

# Introduction



# Tecnologías de Gestión de Información No Estructurada

Prof. Dr. David E. Losada

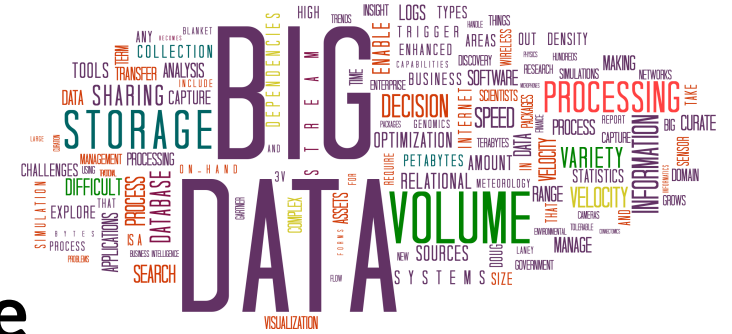


Centro Singular de Investigación  
en **Tecnoloxías Intelixentes**



# Máster Interuniversitario en Tecnologías de Análisis de Datos Masivos: Big Data

# Big Data



raw data => actionable **knowledge**

optimize **decision making** in many application domains  
(health, security/safety, education, science, BI, ...)

“see” useful **hidden** information & knowledge **buried**  
in the data

managing and analyzing **large amounts of text data**  
can help **users** manage and make use of text data in all  
kinds of **applications**



# Text Data

**natural language text**

(e.g., English text)

web pages, social media data (e.g., tweets), news, scientific literature, emails, government documents, enterprise data ...

**production & consumption** of large amounts of text data

**every day**

all kinds of **topics**



# Explosive Growth

## BIG IN GROWTH, TOO.

1 exabyte (EB) = 1,000,000,000,000,000 bytes



**impossible** for people to **consume** all the **relevant** text data in a timely manner

need for **intelligent information retrieval** systems to help people **manage the text data** and get **access** to the needed relevant info



# Text Data

as a **special kind of big data** text data  
offer a great opportunity to discover **knowledge**  
useful for many applications

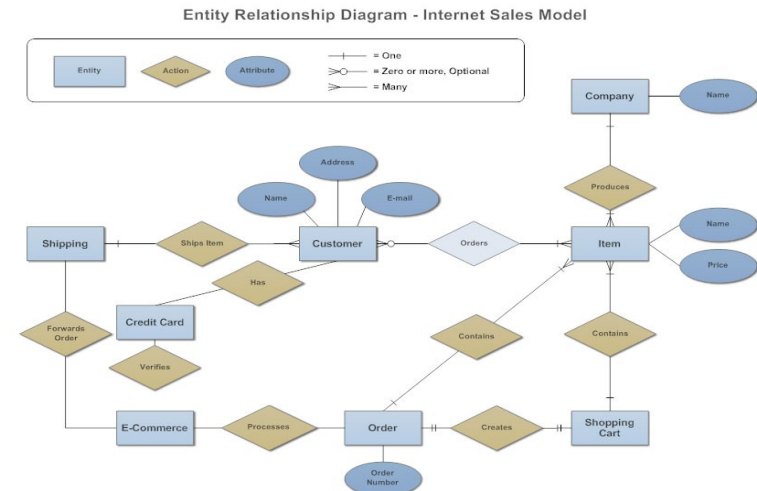


for example, **opinionated text data**  
(product reviews, forum discussions, social media)

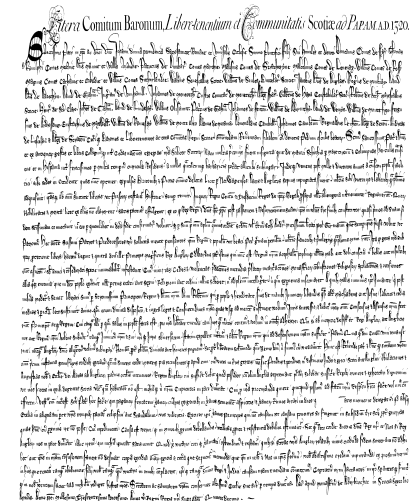


# Structured Data vs Unstructured Data

Structured data:  
**well-defined schemas**  
**easy for computers to handle**



Unstructured data (e.g. text):  
**less explicit structure**  
**requires computer processing**  
**to understand the content**



# Understanding Text



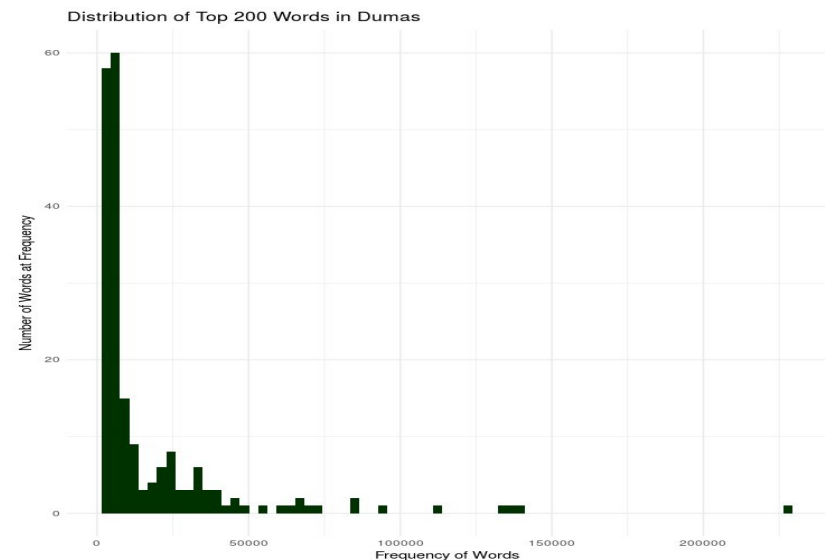
**natural language processing (nlp)** has not yet reached a point to enable a computer to precisely understand text

