Introduction

CiTIUS Centro Singular de Investigación en Tecnoloxías Intelixentes

USC Universidade de Santiago de Compostela

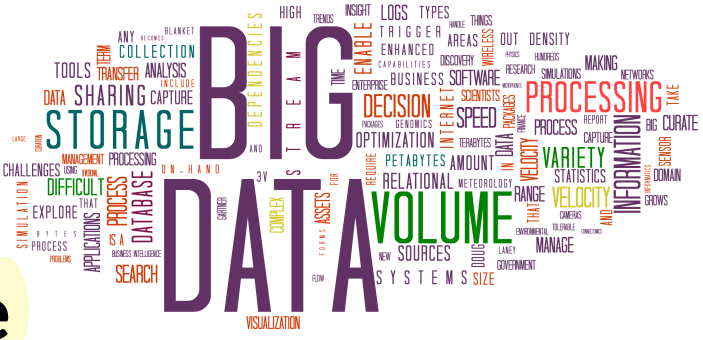
CAMPUS VIDA Campus de Excelencia Internacional

Tecnologías de Gestión de Información No Estructurada

Prof. Dr. David E. Losada

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)

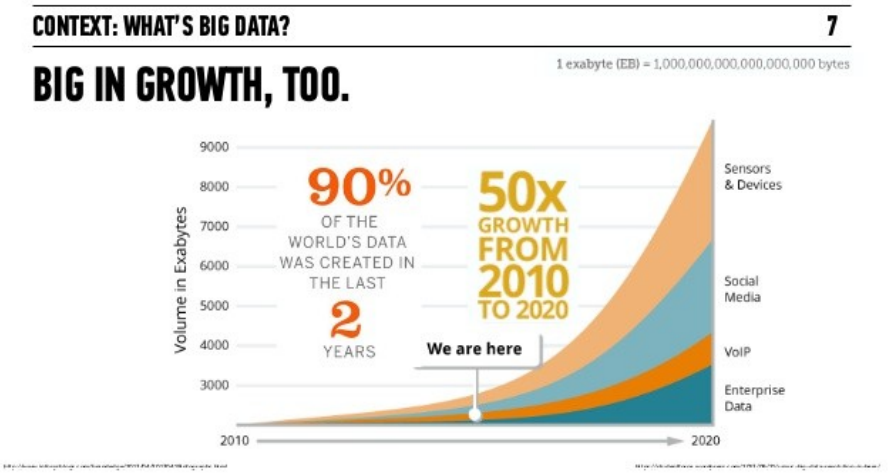
web pages, social media data (e.g., tweets), news, scientific literature, emails, government documents, enterprise data …

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

every day

all kinds of **topics**

Explosive Growth



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**

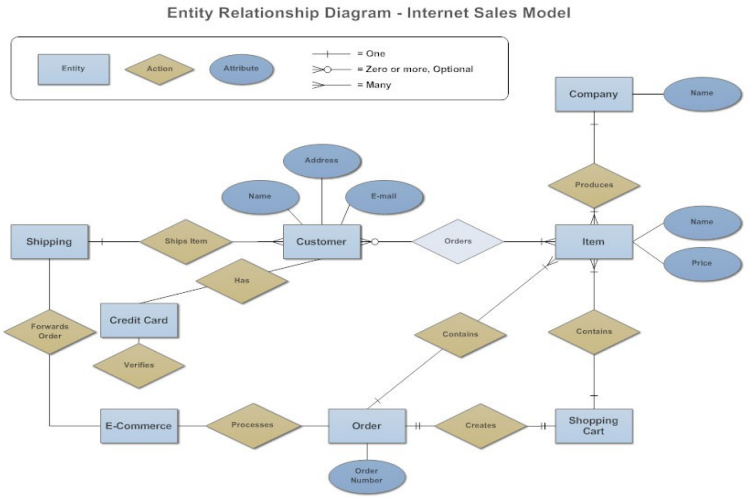
(product reviews, forum discussions, social media)

Structured Data vs Unstructured Data

Structured data:

well-defined schemas

easy for computers to handle

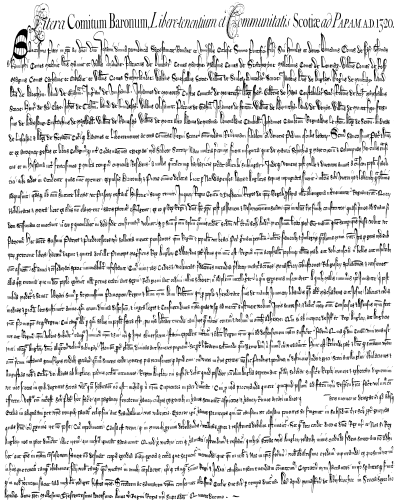


Unstructured data (e.g. text):

