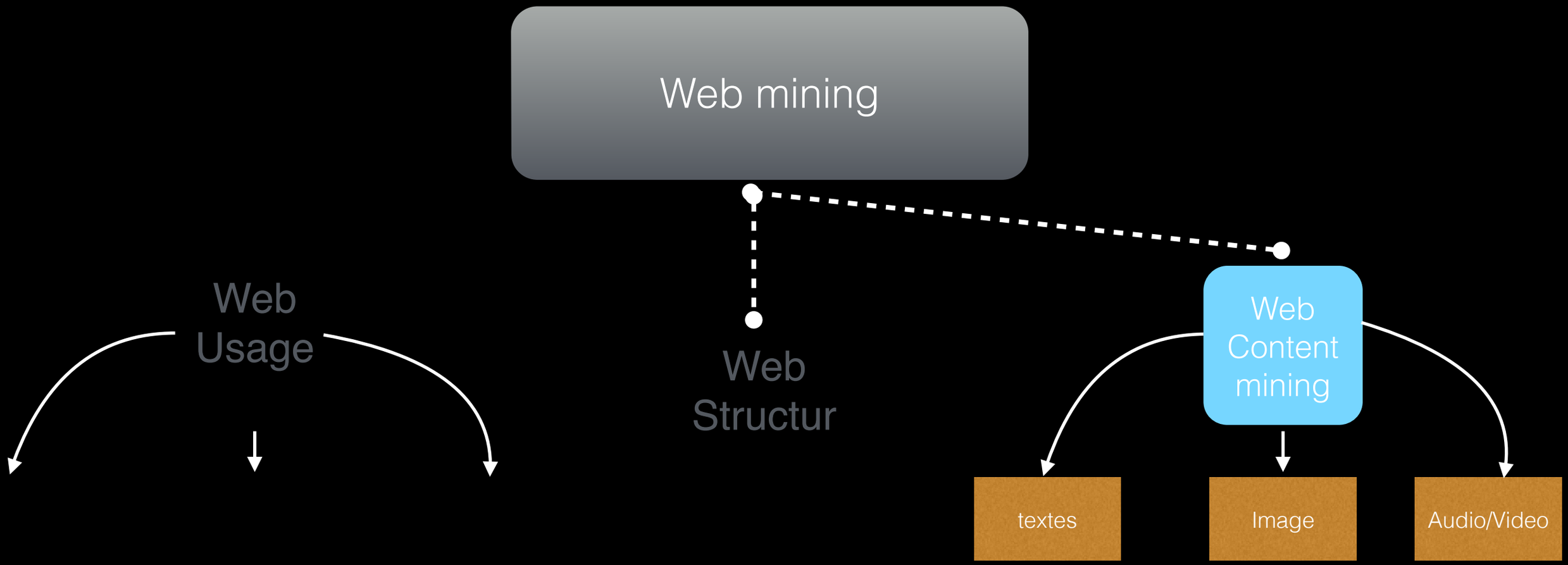
Pour rappel


Web mining
Web Content mining
textes
Image
Audio/Video
Classification, Clustering, Topic modeling, Ranking,