**statistical & heuristic approaches** to management and analysis of text data

robust

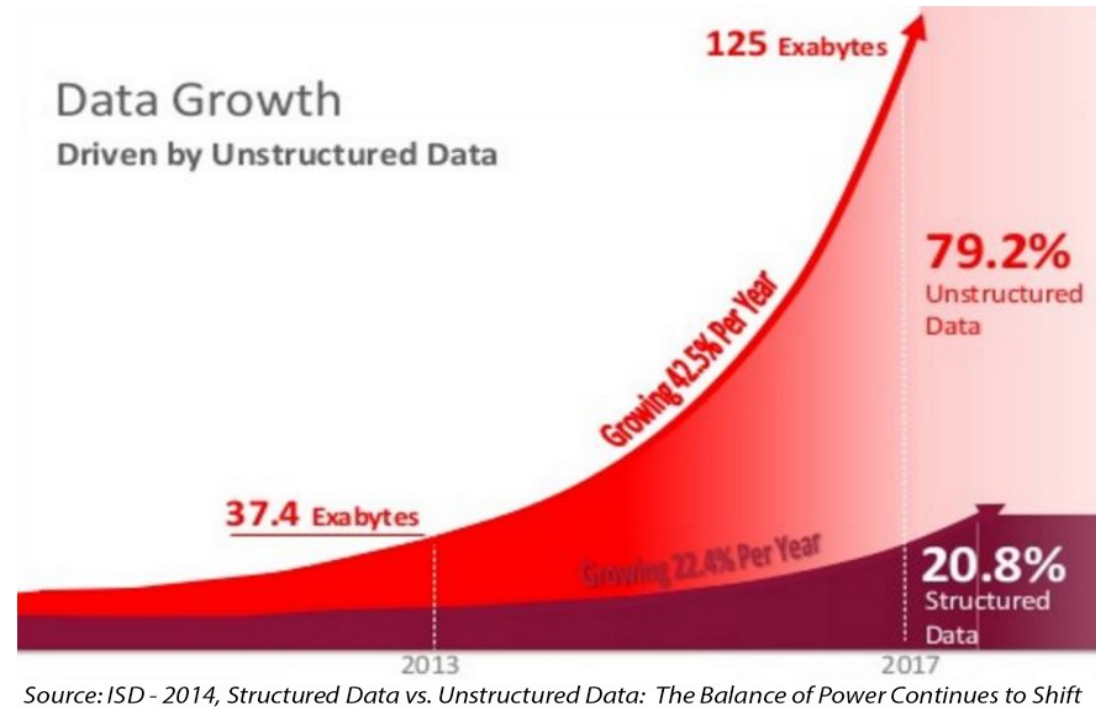
can be applied to any language

& topic





“the world produces between 1 and 2 **exabytes** ( $10^{18}$  petabytes) of unique information per year, which is roughly 250 megabytes for every man, woman, and child on earth. Printed documents of all kinds comprise only .03% of the total.” [Lyman et al. 2003]



A large amount is **textual**

Newspapers, magazines, office documents, emails, blog entries, tweets...



# text is arguably the most useful kind of info

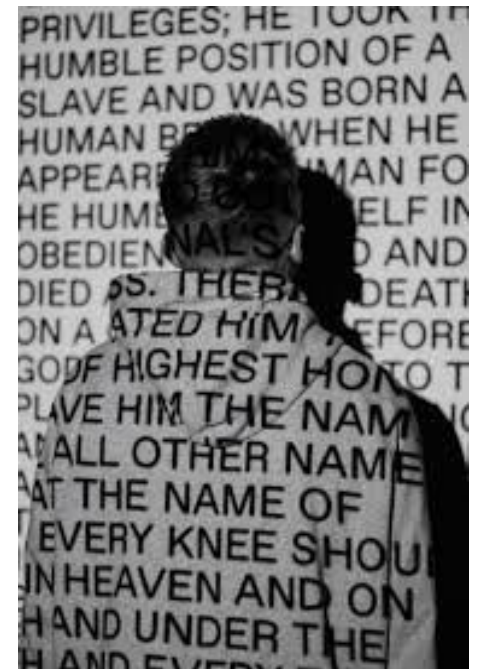
most **natural way of encoding human knowledge**

examples: **scientific knowledge** almost exclusively exists in scientific literature, **technical manuals** detailed explanations of how to operate devices

most **common type of info** encountered

most **expressive form of info**

text used to describe other media (video, images)



# 2 related services to manage & exploit big text data

## Text Retrieval



no one can possibly **digest** all info

urgent need for developing intelligent text retrieval systems to help people get **access to** the needed relevant information **quickly** and **accurately**

search engines



useful not only for the web!

useful anywhere there is a relatively **large amount of text data** (e.g., desktop search, enterprise search or literature search).