less explicit structure

requires computer processing

to **understand** the content



Tecnologías de Gestión de Información No Estructurada

Understanding Text

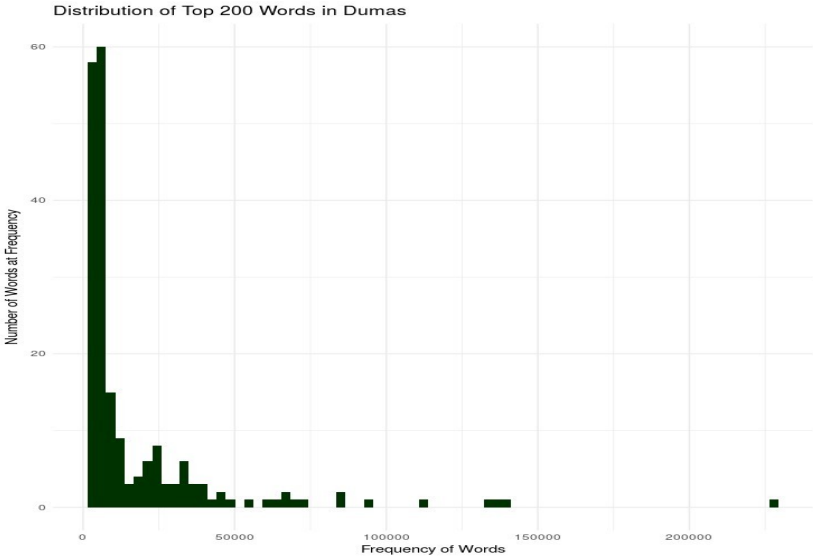
natural language processing (nlp) has not yet reached a point to enable a computer to precisely understand text

statistical & heuristic approaches to management and analysis of text data

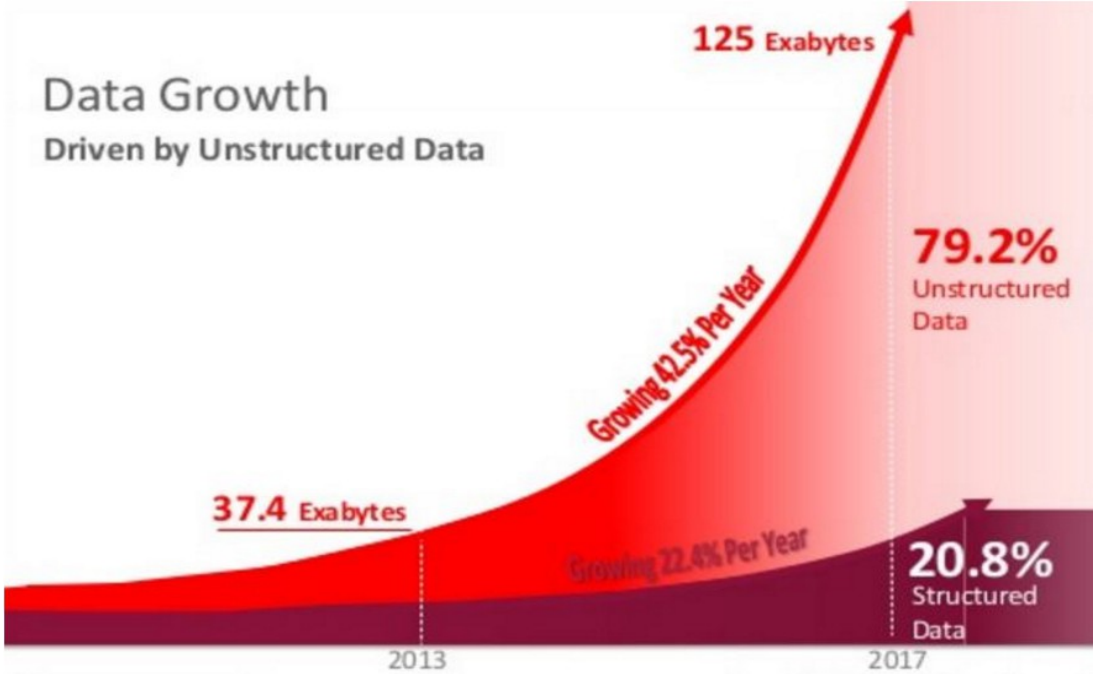
robust

can be applied to any language

& topic



"the world produces between 1 and 2 **exabytes** (1018 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]



Source: ISD - 2014, Structured Data vs. Unstructured Data: The Balance of Power Continues to Shift

A large amount is **textual**

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

text is arguably the most useful kind of info

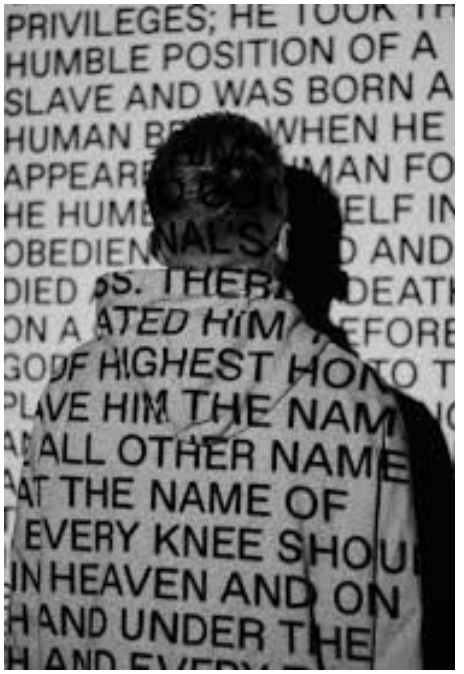
most **natural way** of **encoding human knowledge**