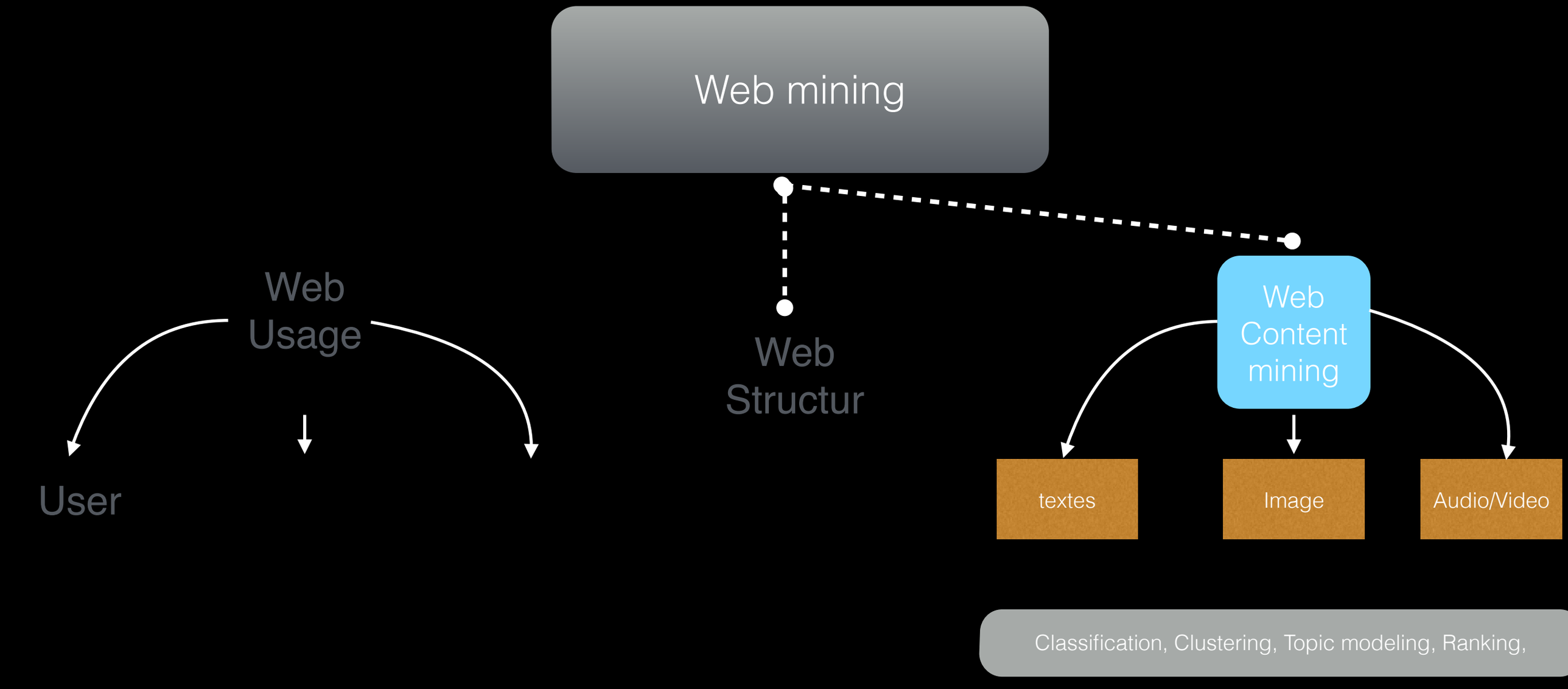
Pour rappel


Web mining
Web Usage
Web Content mining
textes
Image
Audio/Video
Classification, Clustering, Topic modeling, Ranking,

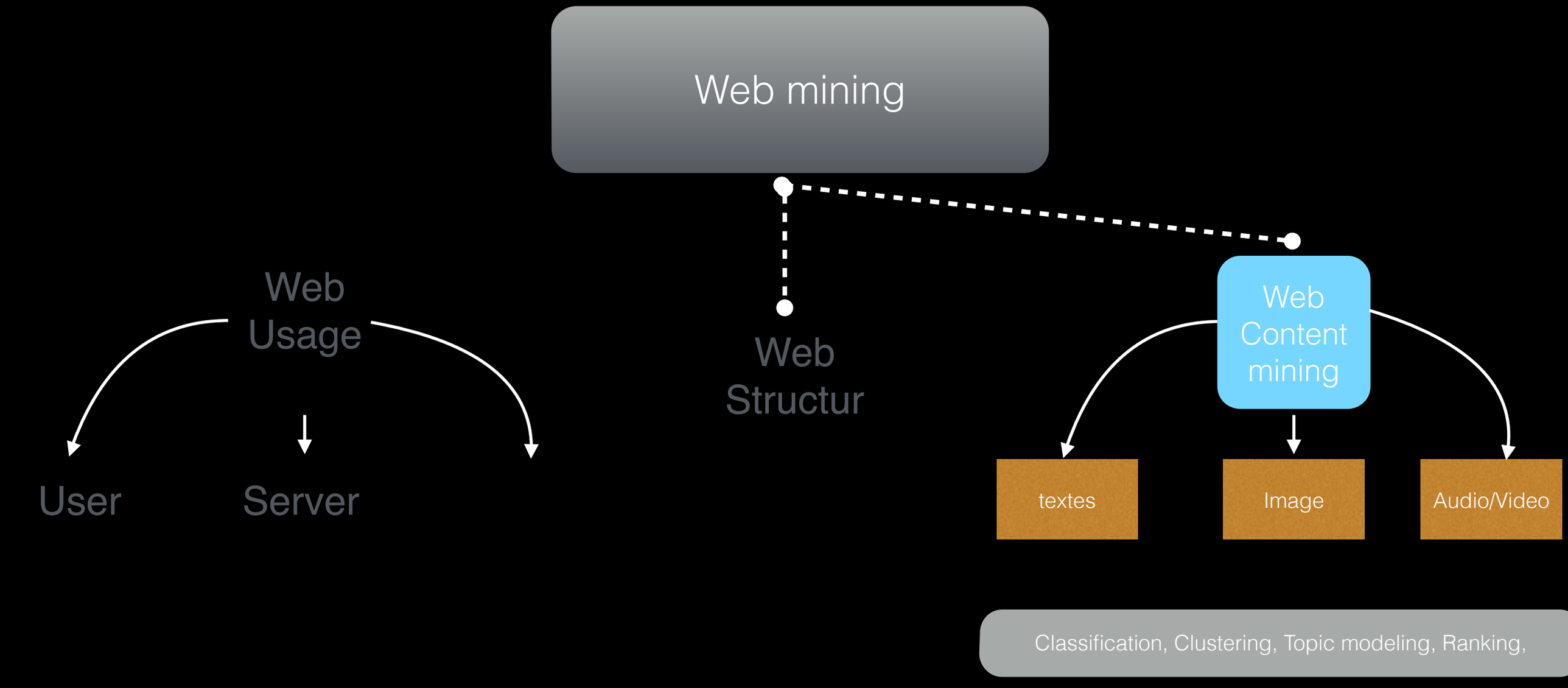
Pour rappel


Web mining
Web Usage
Web Structur
Web Content mining
textes
Image
Audio/Video
Classification, Clustering, Topic modeling, Ranking,

