

Máster en Ciencia de Datos

Recuperación de información

Pablo Castells y Alejandro Bellogín

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Curso 2023 – 2024

Presentación

- ◆ Introducción a la materia
- ◆ Temario
- ◆ Bibliografía
- ◆ Actividades, evaluación, puesta en marcha

Quiénes somos



Pablo Castells

<https://castells.github.io>



Alejandro Bellogín

<https://abellogin.github.io>

- Grupo de recuperación de información, desde ~2005
- Motores de búsqueda, sistemas de recomendación
- Departamento de Ing. Informática & Amazon scholar

¿De qué va la asignatura?

1. Motores de búsqueda (orientada a texto)
2. Sistemas de recomendación

Recuperación de información en el máster en Ciencia de Datos

Aspectos únicos de la IR como problema de aprendizaje automático

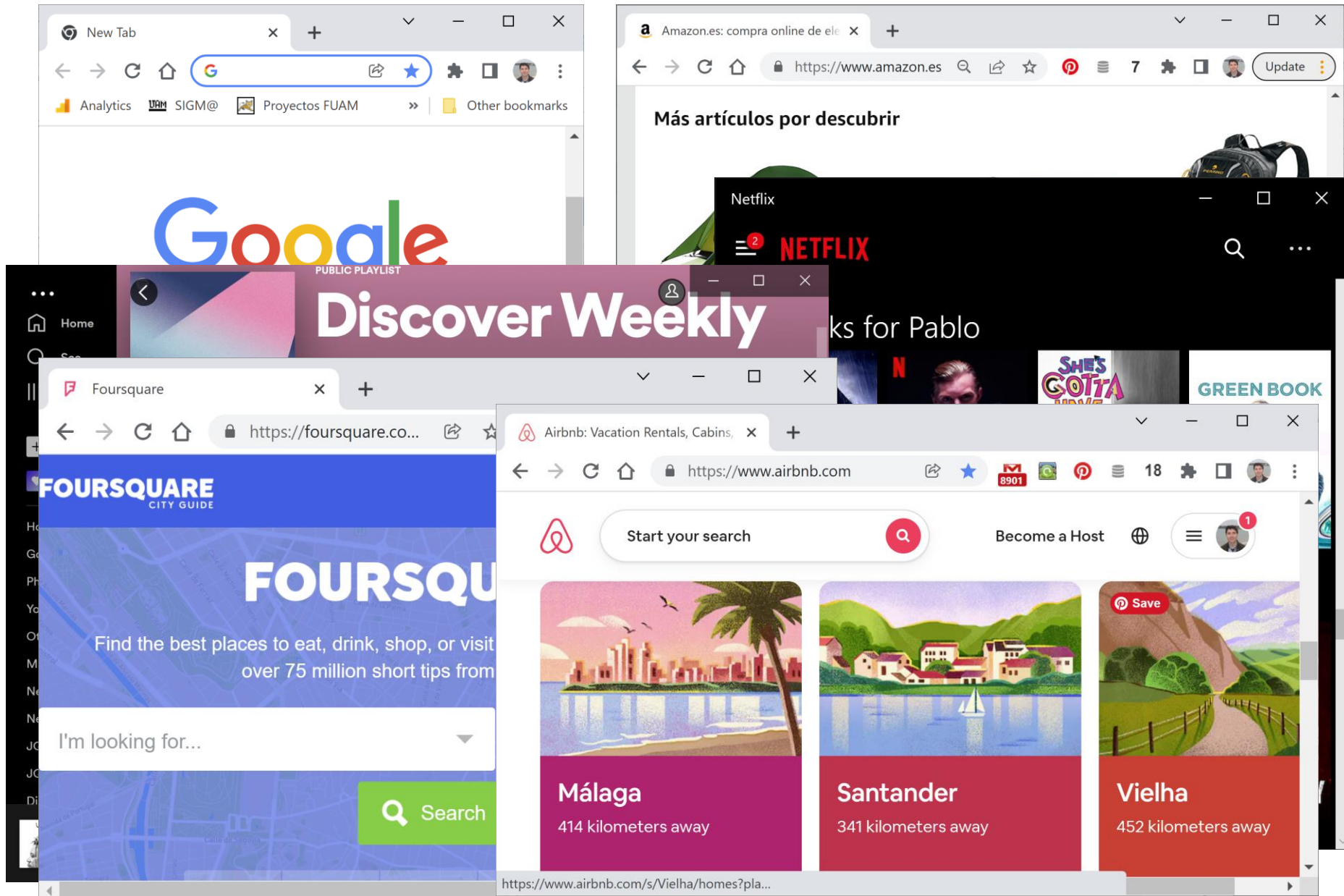
- ◆ Factor **humano**
 - Personas deciden lo que es acertado o no
 - Parte de la información de entrada son acciones humanas espontáneas
 - Subjetividad, inestabilidad, inconsistencia, **incertidumbre**
- ◆ Información **no estructurada**
 - Aspecto común con NLP y procesamiento de señal
 - Escala **masiva**, retos de escalabilidad
- ◆ La tarea de **ranking**
 - No es exactamente regresión ni clasificación
 - Optimización de una ordenación