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

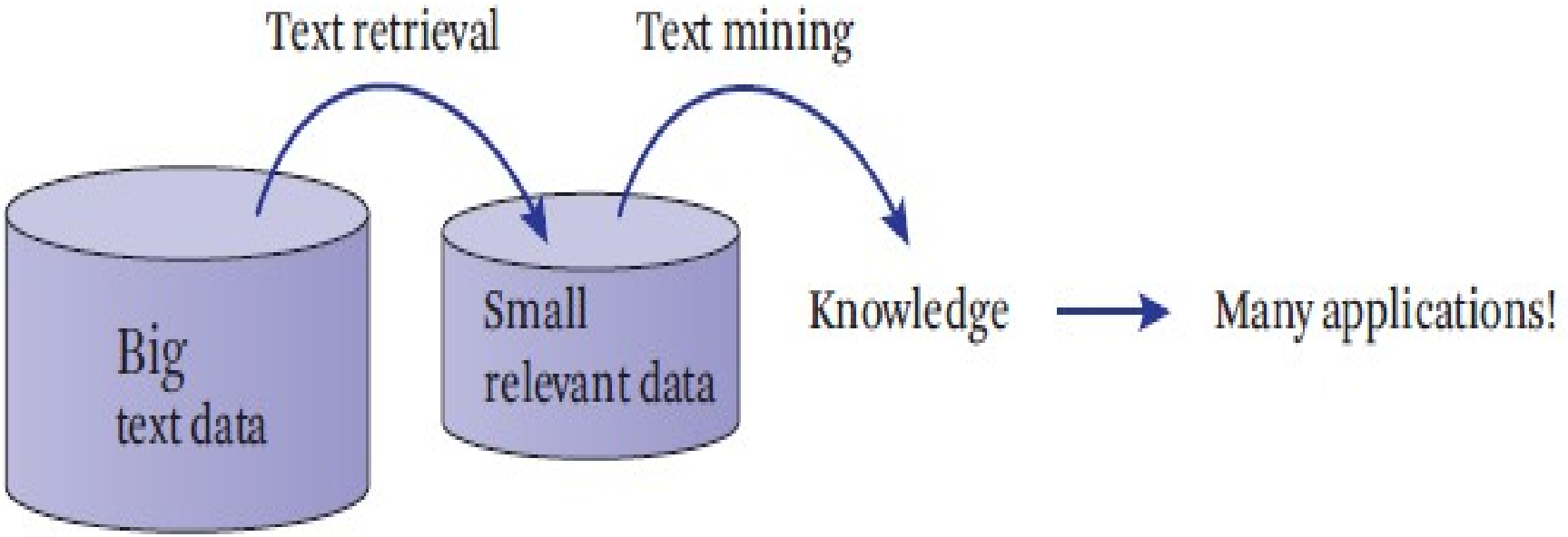


Figure 1.1 Text retrieval and text mining are two main techniques for analyzing big text data.

text information system (TIS)

Information Access

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

- a **search engine** enables a user to access text information through querying
- a **recommender system** can push relevant information to a user as new info 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)

EZOPower **Car Charger** for **Apple iPod Nano** 6G 5G **4G** 3G 2G 1G ...
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

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

interesting **patterns buried in text**

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**

Beyond the Yellow Highlighter:
Teaching Annotation Skills to
Improve Reading Comprehension

Annotating a text can be a powerful strategy to comprehend difficult material and encourage active reading. High school teacher Carol Porter-O'Donnell provides several activities and tools to help students learn to purposefully mark up what they read.

The honor-selling item at the university bookstore when I was a college student was the yellow highlighter. I know because I stocked and sold them while working my way through college. I also purchased them and made the pages of my books very colorful. Annotating, or marking the text to identify important information and record the reader's ideas, was a skill I had observed other learners using but never practiced myself because I did not own the books I read in high school. The yellow marks in my college textbooks, which left little of the page in its original color, did not help me to learn very much. Unfortunately, I was nearly a junior in college before I knew how to highlight key ideas and write marginal notes that helped me make connections, pose questions, and interpret ideas.

I still did not know how to teach this skill effectively until two years ago, when I worked with two teachers—one who had the language for teaching annotating and one who was making the process of active reading. It makes the reader's "dialogue with the text" (Probst) a visible record of the thoughts that emerge while making sense of the reading.

Determining Categories for Responding to Text

Before teaching annotating, we ask that students examine their written responses to a short story and determine the ways readers think about text after reading. Students read a sentence from their responses and we ask the class, "What type of comment is this? What category would that sentence fit into?" Once a category is named and recorded on newsprint or an overhead transparency, we have students read additional sentences that would fit into the category. Students can hear what each type of response sounds like and see the variety of ways to react or respond to a text.