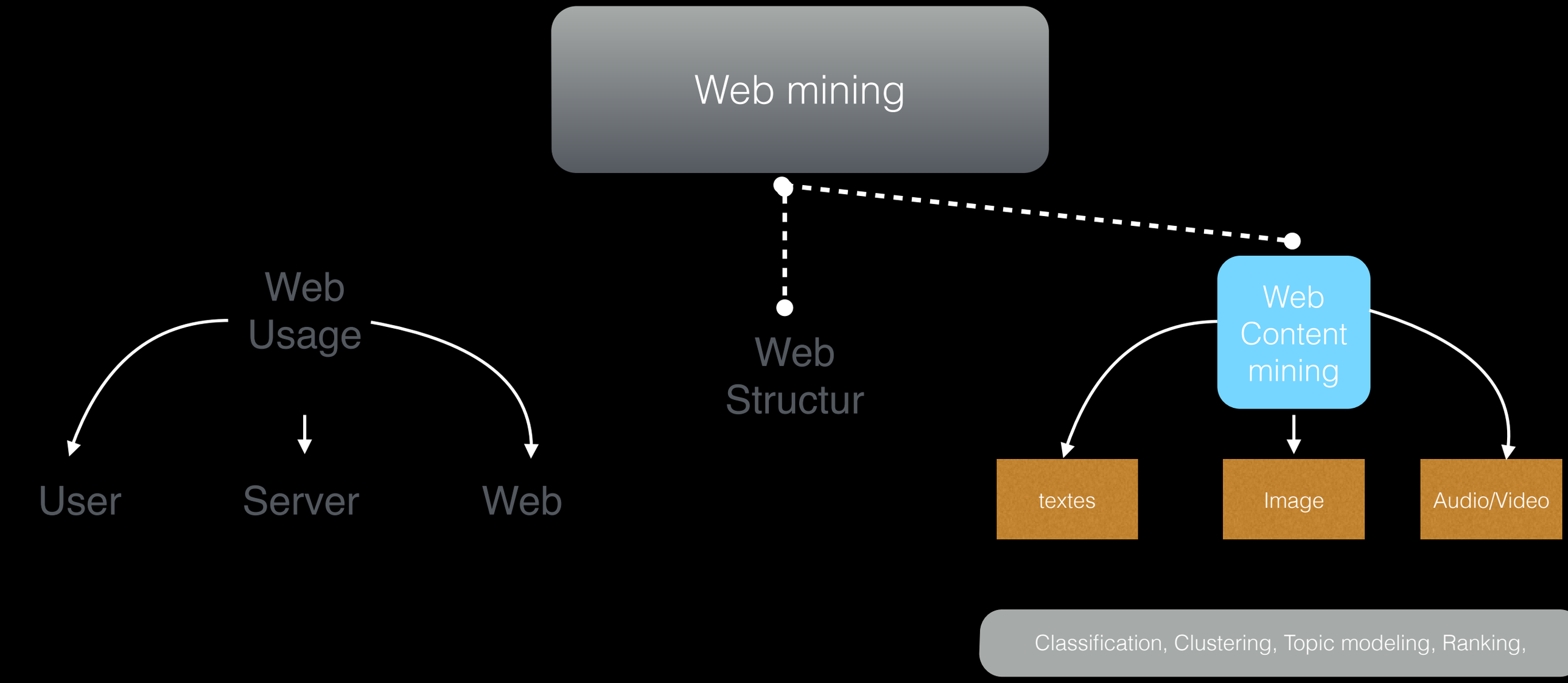
Pour rappel


Web mining
Web Usage
User
Web Structur
Web Content mining
textes
Image
Audio/Video
Classification, Clustering, Topic modeling, Ranking,

