SITA 2019

octubre
16-18

Semana de Innovación
Tecnológica en Archivo

DERECHOS RESERVADOS

ARCHIVO GENERAL DE LA NACIÓN COLOMBIA

SITA 2019 octubre 16-18
Semana de Innovación Tecnológica en Archivo

Mincultura

Procesamiento de Lenguaje Natural y Clasificación de Documentos

PhD(c) Luis Gabriel Moreno Sandoval
gabriel.moreno@Lumon.com.co
morenoluis@javeriana.edu.co
Twitter: @gabrielmoreno10

Agenda

- Conocimiento y lenguaje
- Lenguaje digital y tipos de documentos
- Tareas en textos y TLH
  - NLP & AI
  - Machine Learning
- Extracción de Información, Recuperación de Información y Clasificación de Información

DERECHOS RESERVADOS

ARCHIVO GENERAL DE LA NACIÓN COLOMBIA

SITA 2019

Conocimiento y Lenguaje

DERECHOS RESERVADOS

La cultura es de todos Mincultura

ARCHIVO GENERAL DE LA NACIÓN COLOMBIA

SITA 2019

Conectando personas

"El potencial para compartir conocimientos hoy en día no tiene comparación en la historia: nunca antes se había conectado a tanta gente con conocimientos."

DERECHOS RESERVADOS

La cultura es de todos | Mincultura

¿Estar o no estar conectado?

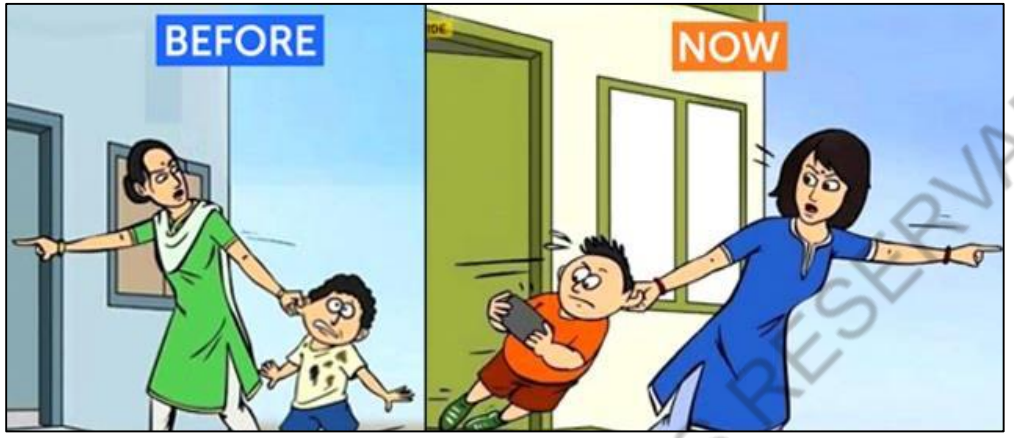
"Estar conectado es bueno, pero estar desconectado durante los últimos millones de años fue la razón principal detrás de nuestra diversidad cultural."

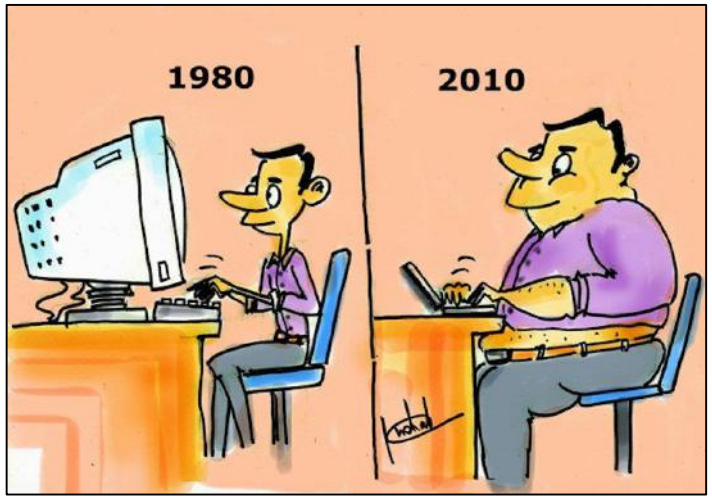
© TODOS LOS DERECHOS RESERVADOS

ARCHIVO GENERAL DE LA NACIÓN COLOMBIA

SITA 2019

¿Estamos desconectados?





TODOS LOS DERECHOS RESERVADOS

La cultura es de todos | Mincultura

La web aun es joven…

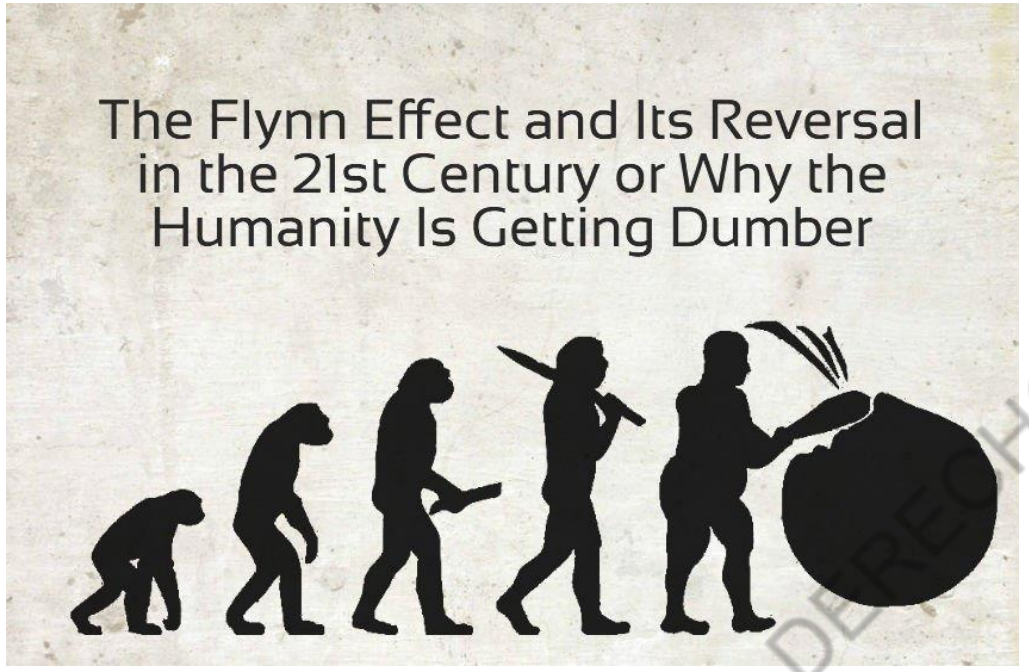
Solo han transcurrido 30 años desde la invención de la Web. Todavía estamos "jugando" con ella, ya que todavía tenemos que descubrir cómo usarlo por completo.

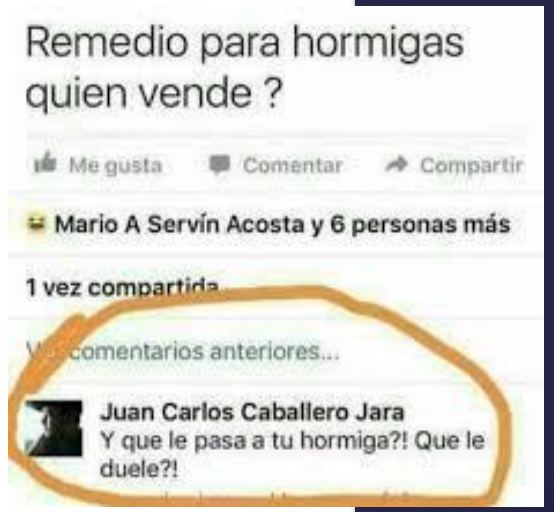
DERECHOS RESERVADOS

ARCHIVO GENERAL DE LA NACIÓN COLOMBIA

SITA 2019

¿Lo estamos usando bien?





DERECHOS RESERVADOS

La cultura es de todos | Mincultura

Una ciudad de conocimiento…

En 1910, Outlet imaginó por primera vez una "ciudad del conocimiento" que serviría como un depósito central para la información del mundo, pero el texto aún no era digital.

https://es.wikipedia.org/wiki/Paul_Otlet

DERECHOS RESERVADOS

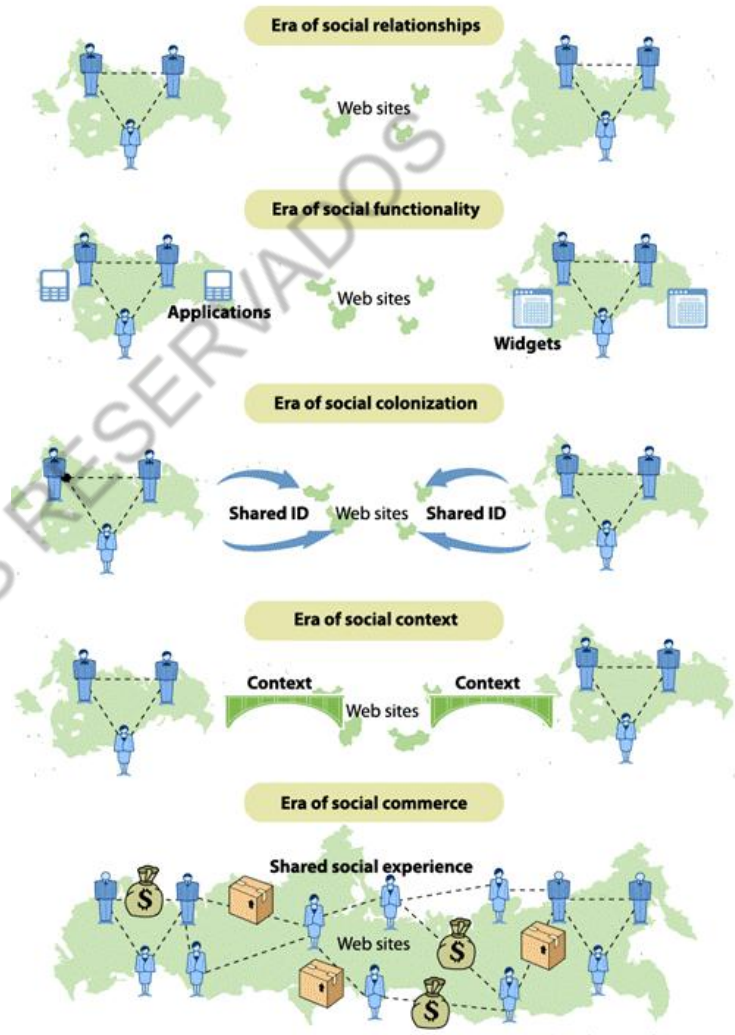
Internet es un laboratorio...

La Web actual no solo representa un almacén de datos ilimitado, sino también un entorno de laboratorio multidisciplinario para experimentos a escala mundial.

DERECHOS RESERVADOS

Era web…

La Web está evolucionando hacia una experiencia social compartida, en la cual los consumidores dependerán de sus pares mientras toman decisiones en línea y darán forma a futuros productos.