# 2 related services to manage & exploit big text data

## Text Mining

text data: rich in **semantic** content

valuable knowledge, info, opinions, preferences

opportunity for discovering **knowledge**

useful for many **applications**

### intelligent software tools

discover relevant knowledge

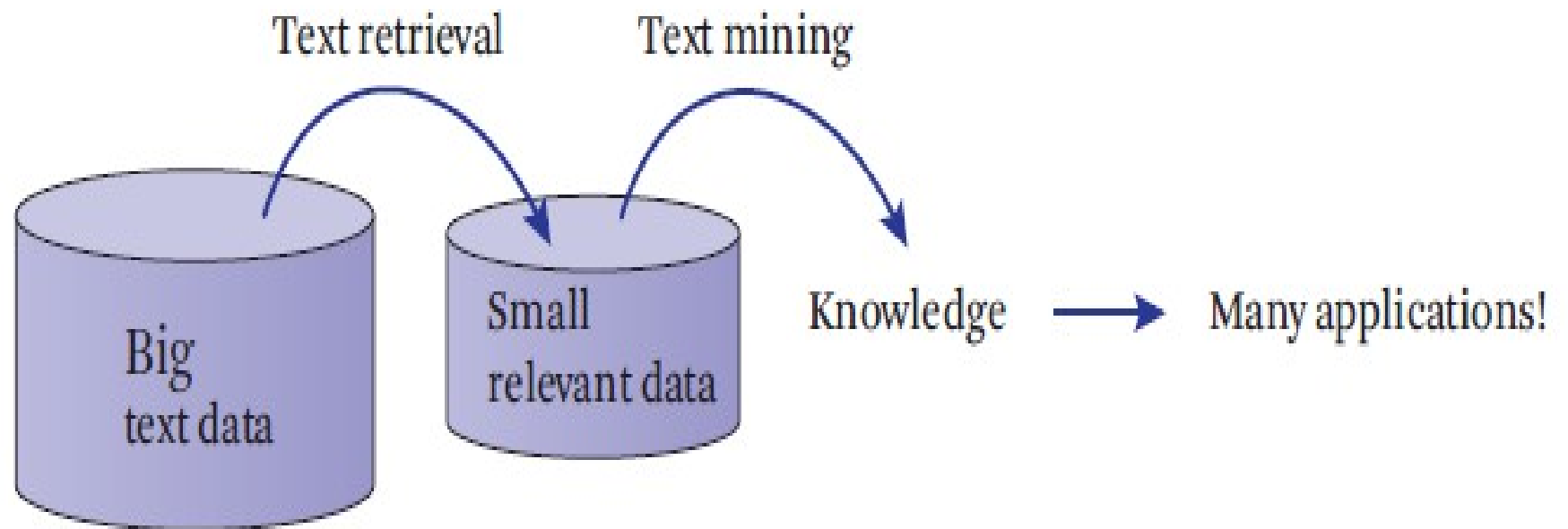
optimize decisions



text mining **is not yet as mature** as search engines

text has less explicit structure

the development of intelligent mining tools requires computers to understand the content encoded in text



**Figure 1.1** Text retrieval and text mining are two main techniques for analyzing big text data.

conectar los tres elementos en **negrita**

# connect the **right info** with the **right user** at the **right time**

a **search engine** enables a user to access text information through querying

modo pull, el usuario tira del sistema, toma la iniciativa (una búsqueda en google p.e.)

a **recommender system** can push relevant information to a user as new info

- modo push (explicado mas adelante)

items become available



**minimum text analysis** sufficient for **matching relevant info** with a **user's info need**

original info items are often delivered to the user in their **original form** (though summaries of the delivered items are often provided)

los summaries son dinamicos (dependen de lo que escriba el usuario en la busqueda))hablaramos mas adelante del tema de los snippets

[EZOPower Car Charger for Apple iPod Nano 6G 5G 4G 3G 2G 1G ...](#)

[www.amazon.com/EZOPower-Charger-Generation.../dp/B004VBJSHA](http://www.amazon.com/EZOPower-Charger-Generation.../dp/B004VBJSHA) ▾

EZOPower Car Charger for Apple iPod Nano 6G 5G 4G 3G 2G 1G / 6th 5th 4th 3rd ...

Save \$2.00 on EZOPower 3.1A Dual-Outlet USB Car Charger - White when ...

users need to read info items to **digest** and exploit the delivered info

# Tecnologías de Gestión de Información No Estructurada



# text information system (TIS)

## Knowledge Acquisition (Text Analysis)

acquire useful **knowledge** encoded in the text data that is not easy for a user to obtain without **synthesizing** and **analyzing** a large portion of the data

por ejemplo basarse en muchas opiniones para dar una recomendacion o respuesta

interesting **patterns buried in text**

a veces la maquina detecta mejor que humanos algunos patrones, incluso revela algunos que se escapan

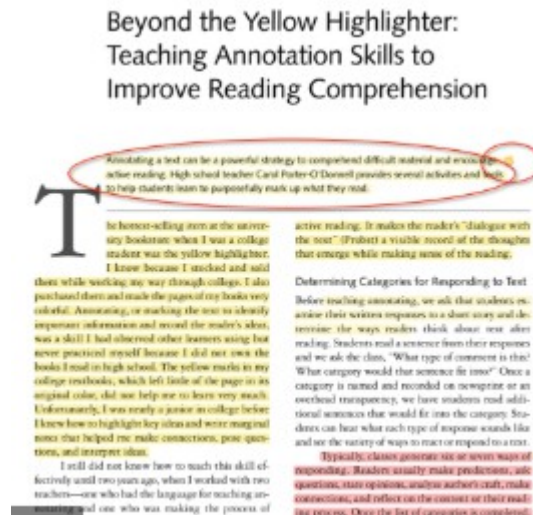


example: a **search engine** returns relevant reviews of a product vs an **analysis engine** that extracts the major positive or negative opinions about the product and compares opinions about multiple products