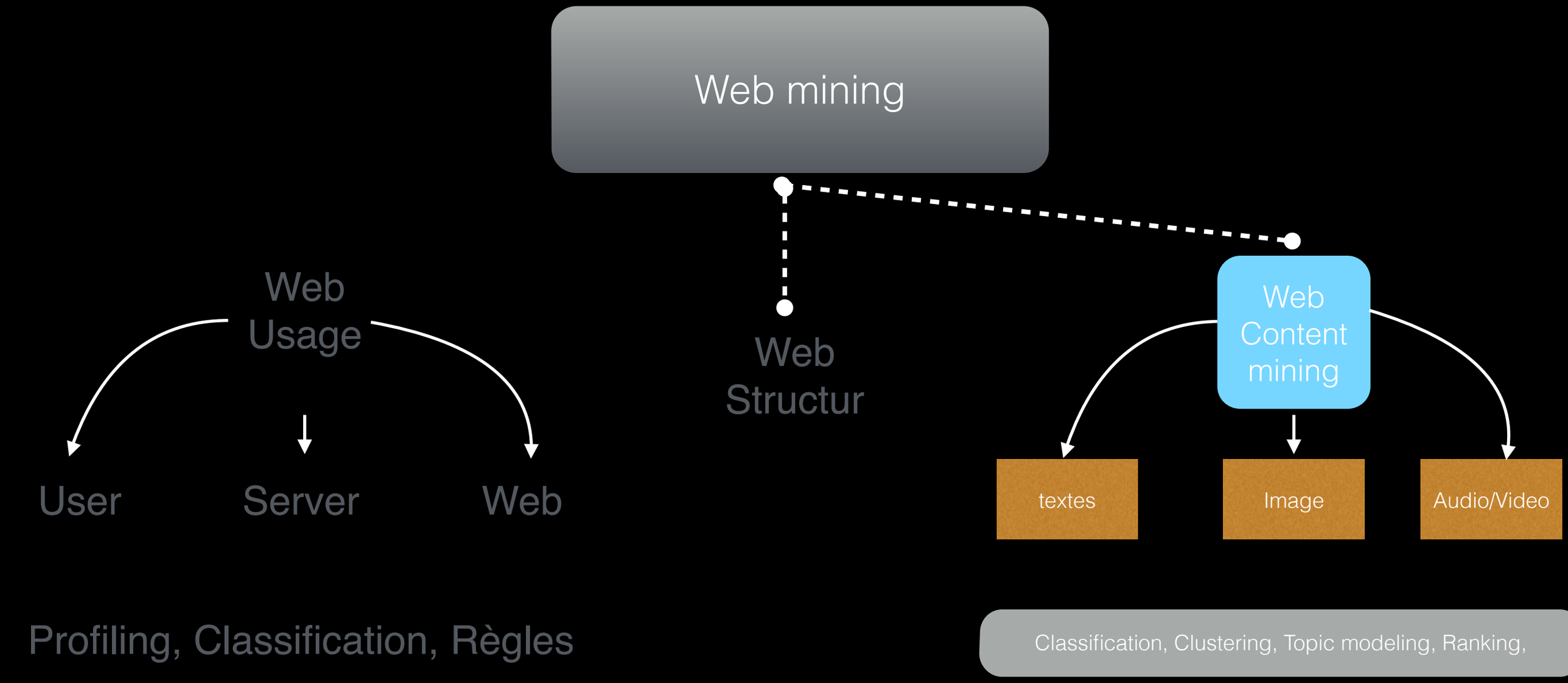
Pour rappel


Web mining
Web Usage
Web Structur
Web Content mining
User
Server
textes
Image
Audio/Video
Classification, Clustering, Topic modeling, Ranking,

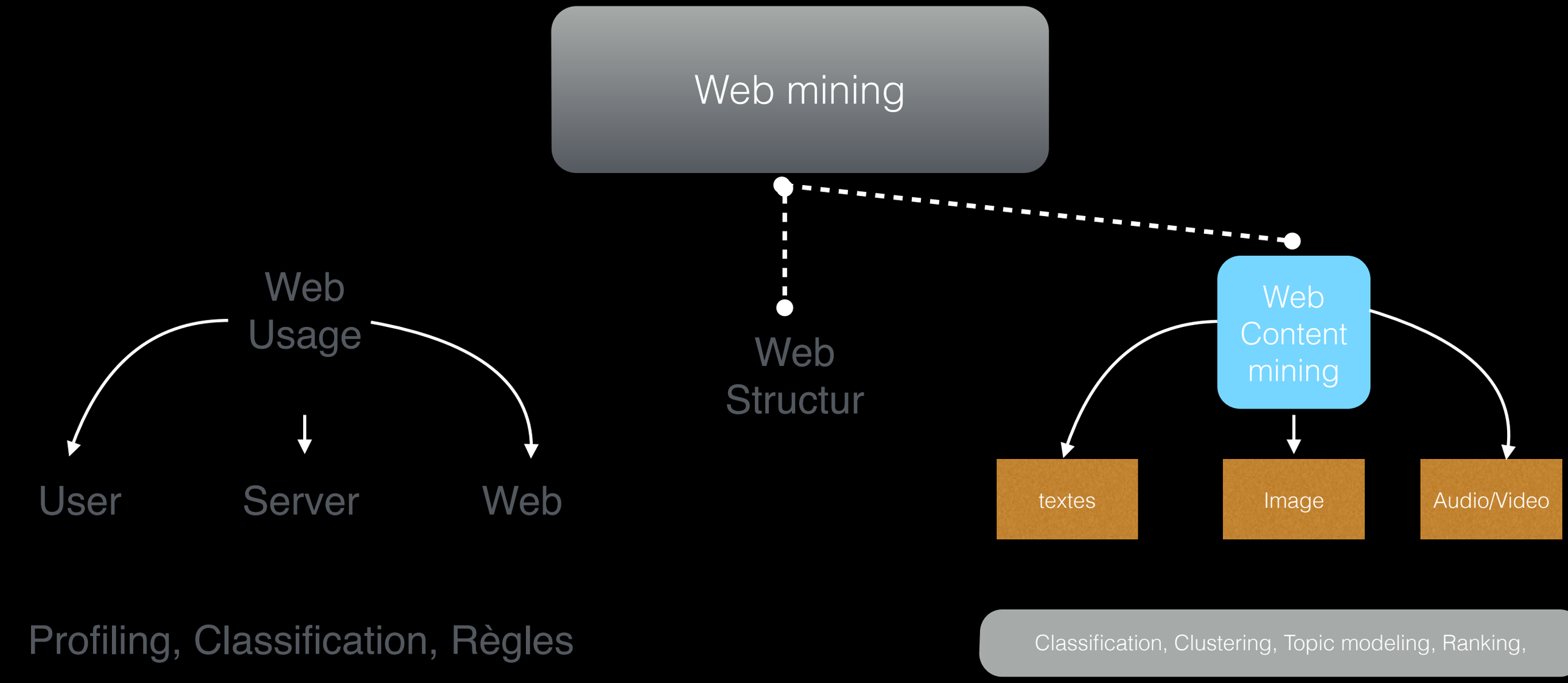
Pour rappel


Web mining
Web Usage
Web Structur
Web Content mining
User
Server
Web
textes
Image
Audio/Video
Classification, Clustering, Topic modeling, Ranking,