Typically, classes generate six or seven ways of responding. Readers usually make predictions, ask questions, state opinions, analyze author's craft, make connections, and reflect on the content or their reading process. Once the list of categories is completed,

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

Arts
Movies Television Music Performing Arts ...

Business
Jobs B2B Construction Investing ...

Computers
Internet Software Hardware Programming ...

Finance
Insurance Banking Loans Mortgages ...

Games
Video RPGs Gambling Card Games ...

Health
Conditions Fitness Medicine Beauty ...

Home
Cooking Family Consumers Gardening ...

Kids & Teens
Games Entertainment Science & Math Society ...

News
Current Events Media Newspapers College ...

Recreation
Travel Outdoors Humor Pets ...

Reference
Museums Maps Education Libraries ...

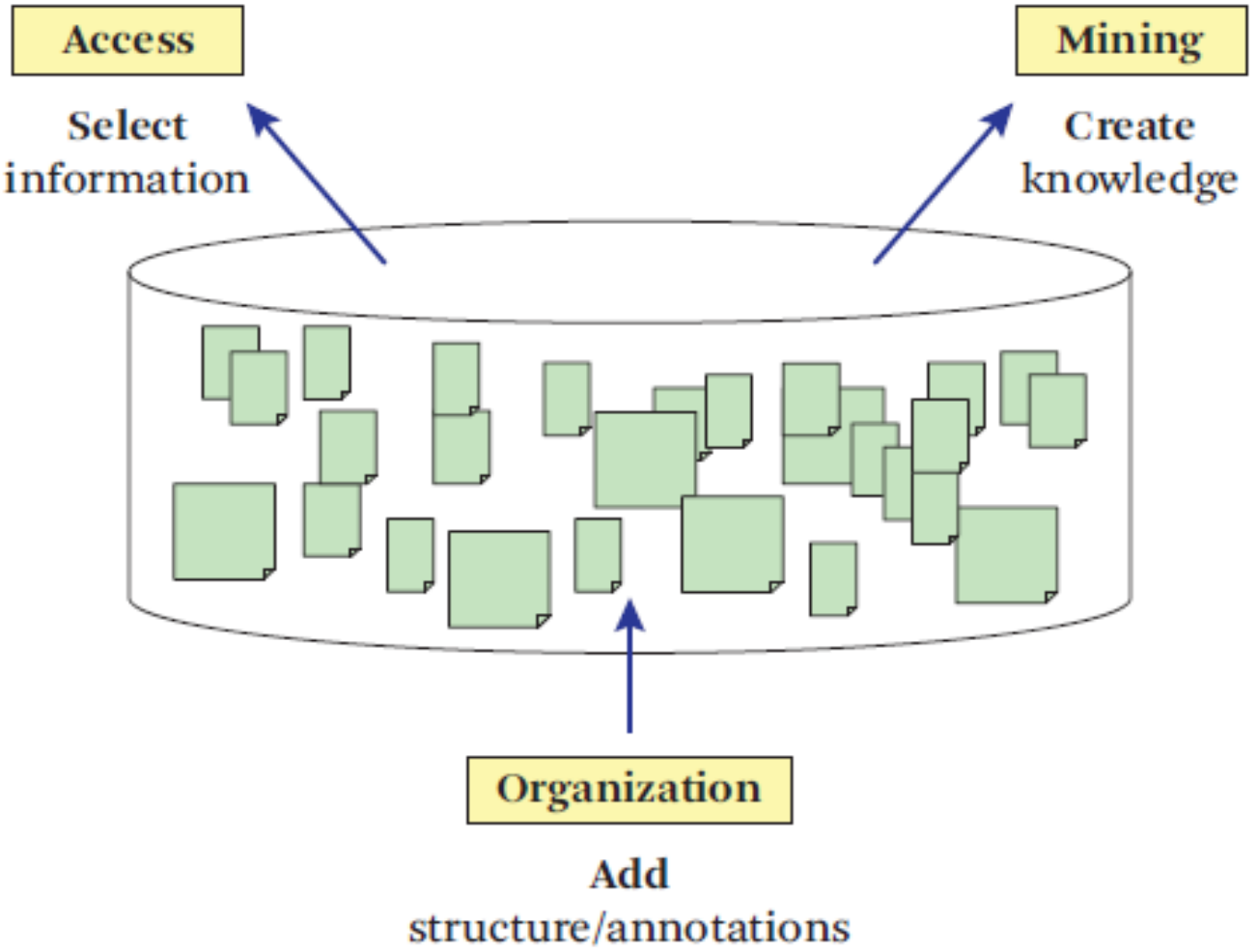
Science
Biology Physics Technology Social Sciences ...

Shopping
Gifts Electronics Vehicles Apparel ...

Society
Law People Religion Issues ...