# text information system (TIS)

## Text Organization

annotate a collection of text documents with meaningful (topical) structures

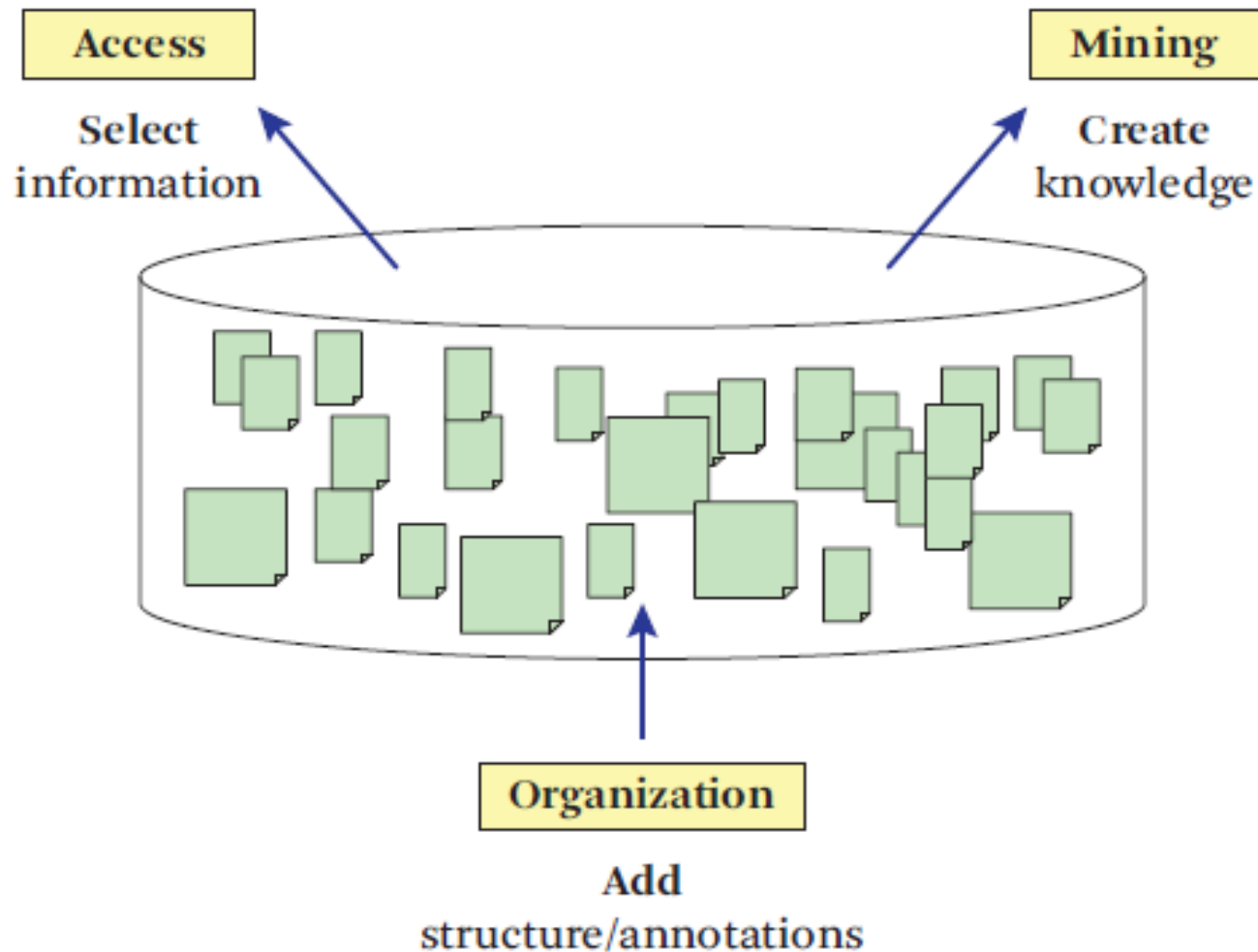


the added structures can allow a user to **search with constraints** on structures or **browse** by following structures

<b>Arts</b> Movies Television Music Performing Arts ...	<b>Business</b> Jobs B2B Construction Investing ...	<b>Computers</b> Internet Software Hardware Programming ...
<b>Finance</b> Insurance Banking Loans Mortgages ...	<b>Games</b> Video RPGs Gambling Card Games ...	<b>Health</b> Conditions Fitness Medicine Beauty ...
<b>Home</b> Cooking Family Consumers Gardening ...	<b>Kids &amp; Teens</b> Games Entertainment Science & Math Society ...	<b>News</b> Current Events Media Newsletters College ...
<b>Recreation</b> Travel Outdoors Humor Pets ...	<b>Reference</b> Museums Maps Education Libraries ...	<b>Science</b> Biology Physics Technology Social Sciences ...
<b>Shopping</b> Gifts Electronics Vehicles Apparel ...	<b>Society</b> Law People Religion Issues ...	<b>Sports</b> Football Baseball Soccer Basketball ...