E Cambria. Affective computing and sentiment analysis. IEEE Intelligent Systems 31(2), 102-107 (2016)

DERECHOS RESERVADOS

ARCHIVO GENERAL DE LA NACIÓN COLOMBIA

SITA 2019

El poder de la gente…

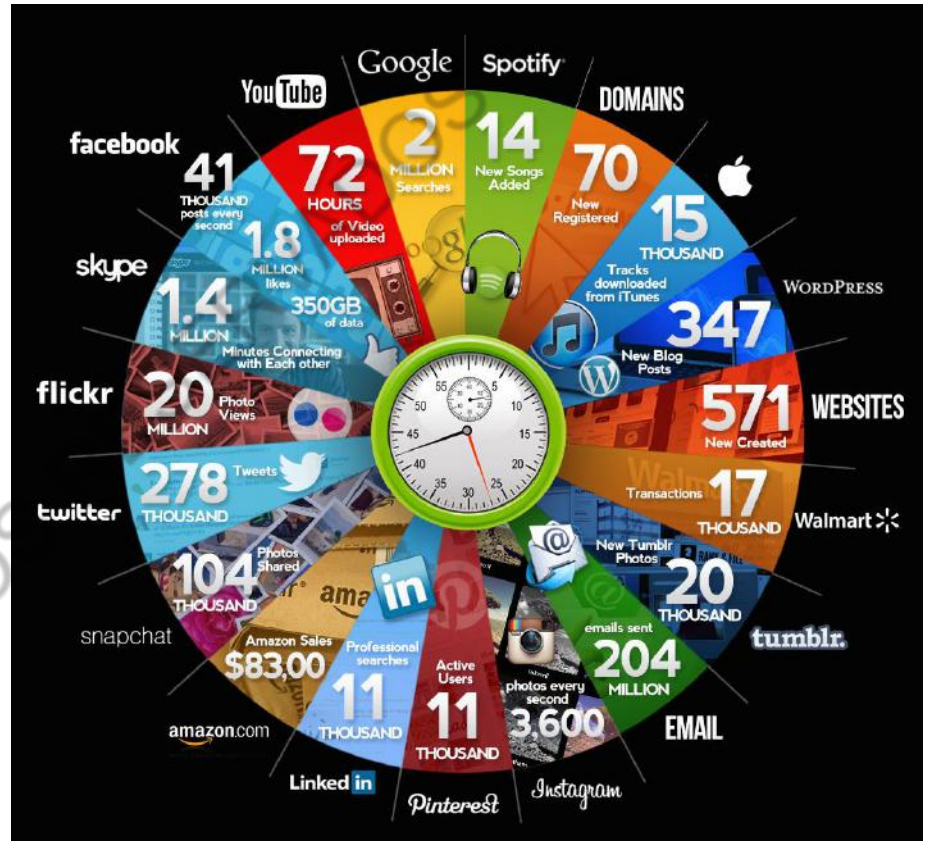
DERECHOS RESERVADOS

Elon Musk Receives Product Suggestion On Twitter, Tesla Implements It 6 Days Later

La cultura es de todos
Mincultura

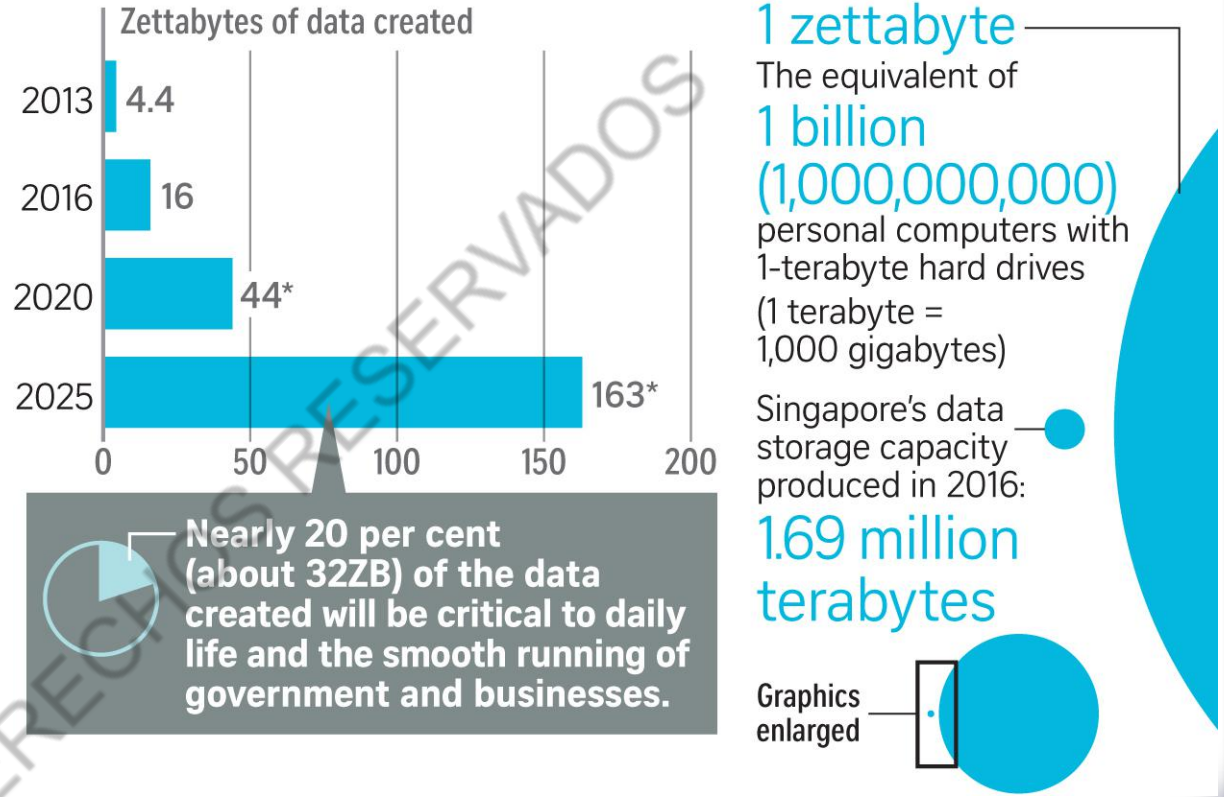
Análisis de datos sociales

Entre los principios de Internet y el año 2003, había cinco exabytes de información en la Web. Ahora, creamos cinco exabytes cada dos días.



DERECHOS

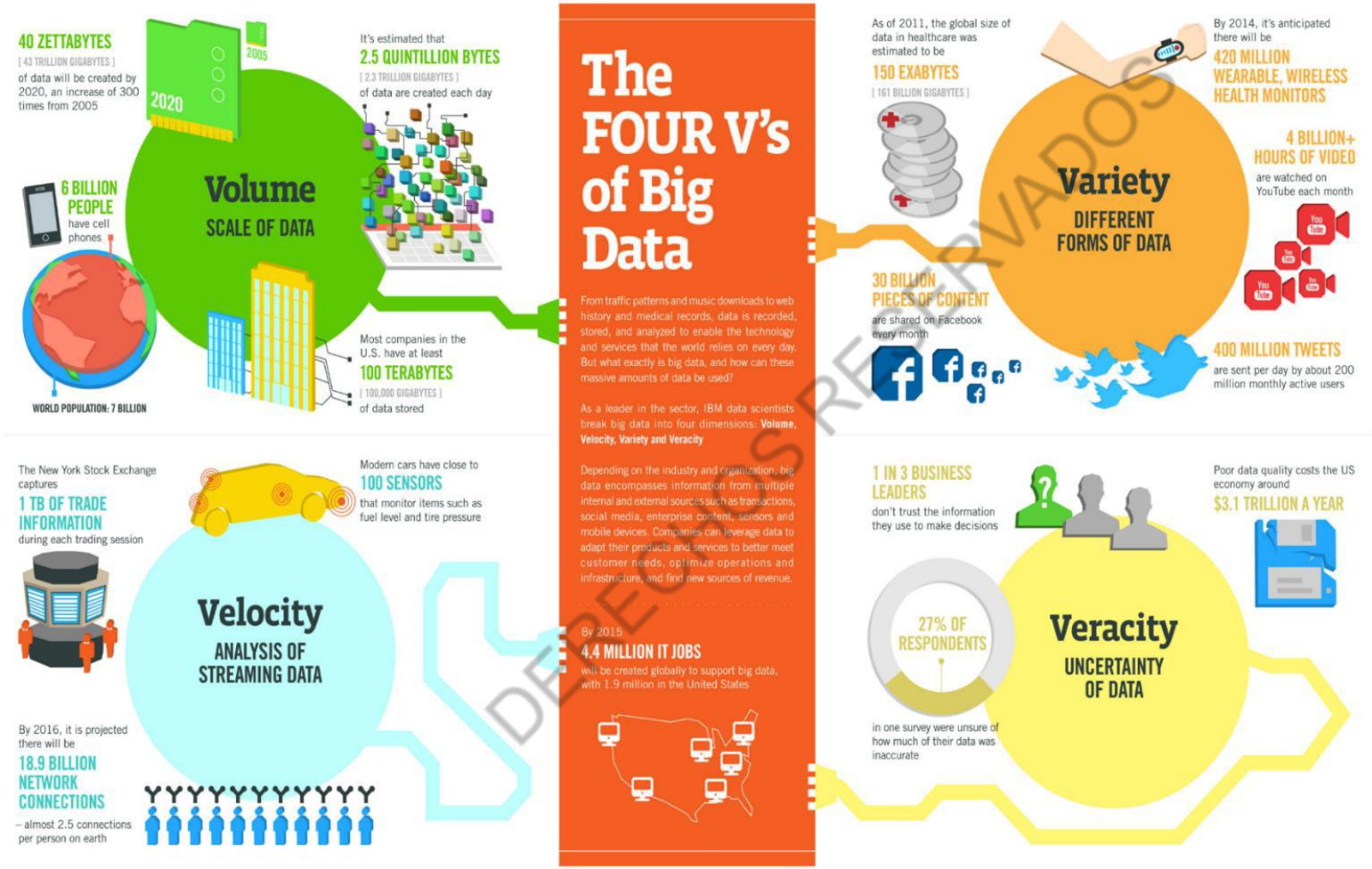
Nos ahogamos en datos?