Sports
Football Baseball Soccer Basketball ...

text information system (TIS)



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

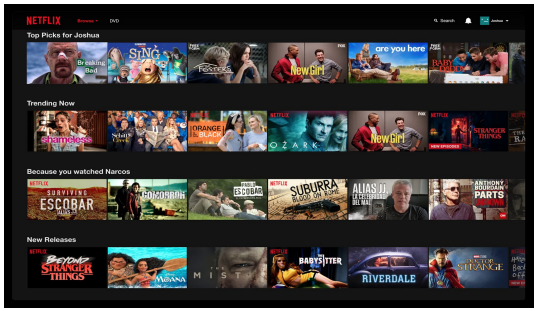
pull vs push

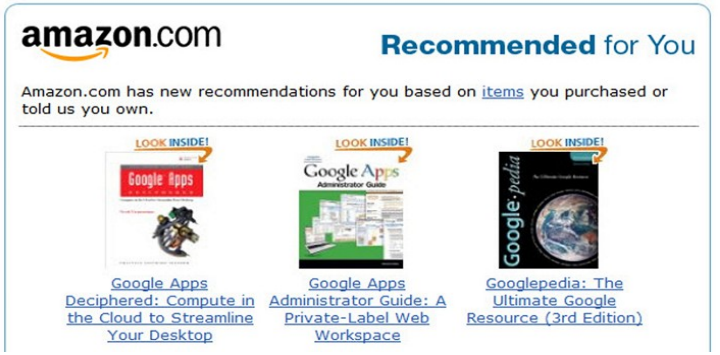
push

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

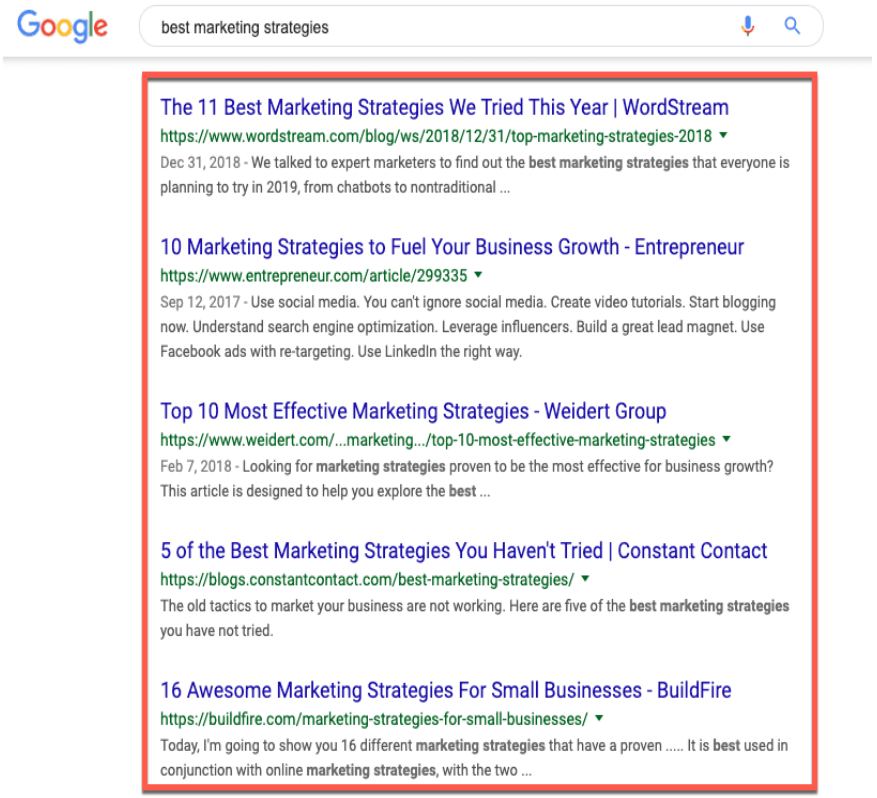




pull mode

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

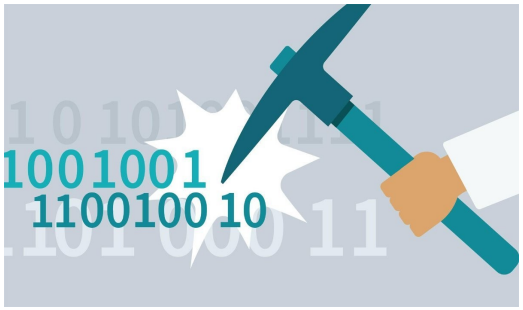
browsing & querying are **interleaved naturally**

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

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

DM perspective

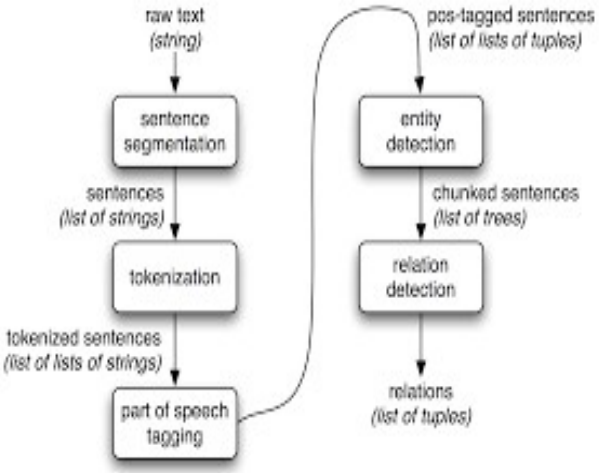
to discover and extract interesting **patterns in text data** (latent topics, topical trends, outliers)