# text information system (TIS)



facilita un servicio de mayor calidad a los usuarios

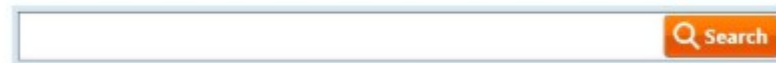
# pull vs push

## pull

the **user takes initiative** to “pull” the useful info out  
from the system

the **system plays a passive role** and waits for a user  
to make a request

e.g. when a user has an ad hoc information need



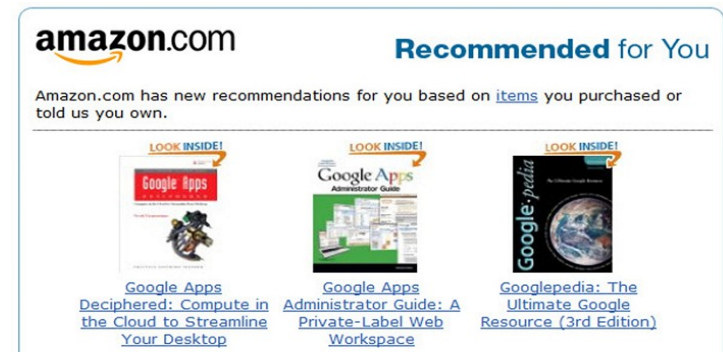
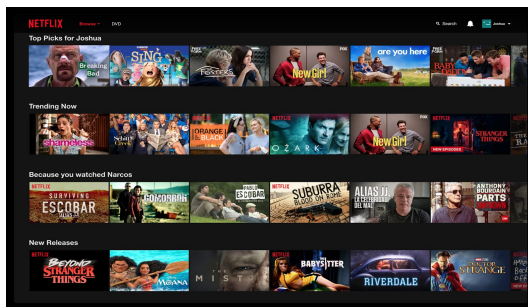
# pull vs push

## push

**the system takes initiative** to “push” (recommend) to the user an info item that the system believes is useful

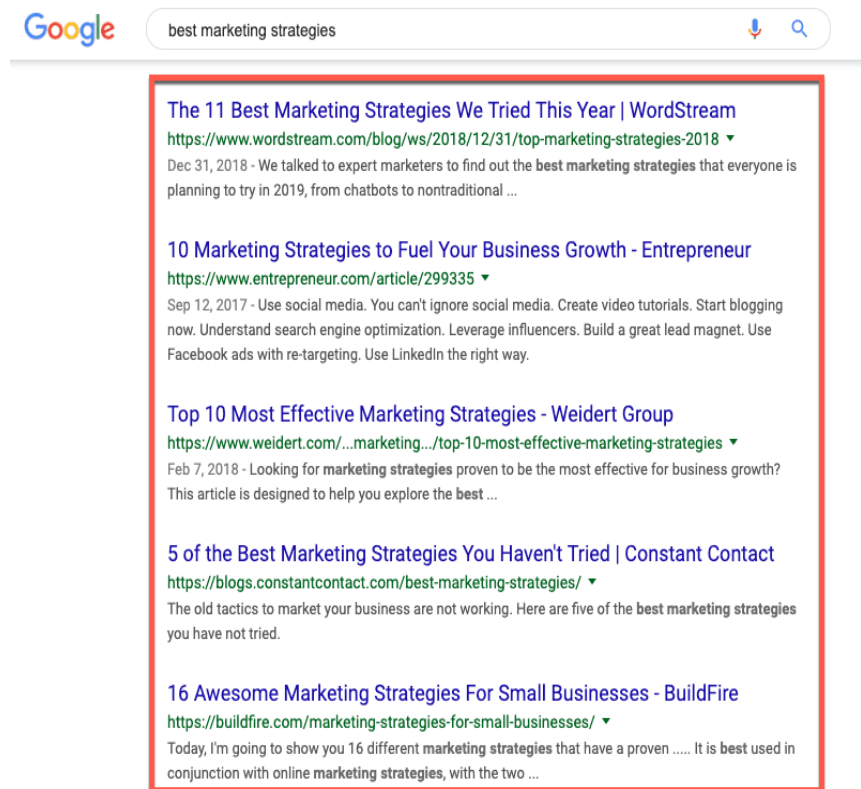
often works well when the user has a relatively **stable info need** (e.g., hobby)

a system can know “in advance” a user’s preferences and interests



# pull mode querying

the user specifies the information need with a (keyword) query, and the system returns docs that are estimated to be relevant



# pull mode

## browsing

the user **navigates** along **structures** that **link** info items together and **progressively** reaches relevant info