DERECHOS RESERVADOS

http://straitstimes.com/singapore/world-faces-data-storage-crunch-ahead

No se trata solo del tamaño



DERECHOS RESERVADOS

La cultura es de todos Mincultura

Inteligencia recopilada

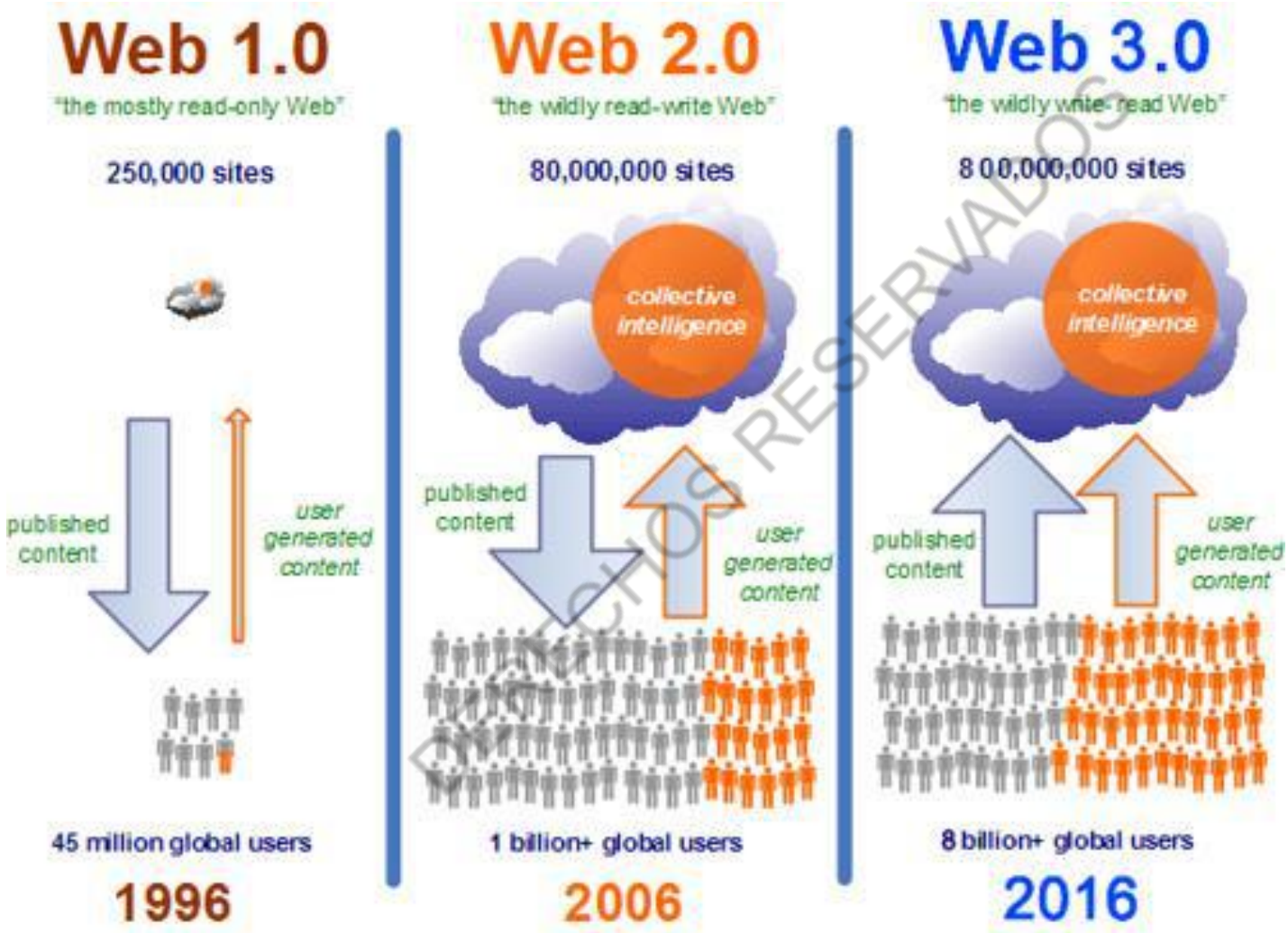
La información actual es extremadamente portátil y procesable. Sin embargo, esta inteligencia recopilada está lejos de tratarse como inteligencia colectiva.

T Gruber. Collective Knowledge Systems: Where the Social Web meets the Semantic Web. Journal of Web Semantics 6(1), 4-13 (2007)

DERECHOS RESERVADOS

ARCHIVO GENERAL DE LA NACIÓN COLOMBIA

The Web 3.0



DERECHOS RESERVADOS

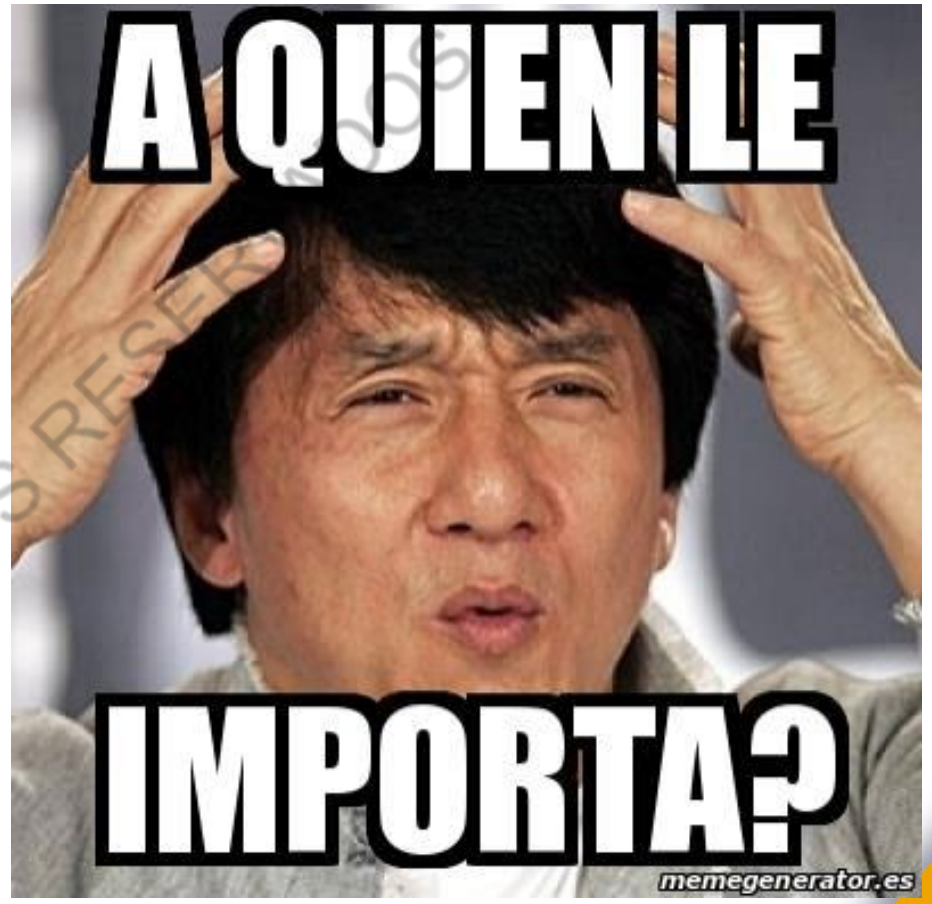
SITA 2019

La cultura es de todos | Mincultura

¿La culpa es de la gente?

Los contenidos en línea están destinados principalmente al consumo humano.

¿Por qué deberían los desarrolladores web y los bloggers preocuparse por hacer que su contenido sea procesable?