NLP perspective

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

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



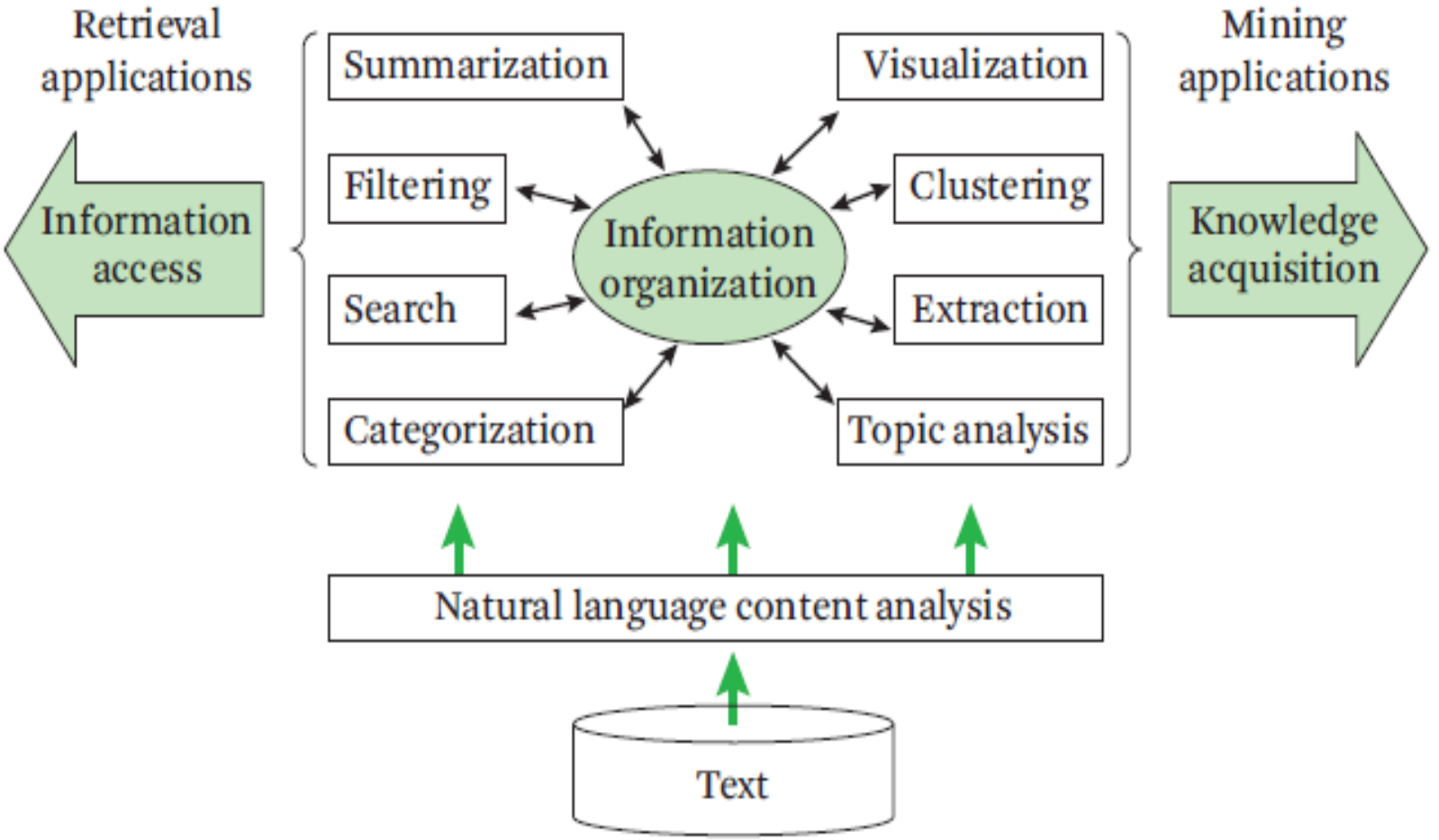


Figure 1.3 Conceptual framework of text information systems.

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”)

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

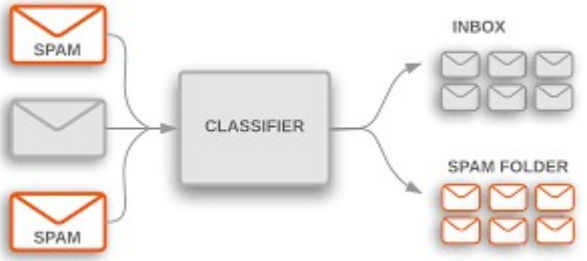
classify a text object into one or several of the **predefined categories**

can annotate text objects with

all kinds of meaningful categories

enriching the representation text data

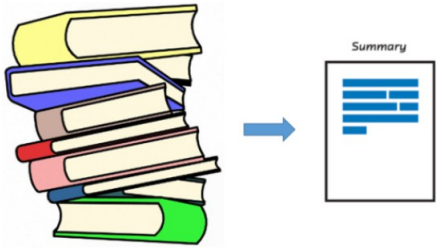
organizing text data and facilitating text access