browsing & querying are **interleaved naturally**

# the process of **text mining** can be defined as mining text data to discover **useful knowledge**

originalmente era contra base de datos, informacion estructurada, no como ahora

## data mining (DM) vs natural language processing (NLP)

### DM perspective

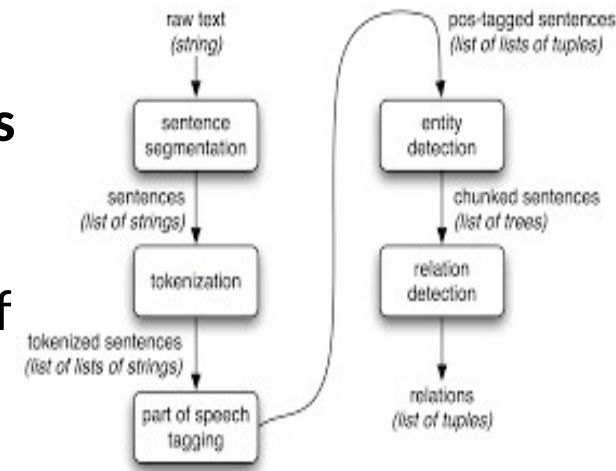
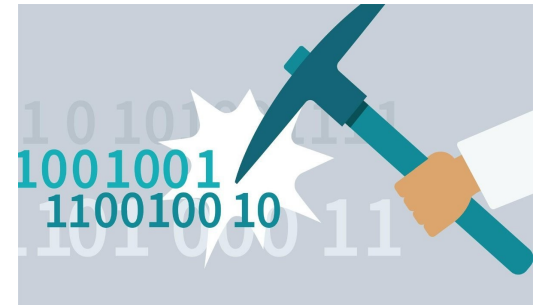
to discover and extract interesting **patterns in text data**  
(latent topics, topical trends, outliers)

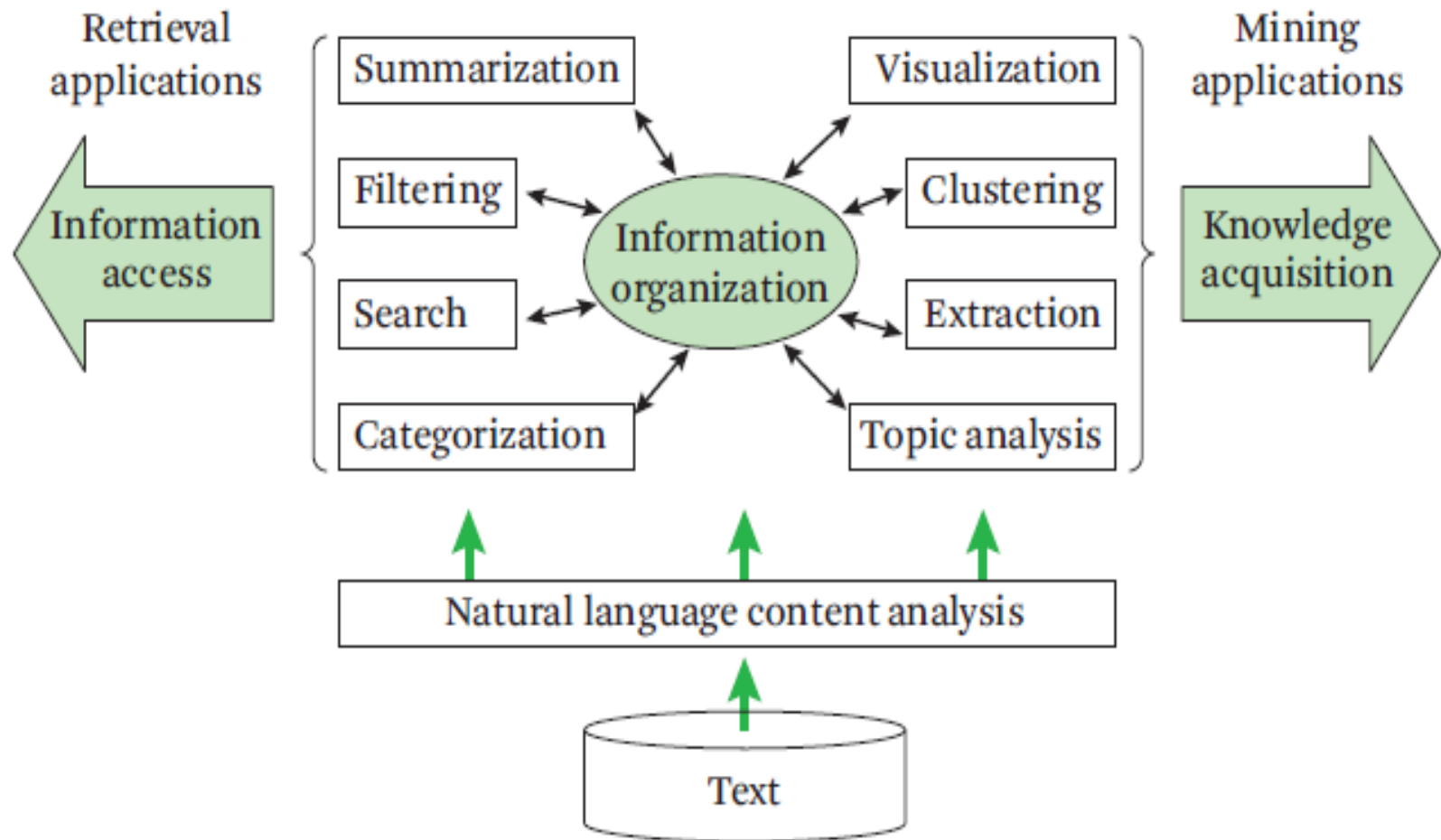
### NLP perspective

to partially **understand NL text**, convert text into some form of **knowledge representation** and make **inferences** based on the extracted knowledge.

problema con las NN profundas que son como una caja negra

**information extraction.** identify and extract mentions of various entities (e.g., people, organization, and location) and their relations (e.g., who met with whom).