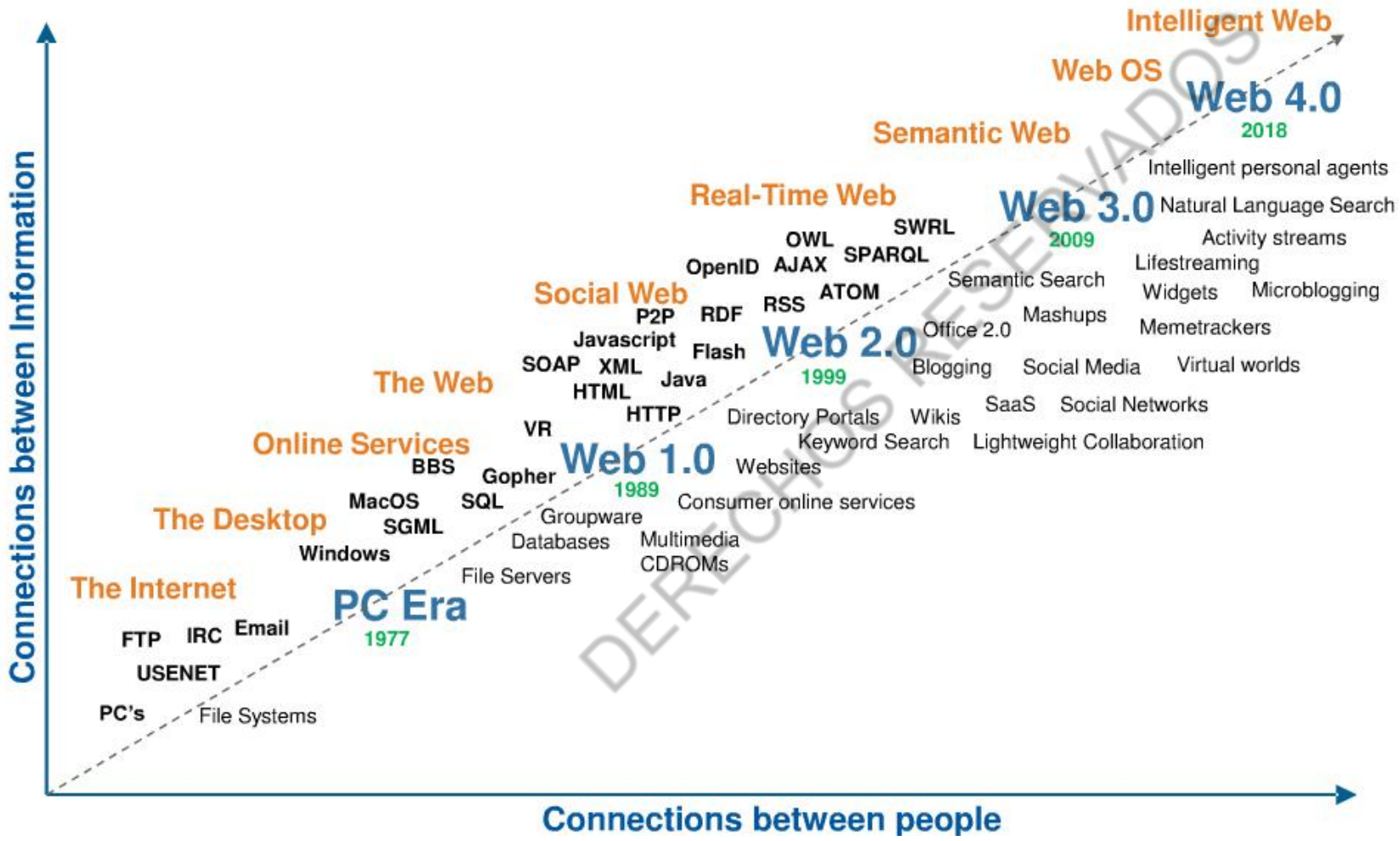


DERECHOS RESERVADOS

ARCHIVO GENERAL DE LA NACIÓN COLOMBIA

SITA 2019

La falla es la tecnología



DERECHOS RESERVADOS

La cultura es de todos
Mincultura

ARCHIVO GENERAL DE LA NACIÓN COLOMBIA

SITA 2019

Lenguaje Digital y Tipos de Documentos

DERECHOS RESERVADOS

La cultura es de todos | Mincultura

La evolución del PLN

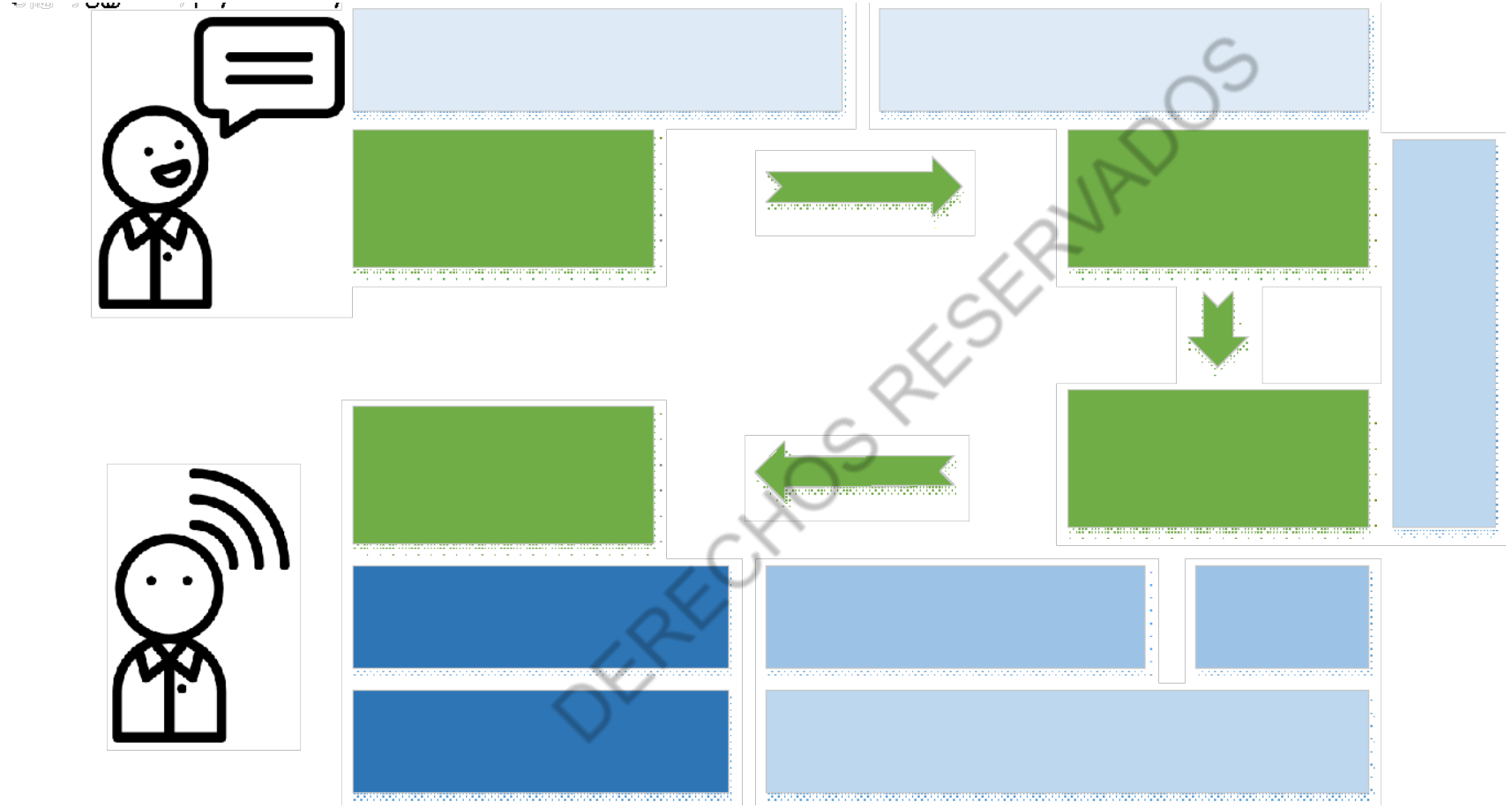
Las tecnologías de PNL evolucionaron desde la era de las tarjetas perforadas (7 minutos por oración) hasta la era de Google y similares (menos de un segundo por oración).

DERECHOS RESERVADOS

ARCHIVO GENERAL DE LA NACIÓN COLOMBIA

Que procesos hay en PLN

SITA 2019



DERECHOS RESERVADOS

La cultura es de todos | Mincultura

Emergencia del PLN

En una web donde UGC (user-generated content) ha alcanzado una masa crítica, el PLN se está convirtiendo en la clave para agregar información, aunque los sistemas todavía están limitados por lo que pueden “ver”.

DERECHOS RESERVADOS

ARCHIVO GENERAL DE LA NACIÓN COLOMBIA

SITA 2019

¿ Realmente estamos viendo todo ?

"El lenguaje está en algún punto intermedio entre la percepción y la comprensión, un material translúcido, de modo que el mundo lleva el tono y el enfoque de lo que expresamos a través de él."

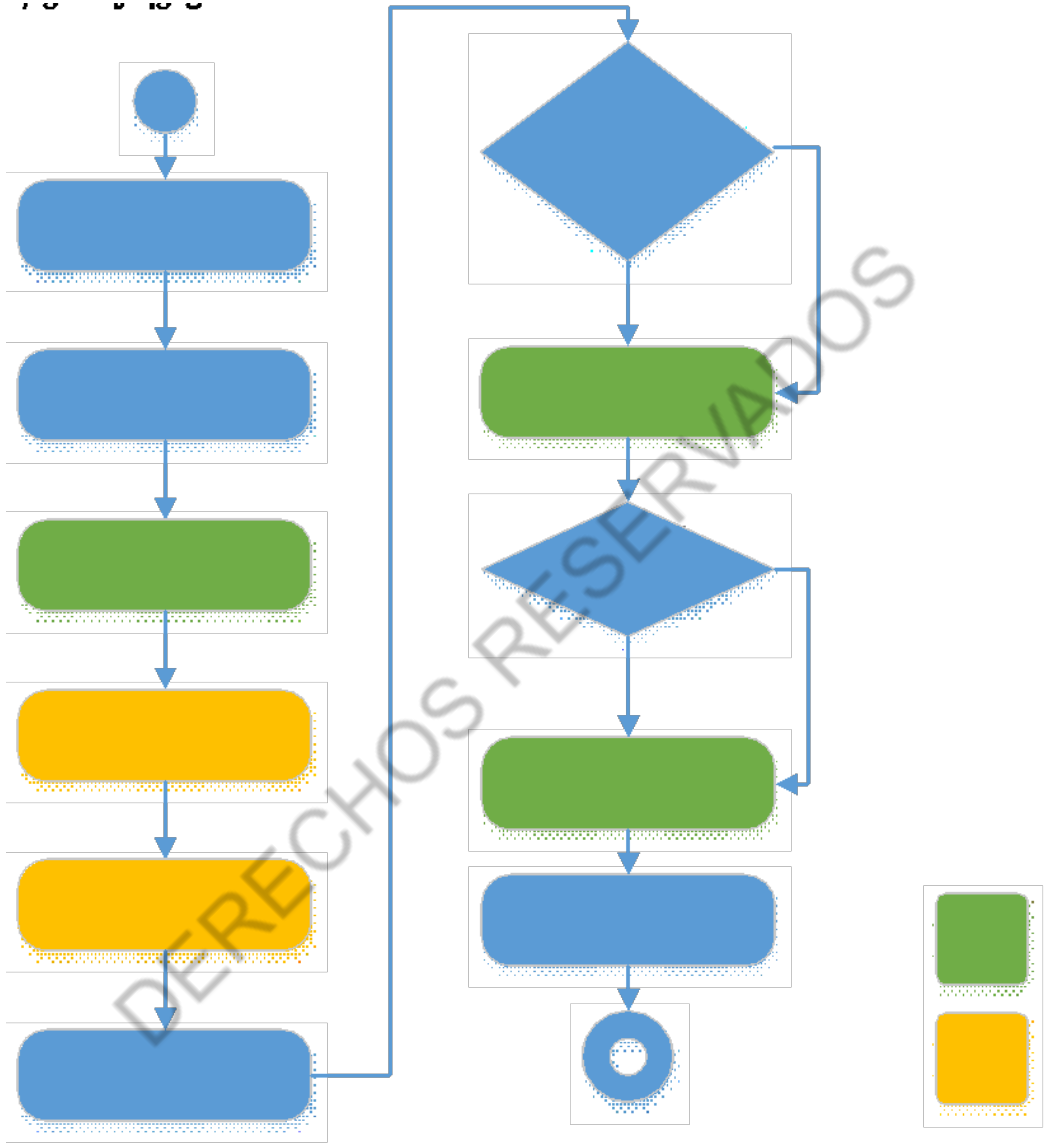
D Davidson. Seeing Through Language. Royal Institute of Philosophy Supplement 42, 15-27 (1997)

DERECHOS RESERVADOS

ARCHIVO GENERAL DE LA NACIÓN COLOMBIA

SITA 2019

Proceso de análisis de Subjetividad Individual



DERECHOS RESERVADOS

La cultura es de todos Mincultura

Entendimiento del Lenguaje Natural

La comprensión del lenguaje natural requiere capacidades simbólicas de alto nivel que la mayoría de las tecnologías de PNL actualmente no poseen.

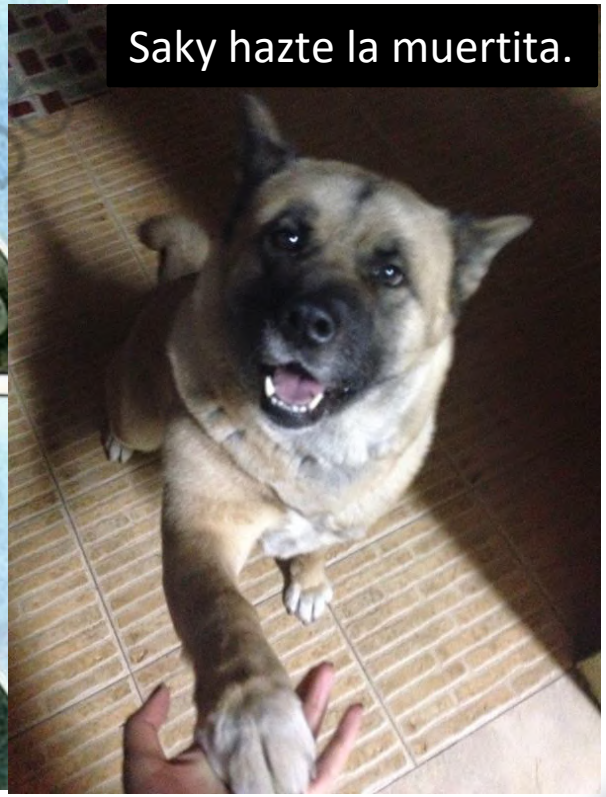
M Dyer. Connectionist natural language processing: A status report. In: Computational Architectures Integrating Neural and Symbolic Processes 292, 389–429 (1995)

- creation and propagation of dynamic bindings;
- manipulation of recursive, constituent structures;
- acquisition and access of lexical, semantic, and episodic memories;
- control of multiple learning/processing modules and routing of information among such modules;
- grounding of basic-level language constructs (e.g., objects and actions) in perceptual/motor experiences;
- representation of abstract concepts.

DERECHOS RESERVADOS

¿ Realmente entendemos ?

Evaluar la inteligencia de los sistemas de inteligencia artificial es como un perro persiguiendo su propia cola: somos los que interpretamos los resultados de los sistemas de inteligencia artificial que construimos.



DERECHOS RESERVADOS

La cultura es de todos Mincultura

Omitiendo un paso

DERECHOS RESERVADOS

http://singularityhub.com/why-hasnt-ai-mastered-language-translation

IA tiene estupidez natural !!!