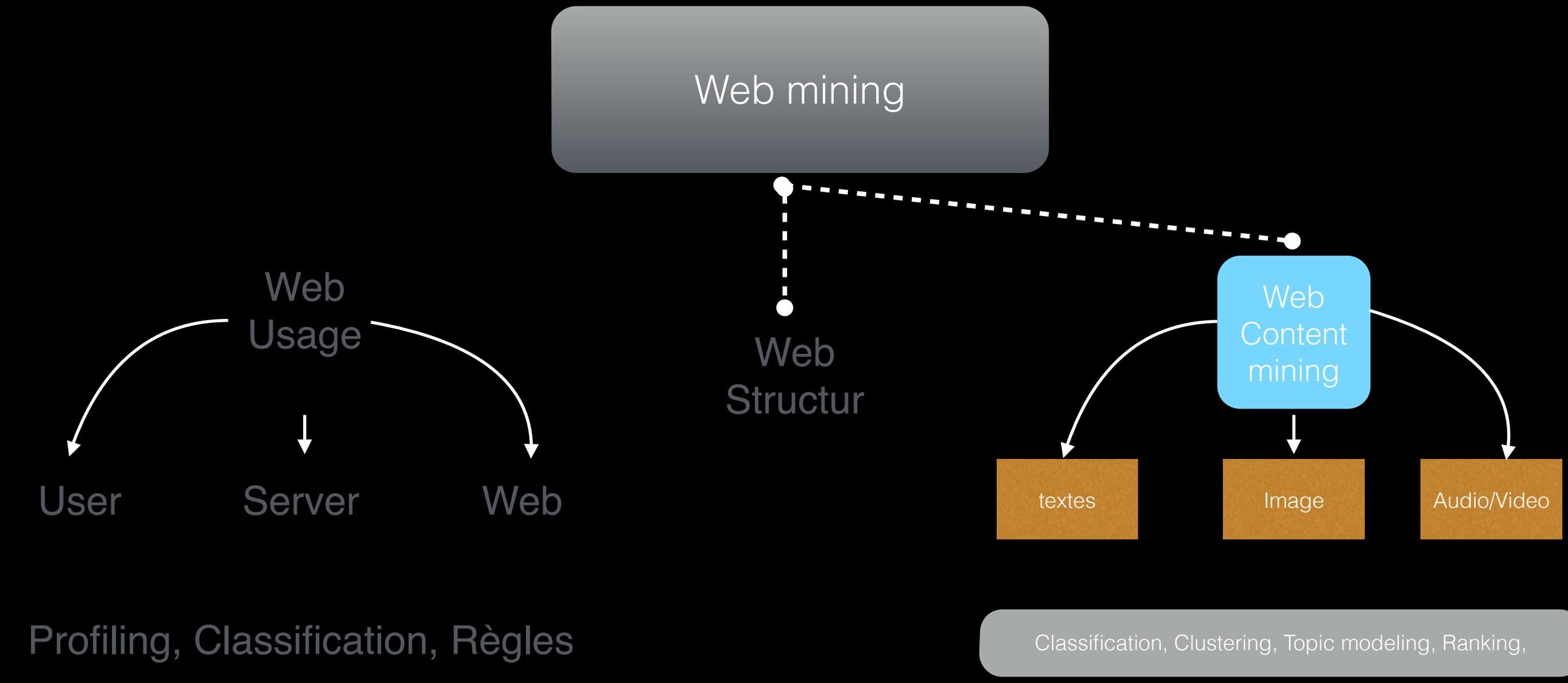
Pour rappel


Web mining
Web Usage
Web Structur
Web Content mining
User
Server
Web
textes
Image
Audio/Video
Profiling, Classification, Règles
Classification, Clustering, Topic modeling, Ranking,