**Figure 1.3** Conceptual framework of text information systems.



# Components of TIS

## content analysis

based on **NLP**

transforms **raw text** data into more **meaningful representations**



**statistical** machine learning enhanced with **limited linguistic knowledge**

**shallow** techniques are robust

**deeper** semantic analysis only feasible for very limited domains

some TIS capabilities (e.g., summarizers) require deeper NLP than others (e.g., search).

most TIS use very shallow NLP (e.g., “bag of words”)

modelos muy superficiales



# Components of TIS

## search

take a user's query and return relevant documents



## filtering/recommendation

monitor an incoming stream, decide which items are relevant (or non-relevant) to a user's interest, and then recommend relevant items to the user (or filter out non-relevant items)



**recommender system** (goal is to recommend relevant items to users) vs **filtering system** (whose goal is to filter out non-relevant items to allow a user to keep only the relevant items)

# Components of TIS

## categorization

super tradicional, se usaba para el spam por ejemplo

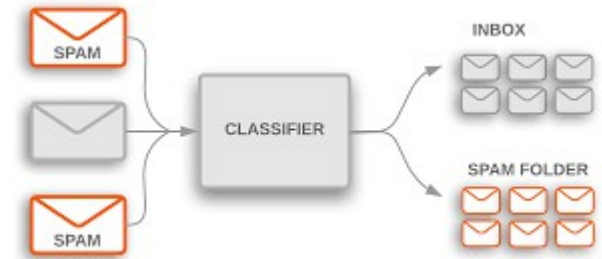
classify a text object into one or several of the **predefined** categories

can annotate text objects with

all kinds of meaningful categories

enriching the representation text data

**organizing** text data and facilitating text access

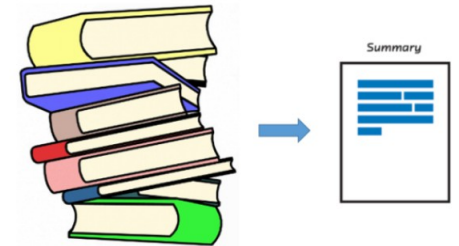


dos tipos de errores aqui: falsos positivos y falsos negativos, a lo mejor hay que dar mas penalty a falsos positivos que a falsos negativos , hay que ver la metrica de rendimiento que se le da al clasificador

## summarization

take one or multiple text documents, and generate a **concise summary** of the essential content.

**reduces human effort** in digesting text information



# Components of TIS

## topic analysis

cojo muchos textos y quiero que la tecnología me descubra los topicos sin yo decirle nada, dando una probabilidad de cada topicico a cada texto

take a set of docs and **extract and analyze topics** in them

tecnología no supervisada no requiere conjunto de entrenamiento

topics directly facilitate **digestion of text data**

los topics NO se los damos nosotros

support **browsing** of text data

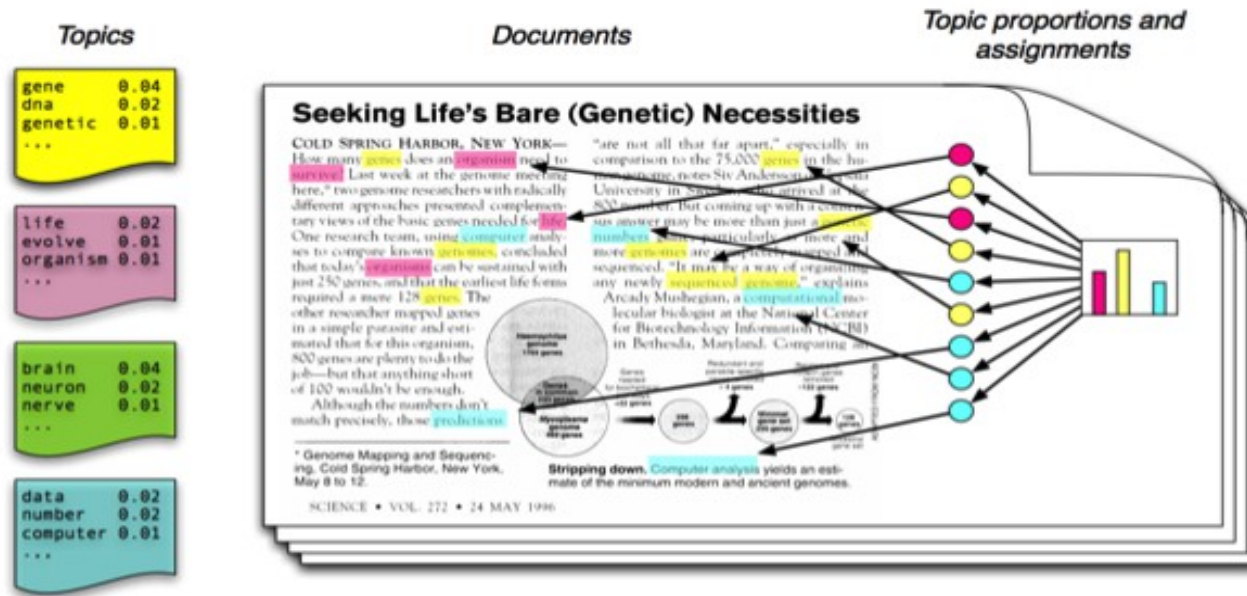


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

# Components of TIS

## topic analysis

can be combined with non-textual data (time, location, authors, and other meta data)

can generate interesting patterns (temporal trends of topics, spatio-temporal distributions of topics, etc).