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

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

topics directly facilitate **digestion of text data**

support **browsing** of text data

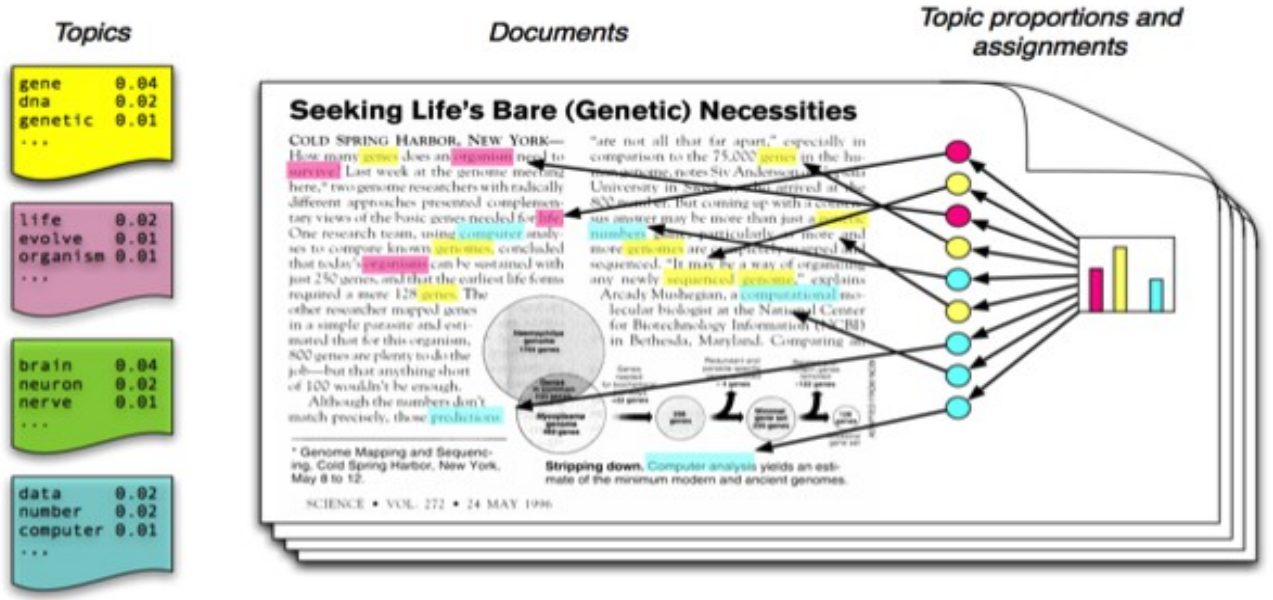


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

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).