Problemas típicos de recuperación de información

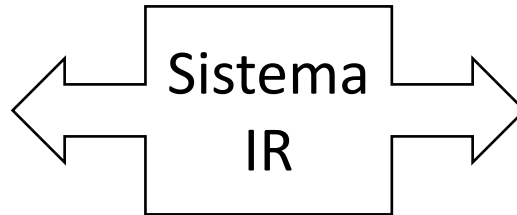
Fundamentos para dar soporte a preguntas o tareas tales como:

- ♦ Cómo funcionan las criptomonedas
- ♦ Por qué se extinguieron los Neandertal
- ♦ Cómo preparar una entrevista de trabajo
- ♦ Dónde cenar este sábado
- ♦ Organizar un viaje
- ♦ Comprar unos auriculares
- ♦ Ver una buena serie
- ♦ Descubrir música
- ♦ Ver vídeos de gatos
- ♦ Buscar pareja

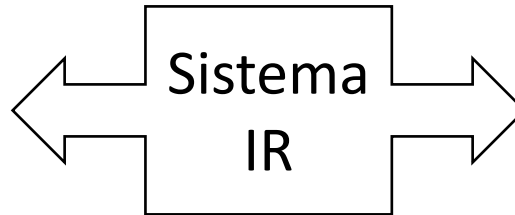
Recuperación de información



Recuperación de información



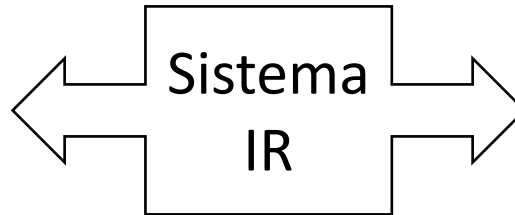
Recuperación de información



- Páginas Web
- Compras
- Vídeo, películas, música
- Personas
- Lugares
- Intranets
- Email, escritorio
- ...

- **No estructurada**: texto, imagen, audio, vídeo, “objetos”
- “Documentos”, “ítems”
- **Escala masiva**