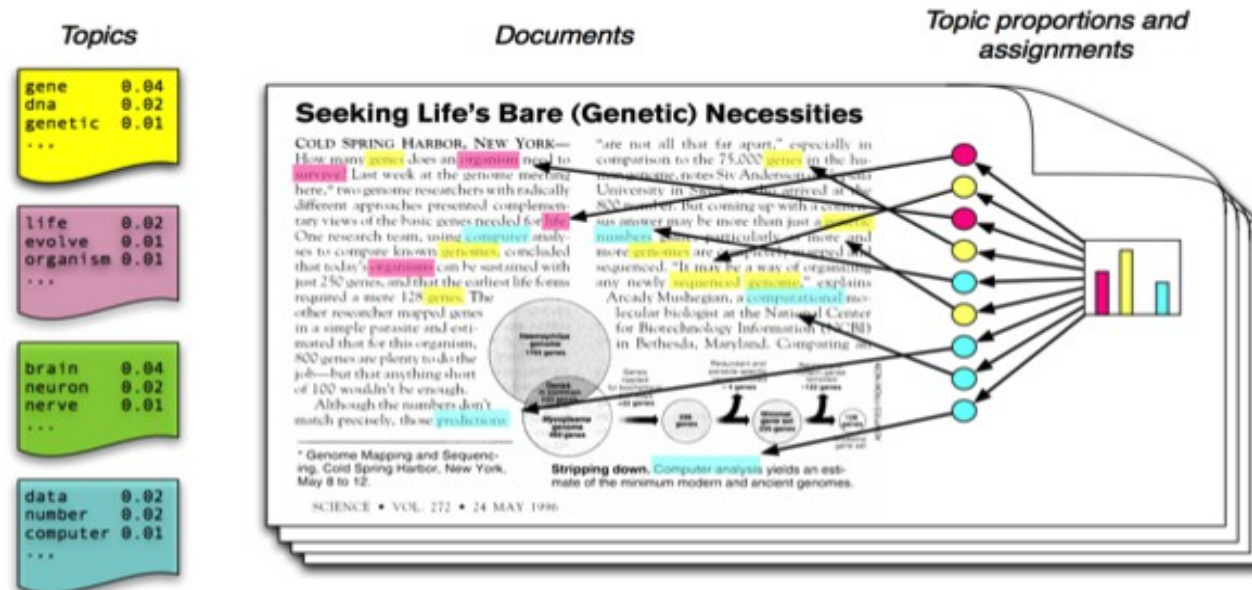


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

# Components of TIS

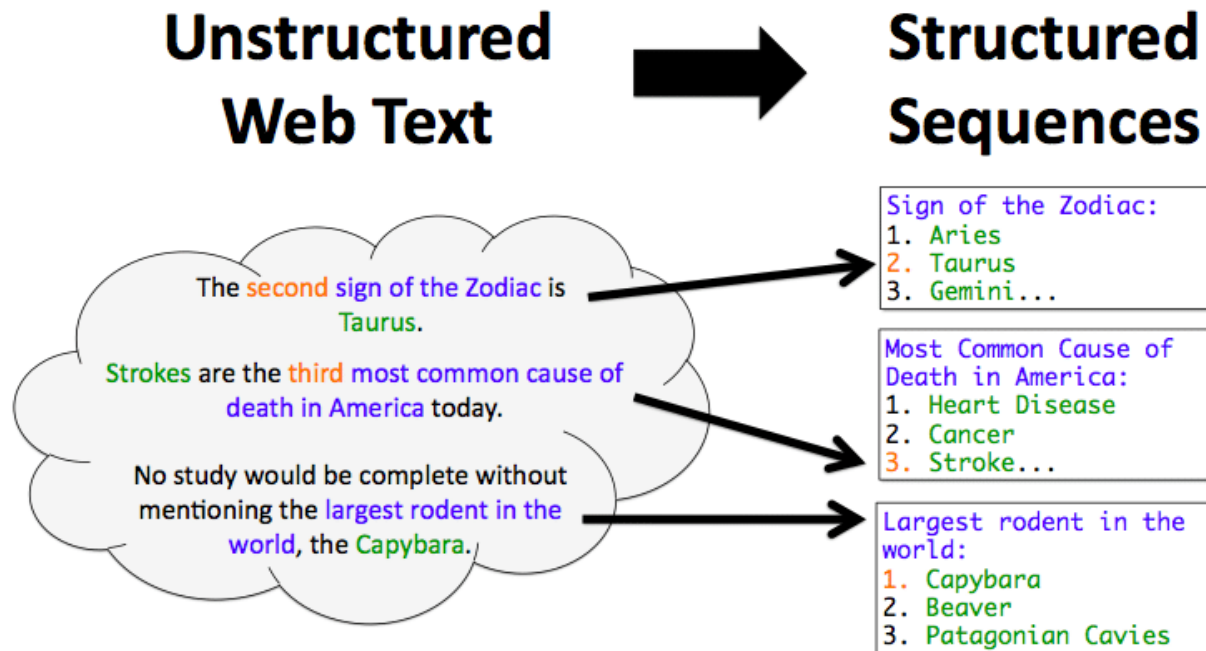
## information extraction

mas ejemplos

extract entities, relations of entities or other  
“knowledge nuggets” from text

## entity-relation graphs

comentaremos mucho acerca de powerlaws



# Components of TIS

## clustering

**discover groups** of similar text objects (e.g., terms, sentences, docs)

**helping users explore** an information space

also useful for discovering **outliers**