Una falla clave de la IA es la persistencia en la búsqueda de la mejor manera de resolver un problema, lo que lleva a la creación de sistemas expertos (no inteligentes)

D McDermott. Artificial Intelligence Meets Natural Stupidity, ACM SIGART Bulletin 57 (1976)

DERECHOS RESERVADOS

ARCHIVO GENERAL DE LA NACIÓN COLOMBIA

SITA 2019

Tareas en textos y TLH

DERECHOS RESERVADOS

La cultura es de todos Mincultura

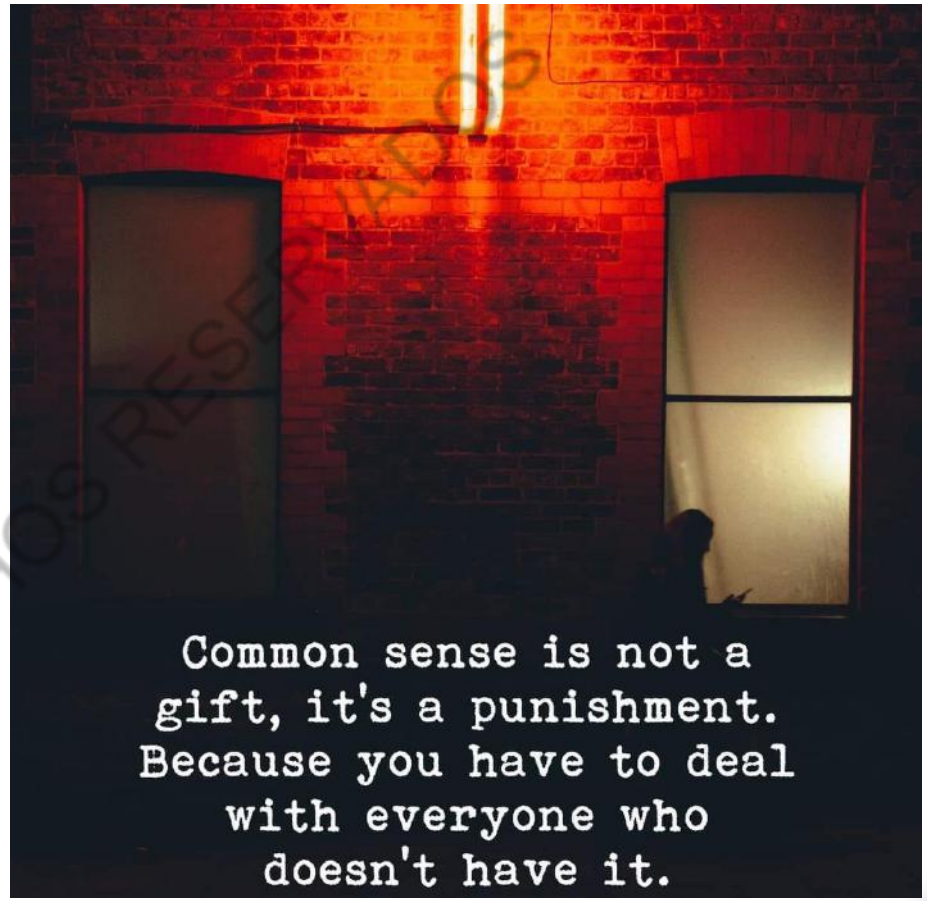
Ceguera es el sentido común

La definición de la IA de hoy es una máquina que puede hacer un movimiento de ajedrez perfecto mientras la sala está en llamas… ;)

DERECHOS RESERVADOS

Conocimiento de sentido común

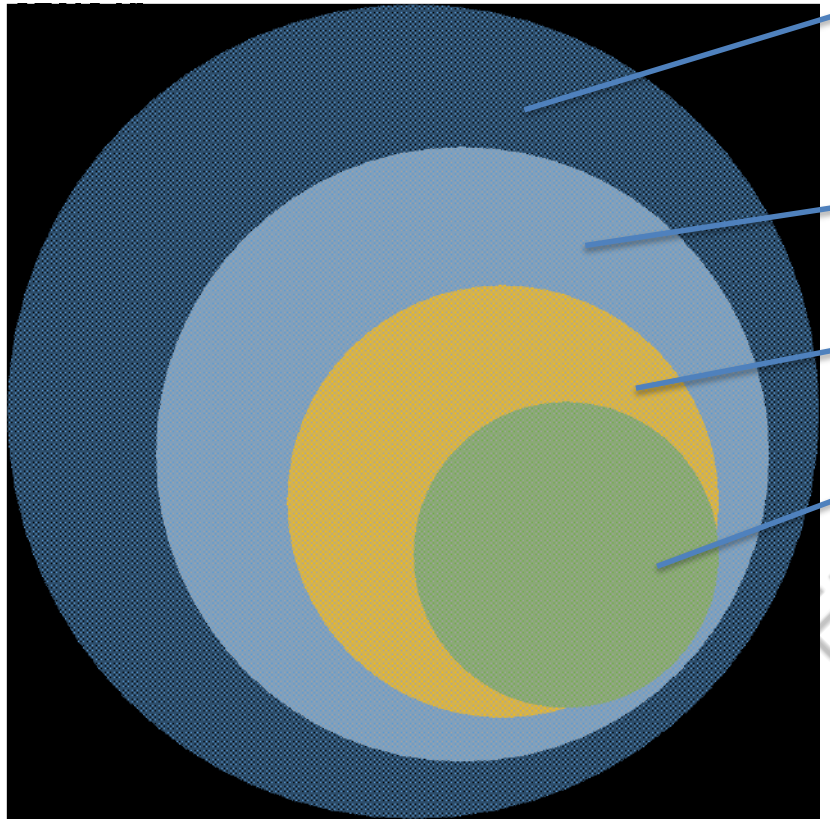
Las personas generalmente solo brindan información útil y dan el resto por sentado. El resto es sentido común: cosas obvias que la gente sabe y generalmente deja sin expresar.



E Cambria, A Hussain, C Havasi, C Eckl. Common Sense Computing: From the Society of Mind to Digital Intuition and Beyond. In: LNCS 5707, 252-259 (2009)

DERECHOS RESERVADOS

Tipos de sentido común



En física, la **inercia** (del latín inertĭa) es la propiedad que tienen los cuerpos de permanecer en su estado de reposo relativo o movimiento relativo.

En la india la **vaca** es un **animal sagrado,** en Colombia no.

El Machine Learning apoya el PLN.

A las 12 siempre almuerzo ;).

DERECHOS RESERVADOS

ARCHIVO GENERAL DE LA NACIÓN COLOMBIA

Tipos de sentido común

SITA 2019

Tesla's autopilot tricked into driving on the wrong side of the road

12

TECHNOLOGY 1 April 2019

DERECHOS RESERVADOS

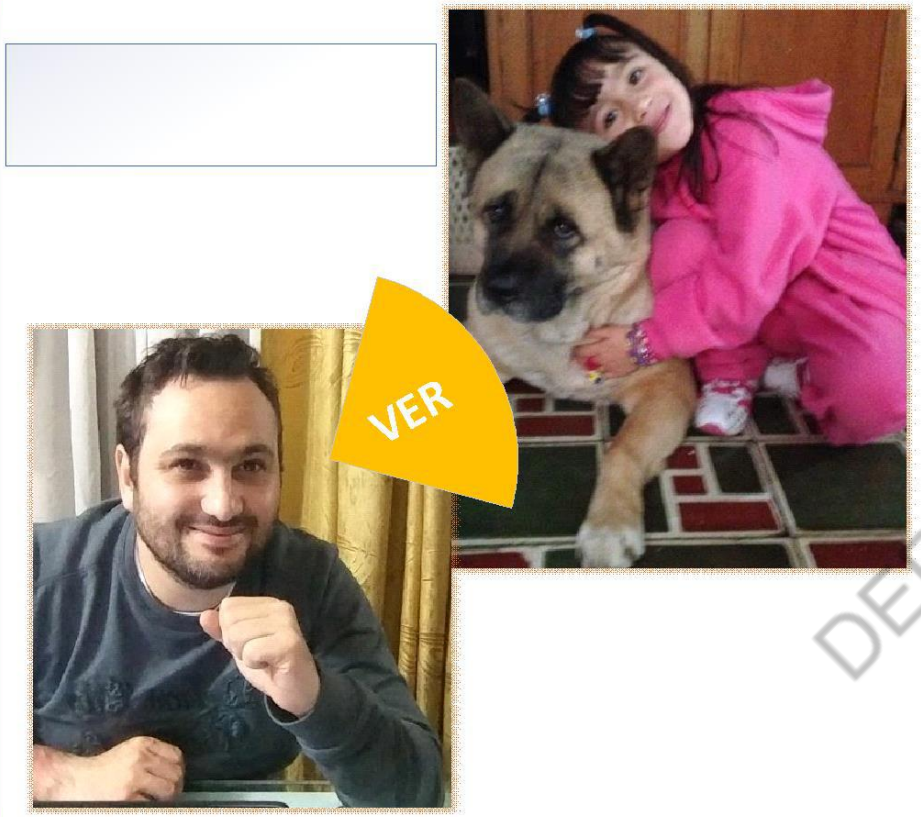
La cultura es de todos Mincultura

Razonamiento de sentido común

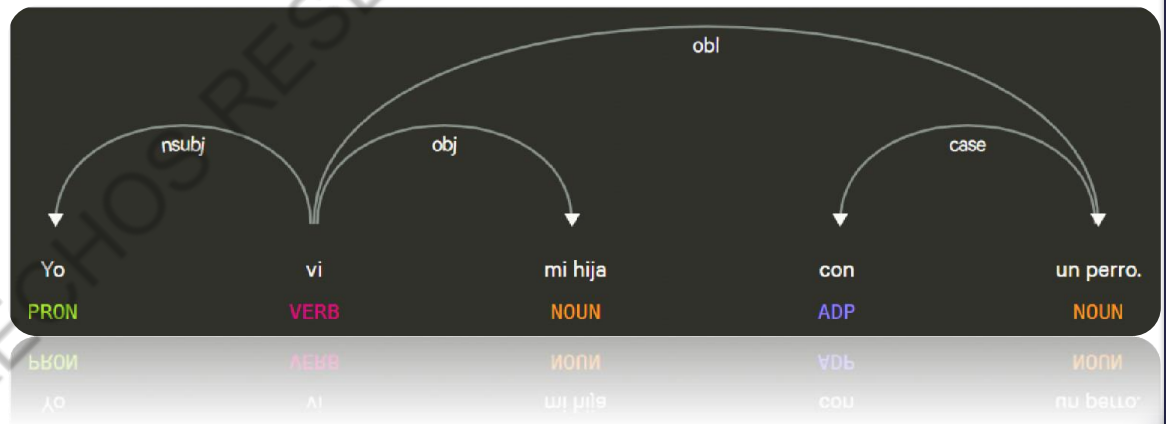
El razonamiento con sentido común es una parte importante de nuestra actividad cognitiva que, en la mayoría de los casos, utilizamos inconscientemente, por ejemplo, para excluir opciones.

DERECHOS RESERVADOS

Pongamos en práctica el sentido común…

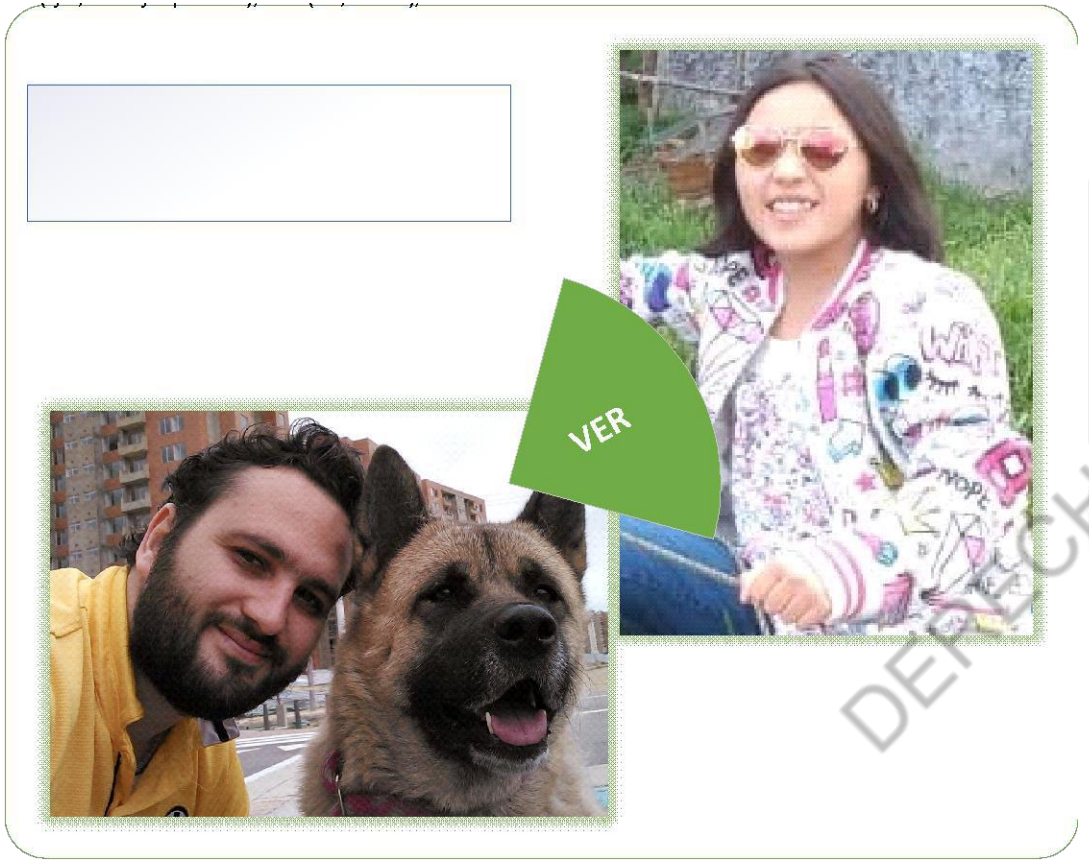


Yo vi a mi hija con un perro.