Recuperación de información

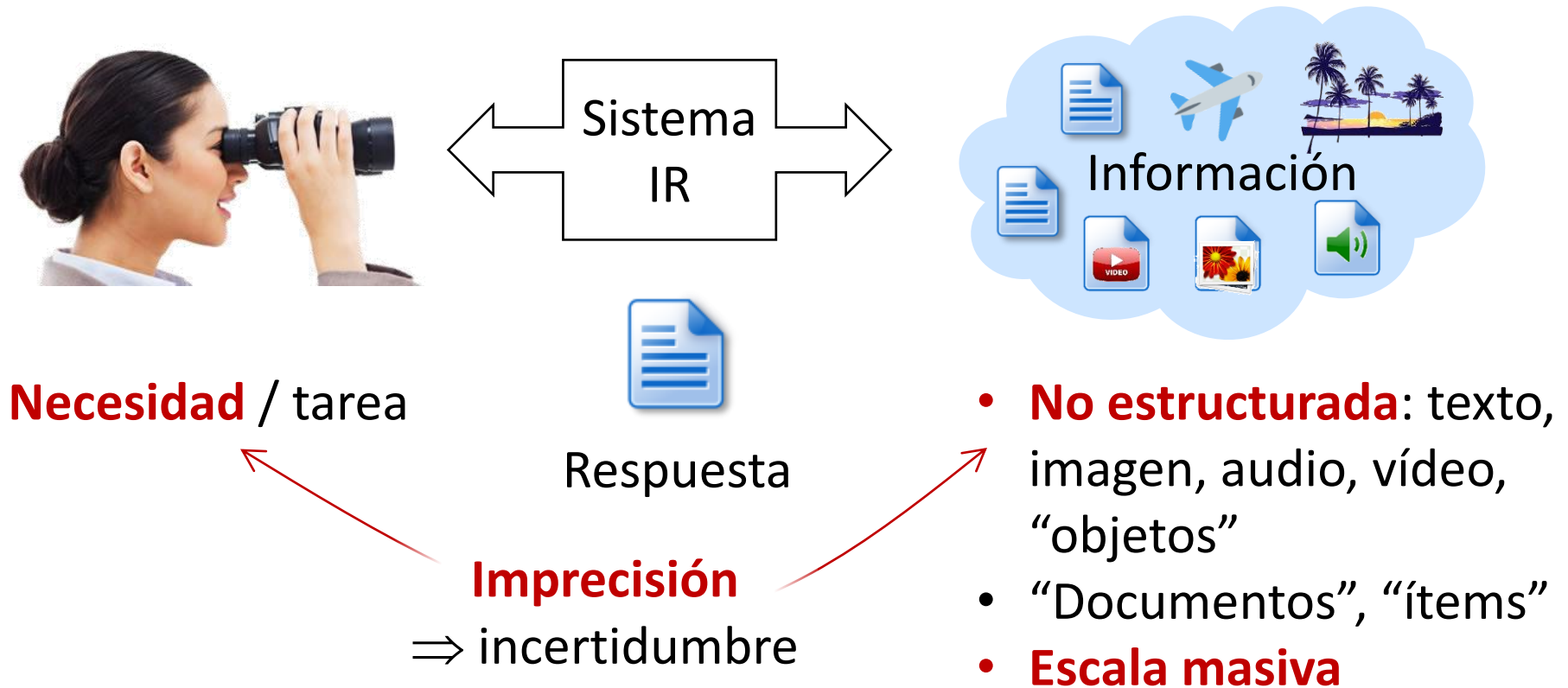


Necesidad / tarea

Imprecisión
⇒ incertidumbre

- **No estructurada**: texto, imagen, audio, vídeo, “objetos”
- “Documentos”, “ítems”
- **Escala masiva**

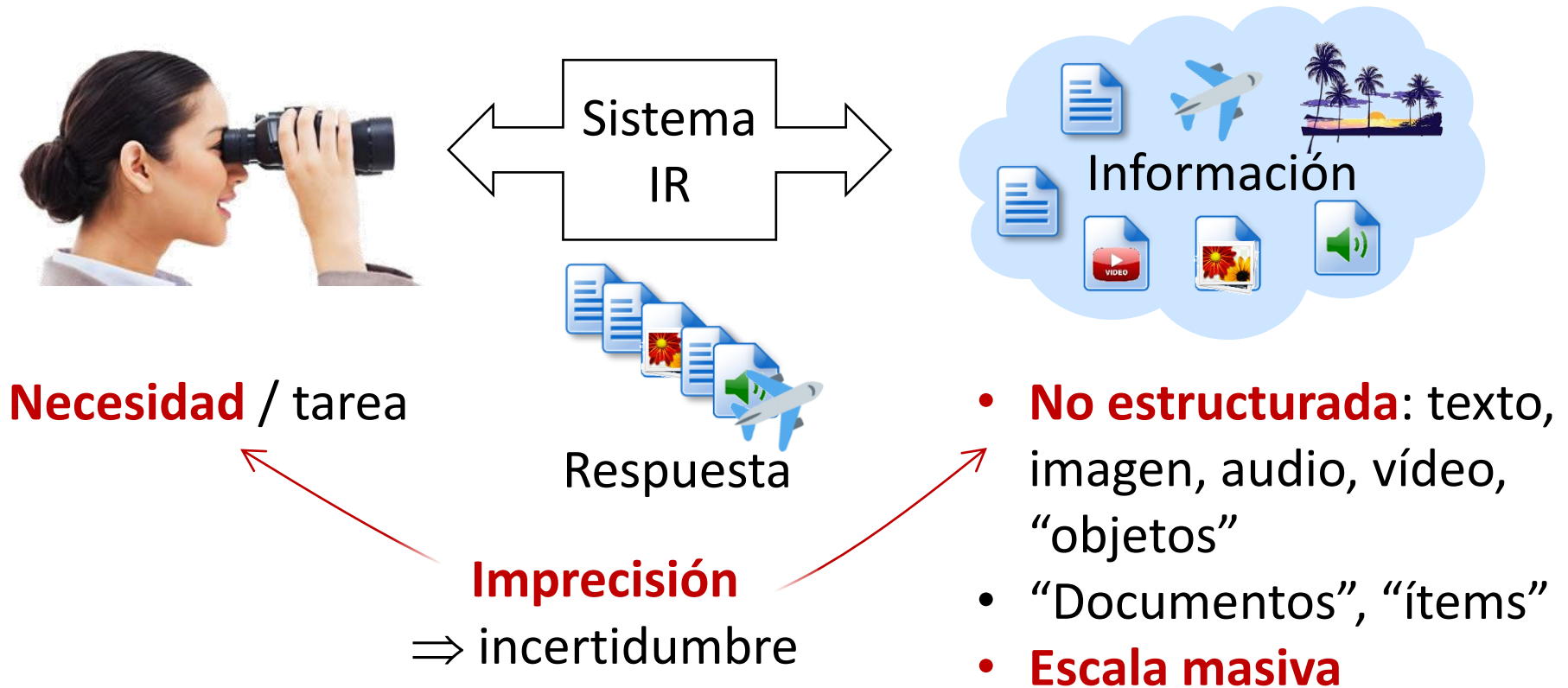
Recuperación de información



¿Cómo lo resolvemos?