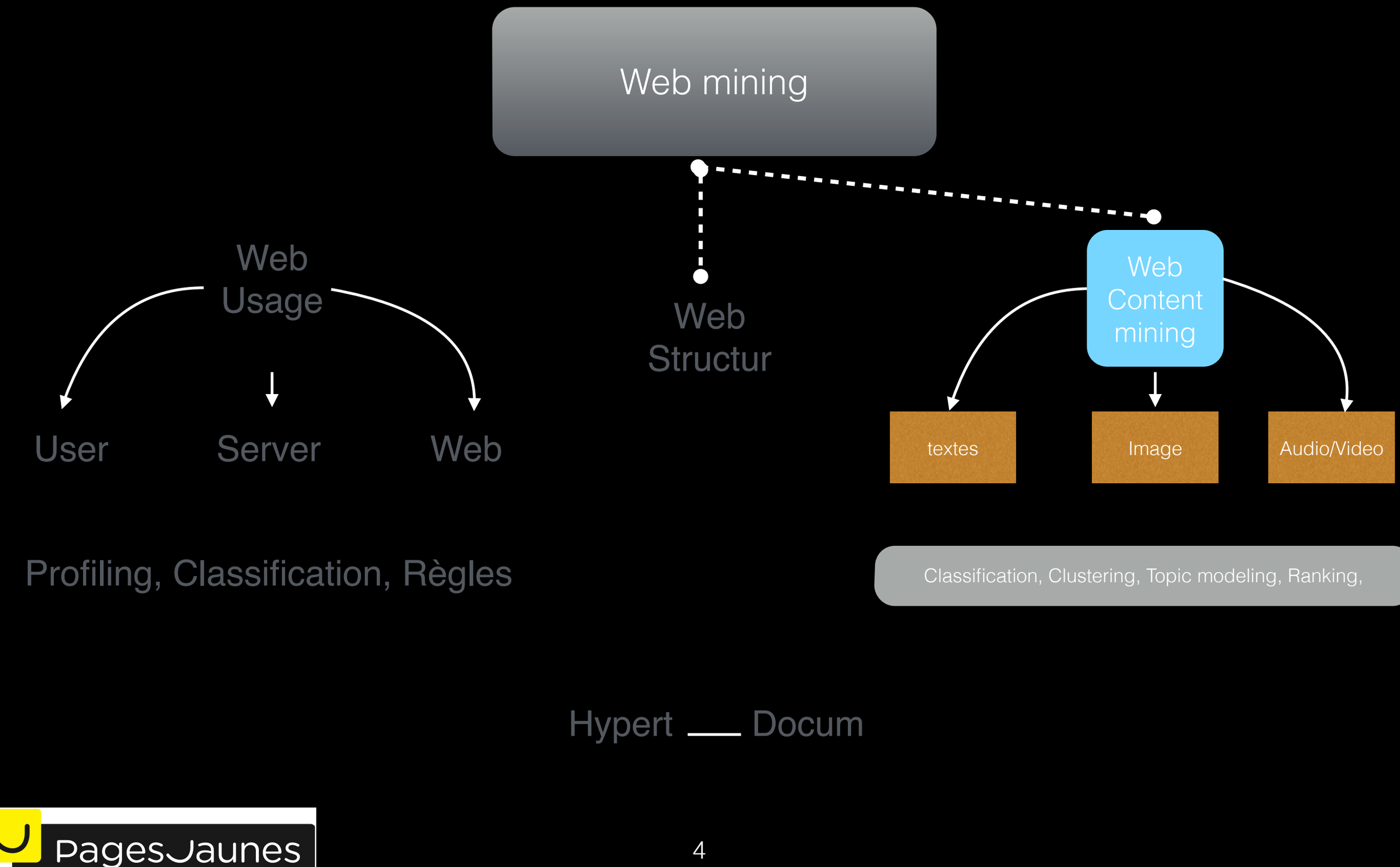
Pour rappel


Web mining
Web Usage
Web Structur
Web Content mining
User
Server
Web
textes
Image
Audio/Video
Profiling, Classification, Règles
Classification, Clustering, Topic modeling, Ranking,