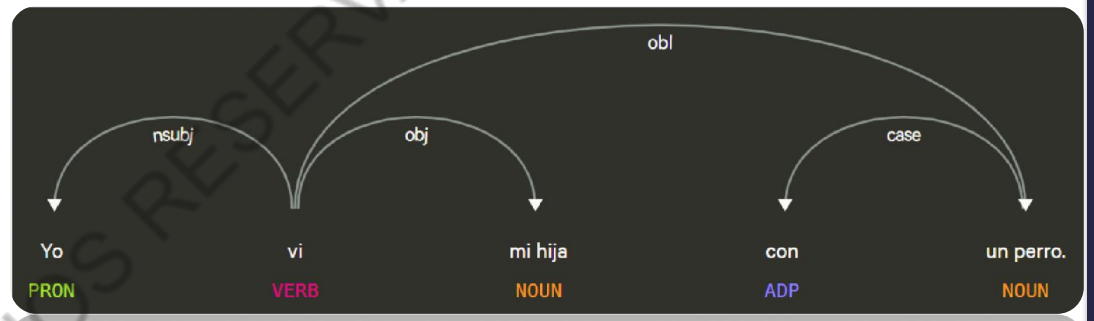


DERECHOS RESERVADOS

Mas sentido común...

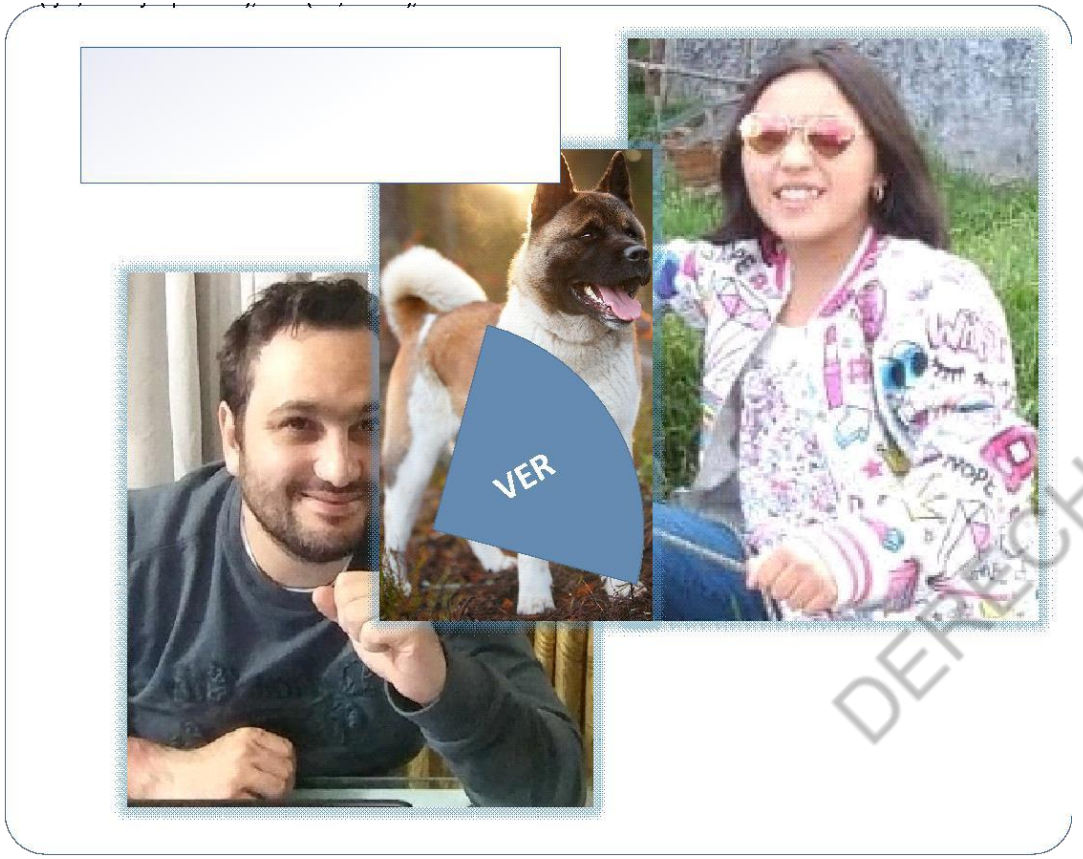


Yo vi a mi hija con un perro.

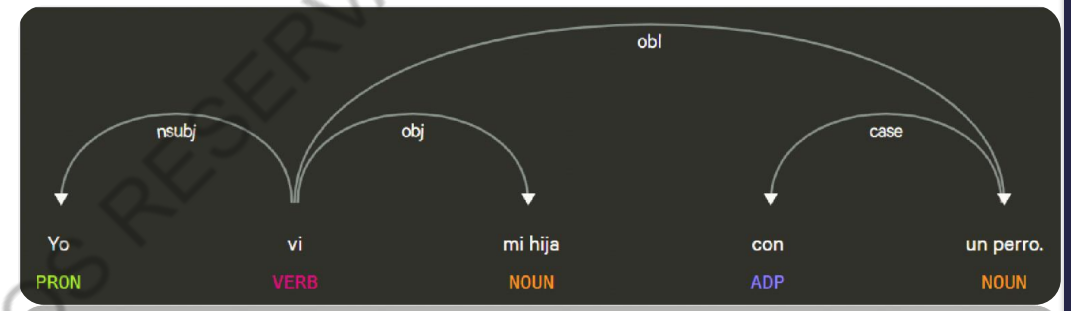


DERECHOS RESERVADOS

Mucho mas sentido común…



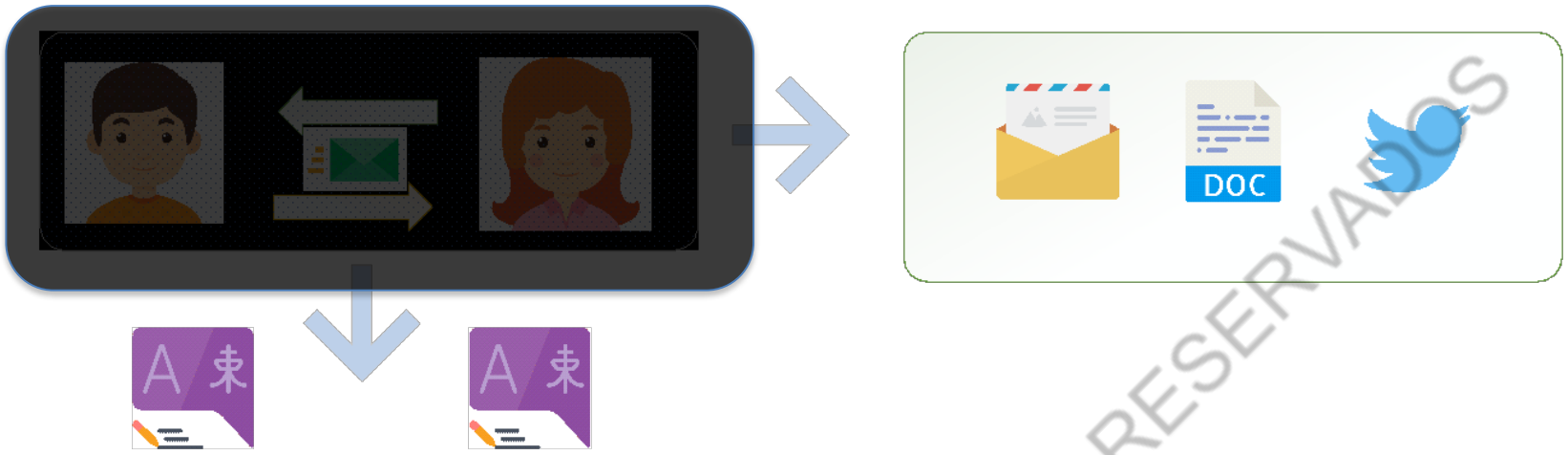
Yo vi a mi hija con un perro.