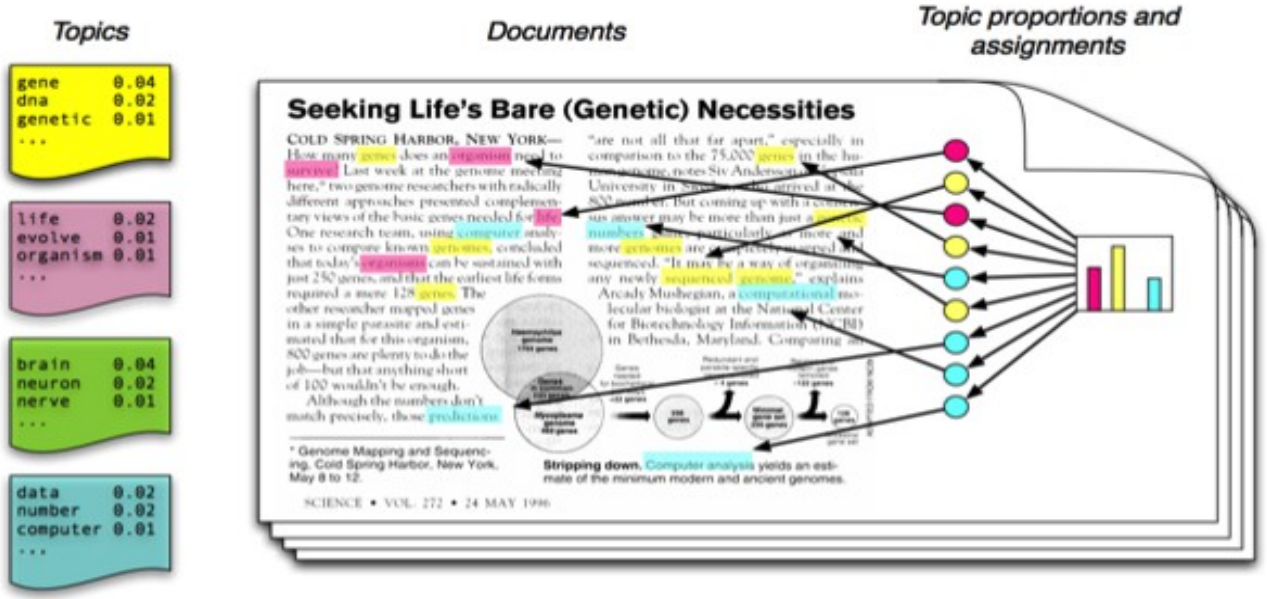


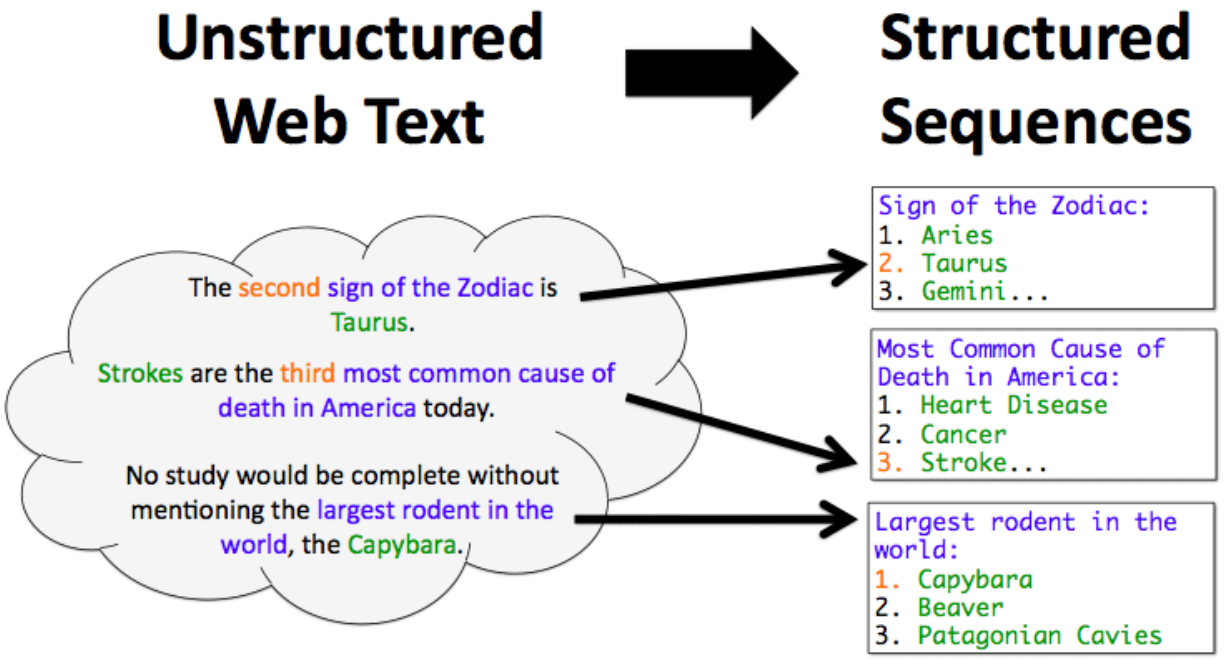
Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Components of TIS

information extraction

extract **entities**, **relations** of entities or other "**knowledge nuggets**" from text

entity-relation graphs



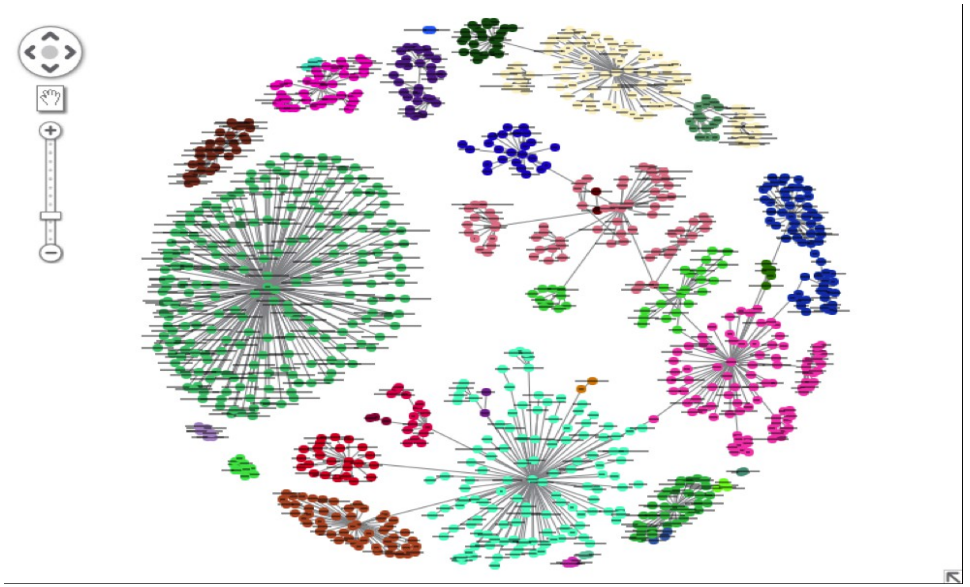
Components of TIS

clustering

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

helping users explore an information space

also useful for discovering **outliers**



Components of TIS

visualization

visually display patterns in text data