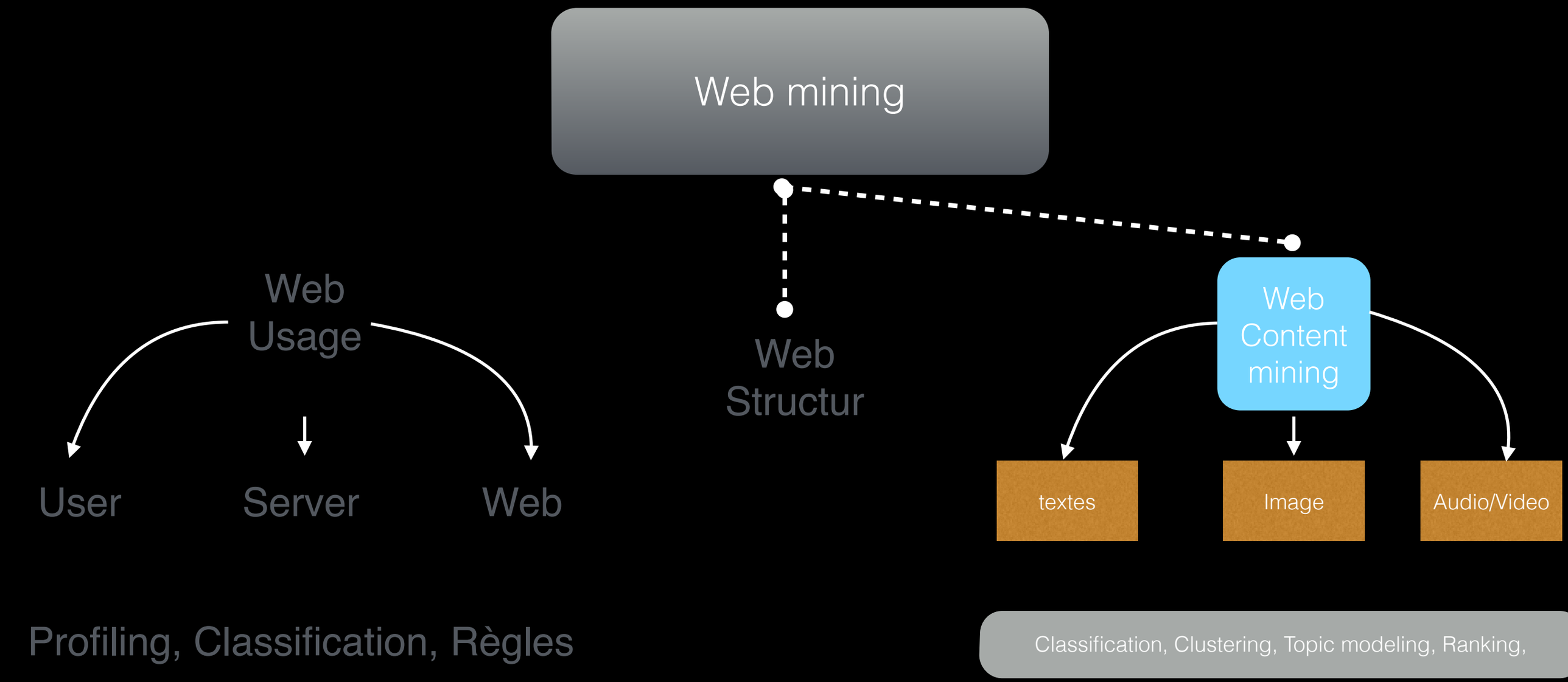
Pour rappel


Web mining
Web Usage
User
Server
Web
Web Structur
Web Content mining
textes
Image
Audio/Video
Profiling, Classification, Règles
Classification, Clustering, Topic modeling, Ranking,
Hypert

Pour rappel


Web mining
Web Usage
User
Server
Web
Web Structur
Web Content mining
textes
Image
Audio/Video
Profiling, Classification, Règles
Classification, Clustering, Topic modeling, Ranking,
Hypert ___ Docum

Pour rappel



Hypert ___ Docum

Classification, Clustering, SEO

PagesJaunes

Programme

- Concepts & Définitions
- Indexation & Crawl
- Introduction aux techniques de webdatamining
- Applications a la classification des documents

PagesJaunes

Concepts et Définitions
Premières définitions

Concepts et Définitions
Premières définitions

- La recherche d’informations est le coeur du search

Concepts et Définitions
Premières définitions

- La recherche d'informations est le coeur du search
- Web search has its root in **information retrieval** (or IR for short), a field of study that helps the user find needed information from a large collection of text documents. [Liu]

Concepts et Définitions
Premières définitions

- La recherche d'informations est le coeur du search
- Web search has its root in **information retrieval** (or IR for short), a field of study that helps the user find needed information from a large collection of text documents. [Liu]
- en RI => Document

Concepts et Définitions
Premières définitions

- La recherche d'informations est le coeur du search
- Web search has its root in **information retrieval** (or IR for short), a field of study that helps the user find needed information from a large collection of text documents. [Liu]
- en RI => Document
- Web => Document = page web

Concepts et Définitions
Premières définitions

- La recherche d'informations est le coeur du search
- Web search has its root in **information retrieval** (or IR for short), a field of study that helps the user find needed information from a large collection of text documents. [Liu]
- en RI => Document
- Web => Document = page web
- RI = Chercher de l'information pertinente à partir d'une requête

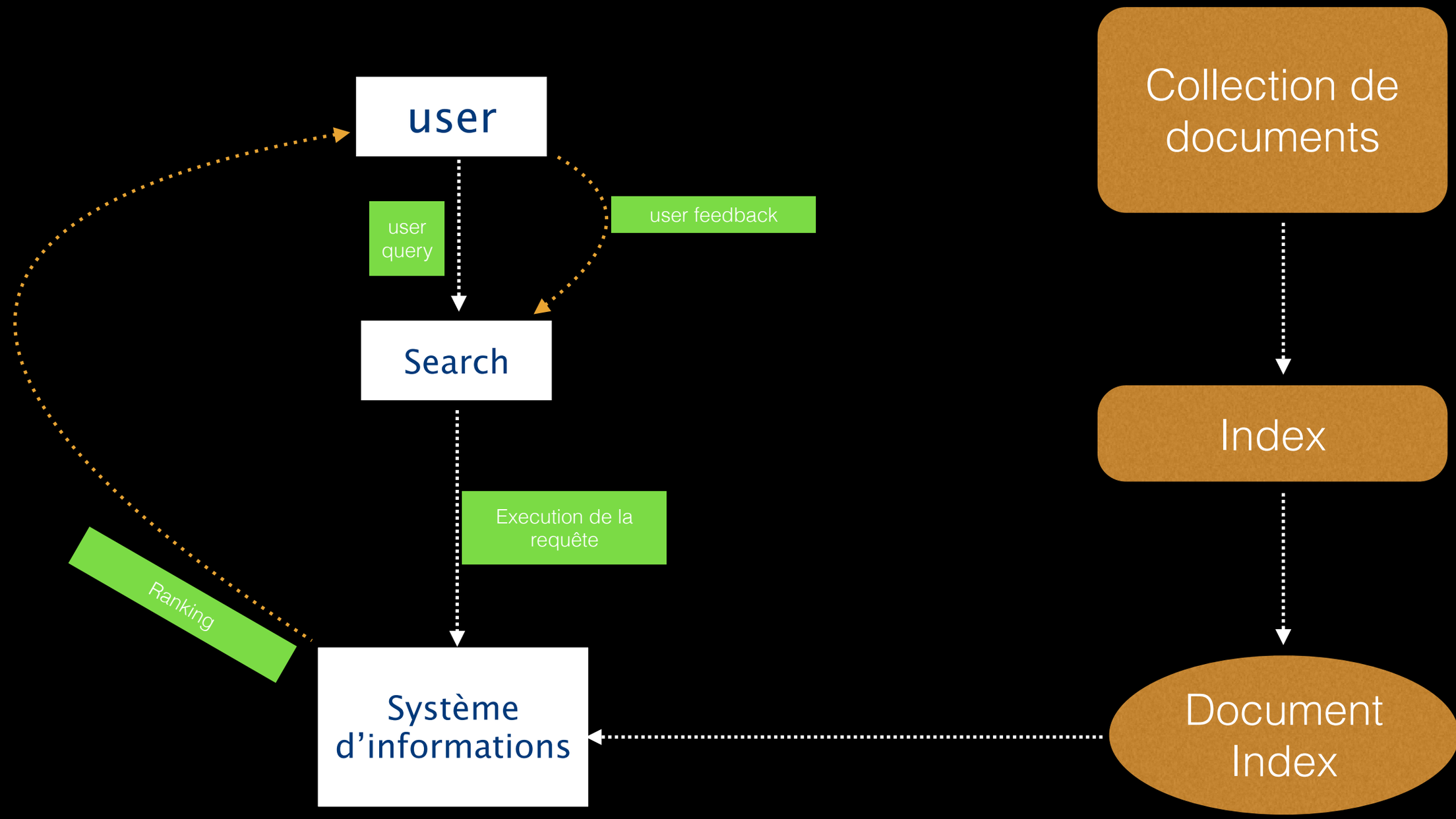
Concepts et Définitions
Premières définitions