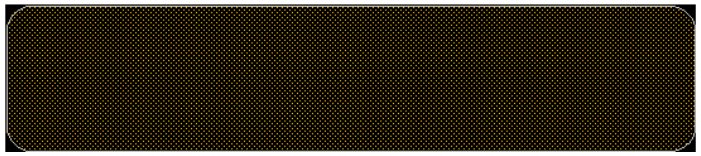


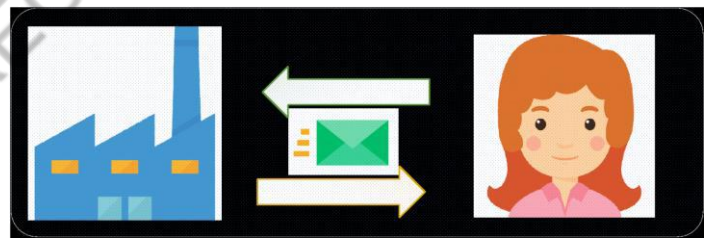
DERECHOS RESERVADOS

La máxima del sentido común…

DERECHOS RESERVADOS



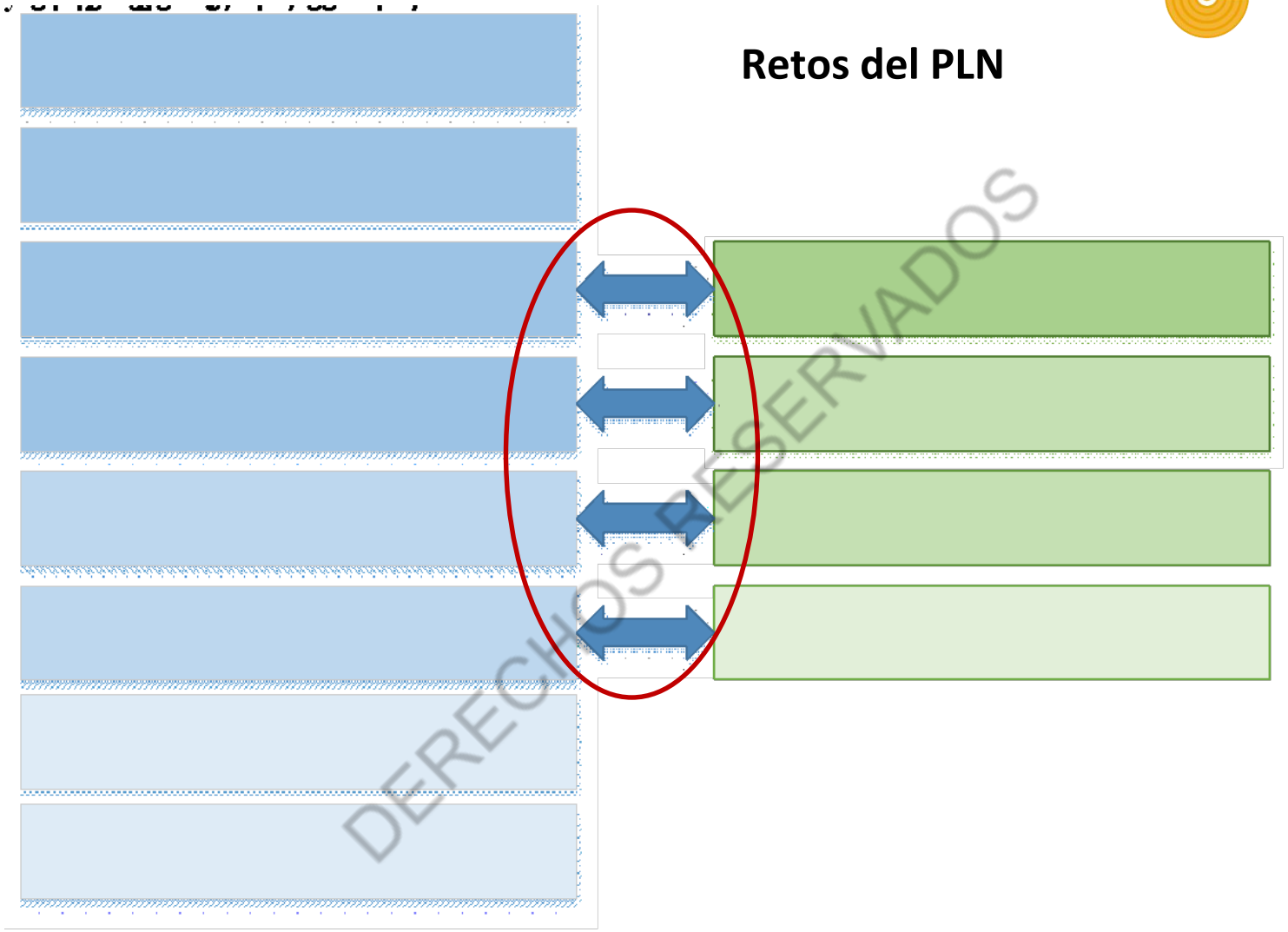




DERECHOS RESERVADOS

DERECHOS RESERVADOS

Retos del PLN



DERECHOS RESERVADOS

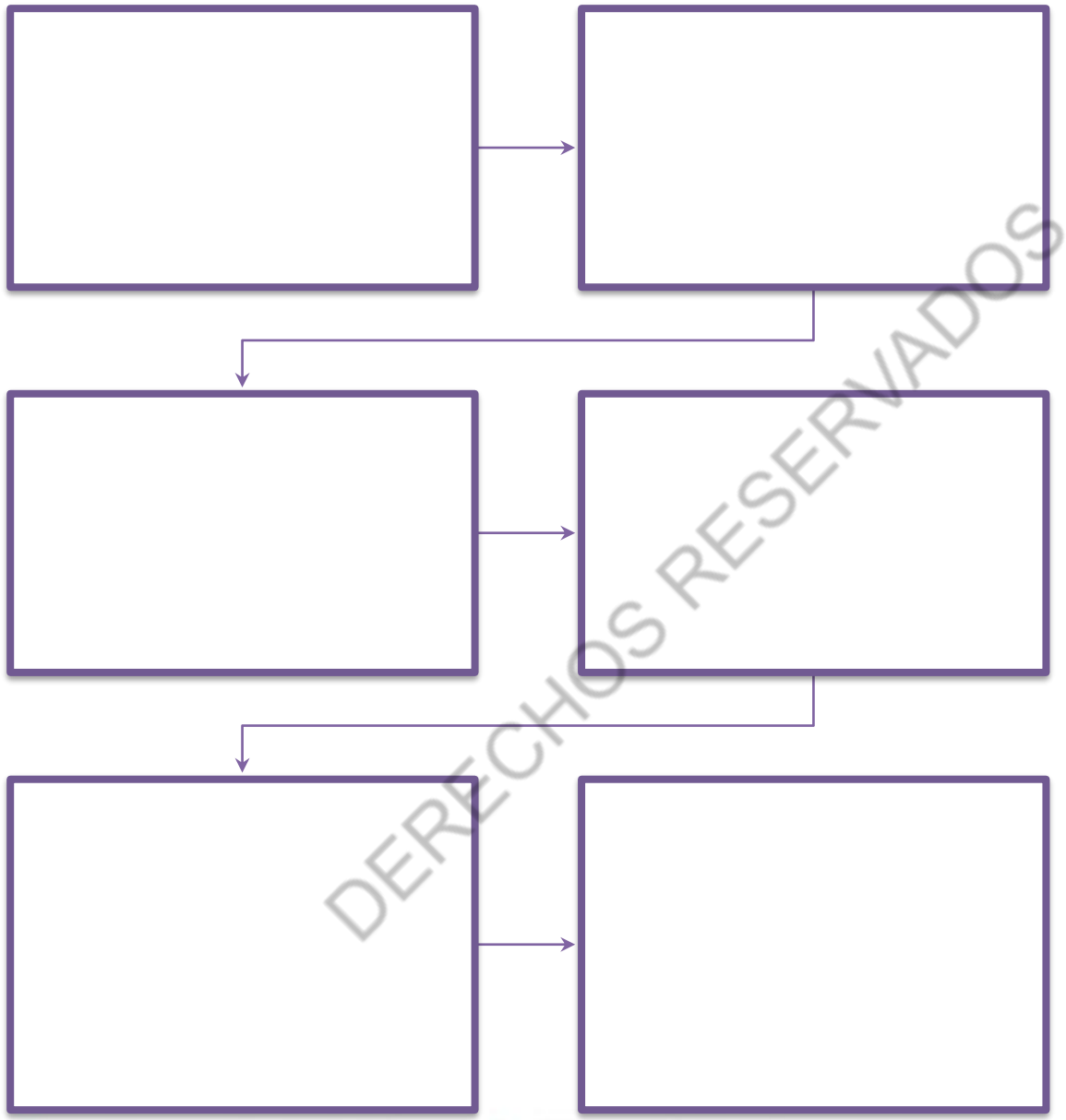
ARCHIVO GENERAL DE LA NACIÓN COLOMBIA

SITA 2019

Extracción, recuperación y clasificación de información

DERECHOS RESERVADOS

La cultura es de todos Mincultura



DERECHOS RESERVADOS

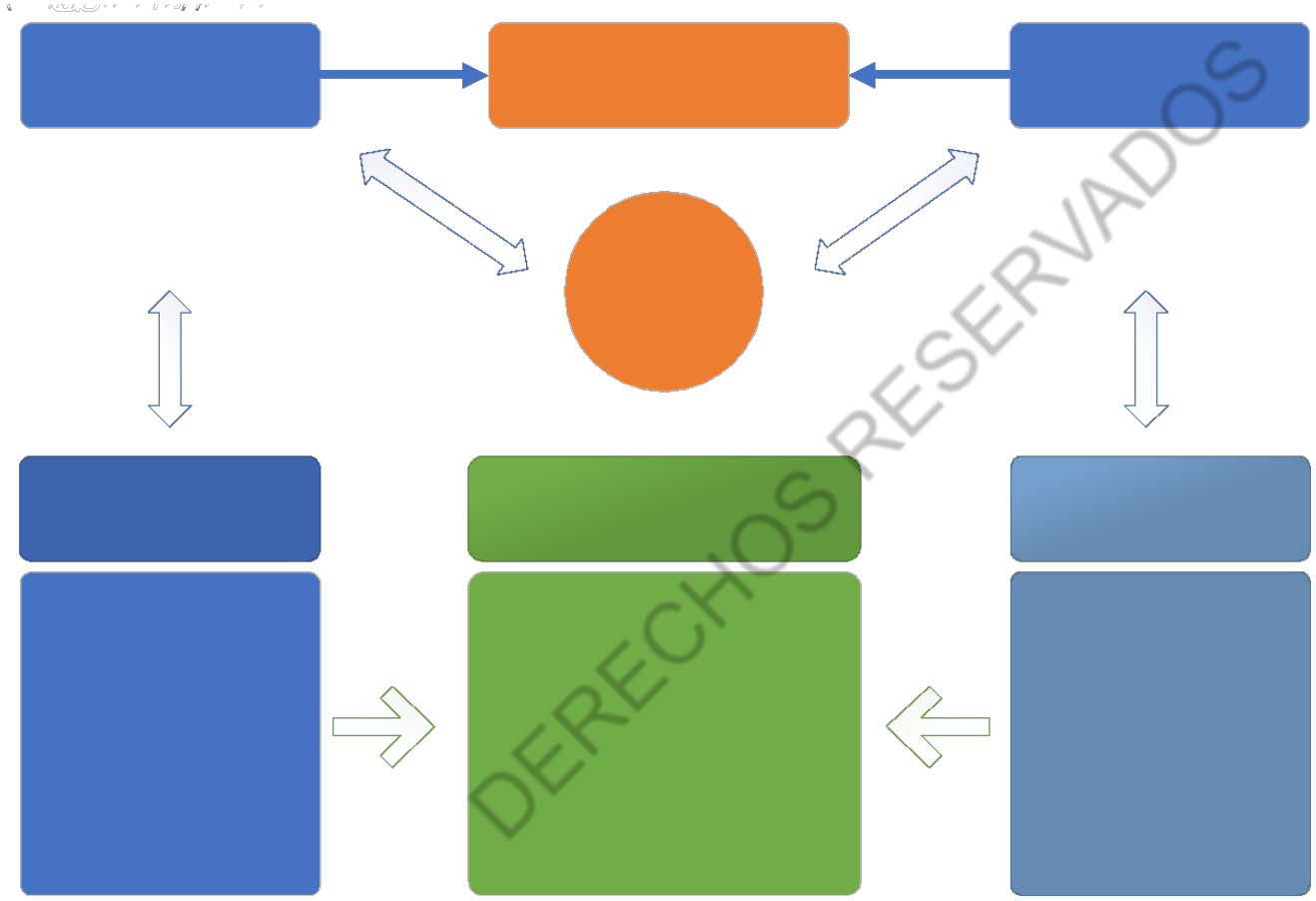
DERECHOS RESERVADOS

“Nos ahogamos de información, pero tenemos hambre de conocimiento”
John Naisbitt

CONTAMINACIÓN INFORMATIVA

La solución: TLH (Tecnologías del Lenguaje Humano)

DERECHOS RESERVADOS

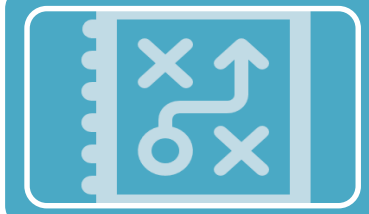


DERECHOS RESERVADOS

Extracción de información

Construir sistemas que encuentren y relacionen información relevante mientras ignoran otras informaciones NO relevantes