- La recherche d'informations est le coeur du search
- Web search has its root in **information retrieval** (or IR for short), a field of study that helps the user find needed information from a large collection of text documents. [Liu]
- en RI => Document
- Web => Document = page web
- RI = Chercher de l'information pertinente à partir d'une requête
- Plusieurs notions en découlent

Concepts et Définitions
Premières définitions

- La recherche d'informations est le coeur du search
- Web search has its root in **information retrieval** (or IR for short), a field of study that helps the user find needed information from a large collection of text documents. [Liu]
- en RI => Document
- Web => Document = page web
- RI = Chercher de l'information pertinente à partir d'une requête
- Plusieurs notions en découlent
  - ➡ Mots clés, Ranking, Pertinence, Base de données, Indexation

Concepts et Définitions
Architecture de la RI



Concepts et Définitions
Requêtes

- Recherches par mots clés : Exemple de PagesJaunes
  - L'utilisateur exprime sa recherche à l'aide de quelques mots clés. Le moteu répond par une liste de documents contenant les mots clés (ex : médecin)

- Recherche full text ou de documents : Exemple de google
  - L'utilisateur exprime sa recherche par une phrase et le moteur remonte de documents qui contiennent au moins une fois la phrase
- Recherche approchée :
  - Comme la recherche full, mais sont remontés les documents qui contiennent ur partie seulement de la phrase sont remontées
- Recherche en langage naturel
- Les opérations de recherche peuvent être très complexes

Programme

- Concepts & Définitions
- Indexation & Crawl
- Introduction aux techniques de webdatamining
- Applications a la classification des documents

PagesJaunes

Concepts et Définitions
Indexation

- Un index est un module qui enregistre les documents bruts et permet ainsi une recherche efficace
- L'action de rechercher l'information consiste donc à :
- Calculer le degré de pertinence de chaque document indexé avec requête de l'utilisateur
- Ordonnancer une liste de documents pertinents à renvoyer l'utilisateur

- Modèle Booléen

In the Boolean model, documents and queries are represented as sets of terms. That is, each term is only considered present or absent in a document. [LIU]

La recherche dans un modèle booléen s'effectue par l'utilisation de l'algèbre des prédicats (et, ou, non)

Le système d'information renvoie des documents pour lesquels la requête est TRUE

Mot clé à retenir : EXACT MATCH

Concepts et Définitions
Représentation de documents

- Modèle à vecteurs espaces-états

A document in the vector space model is represented as a weight vector, in which each component weight is computed based on some variation of TF or TF-IDF scheme.[LIU]

TF : Terms Frequency (Nombre de fois où le terme apparaît dans le document)

IDF : Inverse Document Frequency (Nombre de documents dans lequel le terme apparaît)

La recherche dans le système d'informations se fait sur la base de mesure de similarité entre documents => Cosinus, Indice de Jaccard, etc…

Programme

- Concepts & Définitions
- Indexation & Crawl
- Introduction aux techniques de webdatamining
- Applications a la classification des documents

Introduction aux techniques de webmining
1. Term Document Matrix

- TDM ou DTM

Une matrice de matrice document-terme ou terme-document est une matrice au sens mathématique du terme qui décrit la fréquence des termes qui apparaissent dans une collection de documents.

Dans une tdm(term-document-matrix), les lignes correspondent aux documents de la collection et les colonnes correspondent à des termes. Il existe différents systèmes pour déterminer la valeur que chaque entrée de la matrice doit prendre. le plus répandu est la mesure tf-idf(Term Frequency-Inverse Document Frequency), mais aussi, tf(Term Frequency).

PAUSE

Web Datamining

Information Retrieval and Machine Learning for the Web

Guibert J. TCHINDE
ENSAI | Solocal group – 2014

PagesJaunes