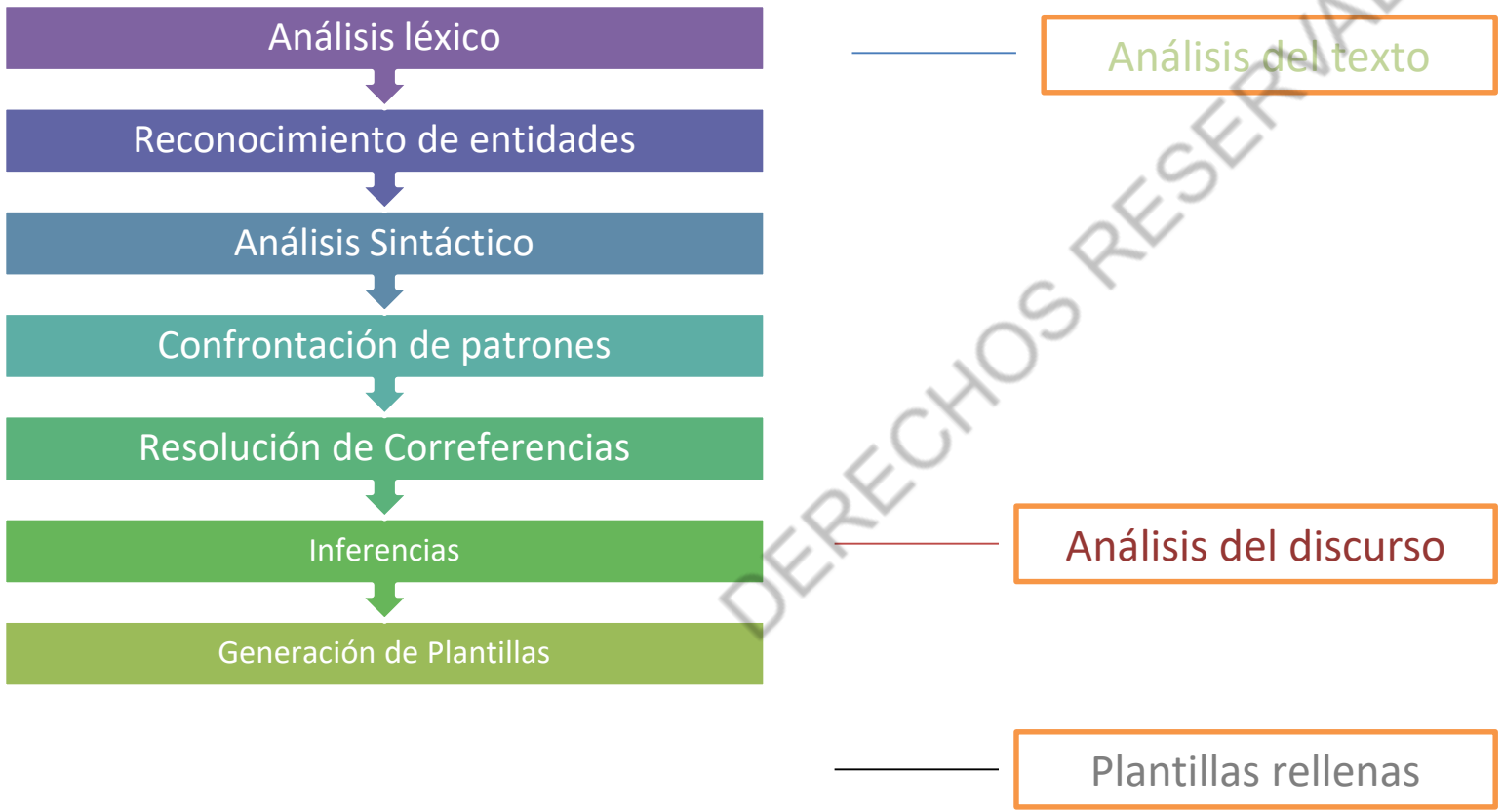
La relevancia se determina a partir de guías predefinidas de dominio, las cuales deben especificar con la mayor exactitud posible el tipo de información a extraer



Desde la perspectiva del PLN, los sistemas de EI deben trabajar a distintos niveles: desde el reconocimiento de palabras hasta el análisis de frases y desde el entendimiento a nivel de frase hasta el texto completo

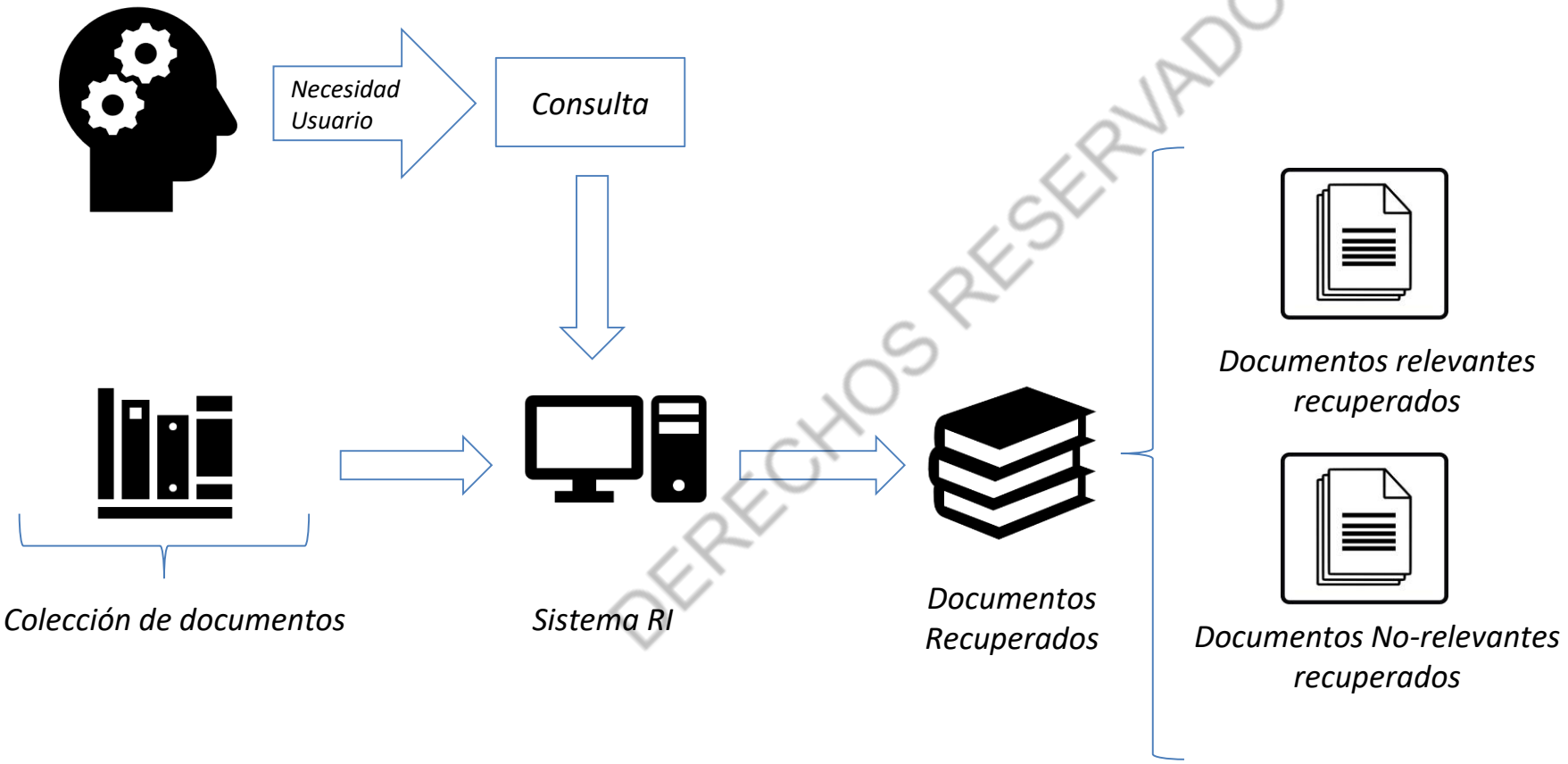
DERECHOS RESERVADOS

Extracción de información



DERECHOS RESERVADOS

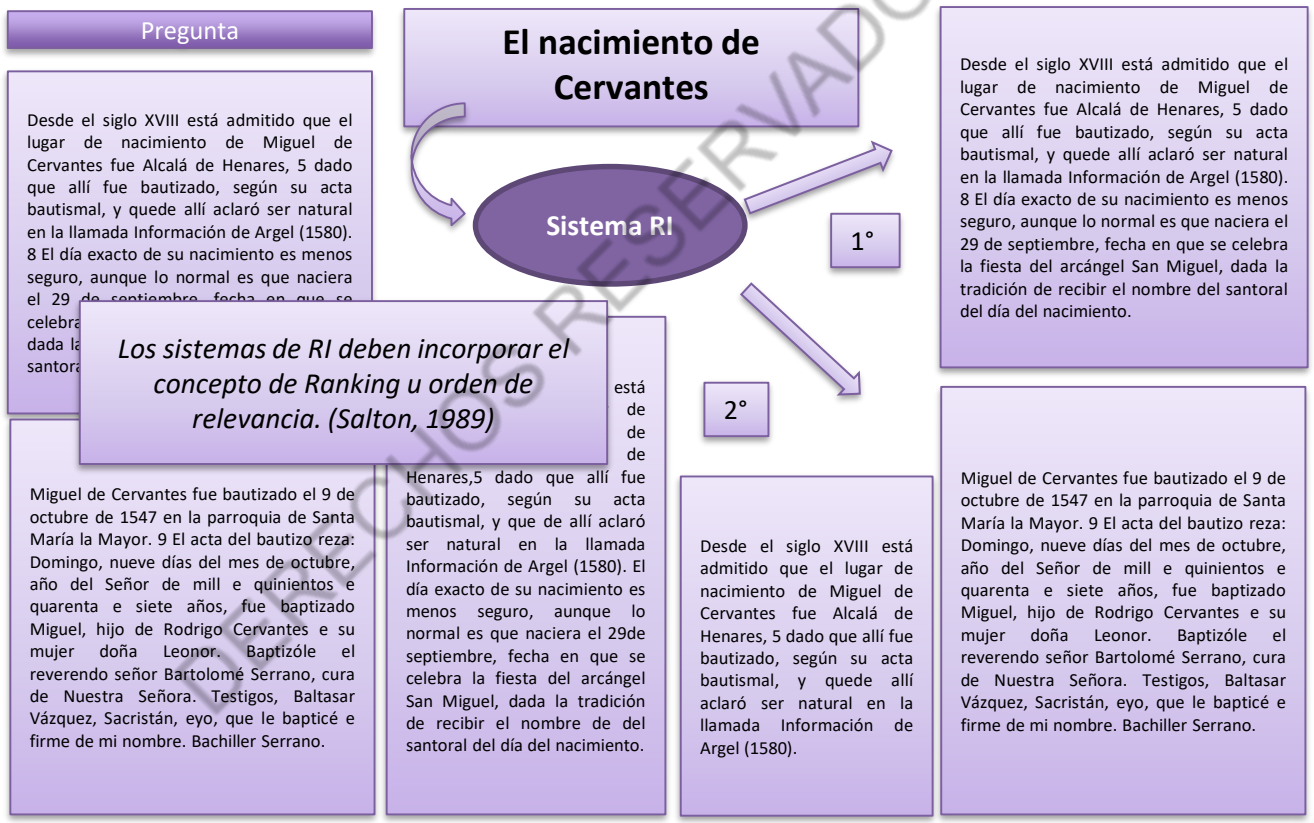
Recuperación de información



DERECHOS RESERVADOS

Dada una pregunta y una colección de documentos

Objetivo: Determinar qué documentos son relevantes.



DERECHOS RESERVADOS

Preguntas ??

DERECHOS RESERVADOS

SITA 2019 octubre 16-18

Semana de Innovación Tecnológica en Archivo

GRACIAS!

gabriel.moreno@lumon.com.co
morenoluis@javeriana.edu.co

DERECHOS RESERVADOS