- Un problema de clasificación (acierto/fallo)

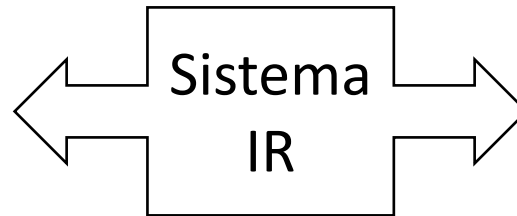
Recuperación de información



¿Cómo lo resolvemos?

- Un problema de clasificación (acierto/fallo)
- O “aprender un **ranking**”
- ¿O clasificación de pares de ítems?

Recuperación de información



Respuesta

Necesidad / tarea

Factor humano

- Una persona decide lo que es útil o no

Imprecisión

⇒ incertidumbre

- **No estructurada**: texto, imagen, audio, vídeo, “objetos”
- “Documentos”, “ítems”
- **Escala masiva**

¿Cómo lo resolvemos?

- Un problema de clasificación (acierto/fallo)
- O “aprender un **ranking**”
- ¿O clasificación de pares de ítems?

En resumen...

La recuperación de información trata problemas de:









- ♦ Acceso a información **masiva** (p.e. Web)
- ♦ **No estructurada** (p.e. texto)
- ♦ Con expresión **imprecisa** (p.e. consulta) del resultado deseado

Se aborda como un problema de aprendizaje automático

- ♦ Supervisado o no supervisado
- ♦ Típicamente el objetivo es generar un **ranking**
- ♦ Etiquetas o *ground truth* (resultado correcto) definidas subjetivamente por **personas**

Por todo ello, se caracteriza por altos grados de **incertidumbre**

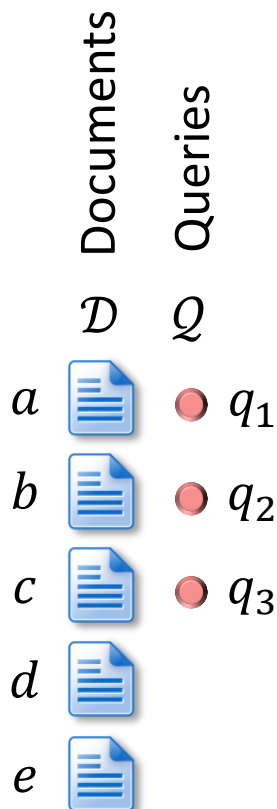
Aprendiendo a recuperar información (búsqueda)

	Documents	Queries
	\mathcal{D}	\mathcal{Q}
a		 q_1
b		 q_2
c		 q_3
d		
e		



Aprendiendo a recuperar información (búsqueda)

How to
find a good f ?



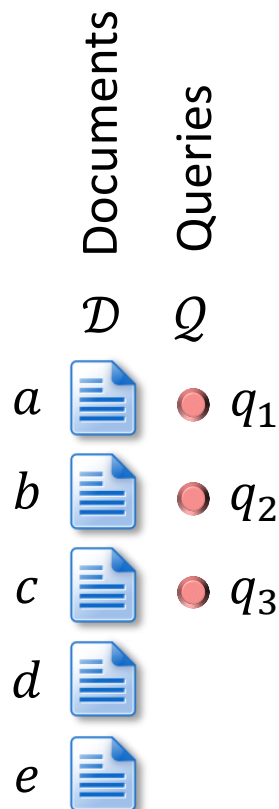
$$f: \mathcal{D} \times \mathcal{Q} \rightarrow \mathbb{R}$$

$$f(\text{document icon}, \text{query icon}) = s$$



Given query q , sort $d \in \mathcal{D}$ by
decreasing score $s = f(d, q)$

Aprendiendo a recuperar información (búsqueda)



$$f: \mathcal{D} \times \mathcal{Q} \rightarrow \mathbb{R}$$

$$f(\text{document icon}, \text{red dot}) = s$$



How to
find a good f ?

1. Unsupervised IR (1950's – 2000)

- Pull it out of our sleeve: intuition + trial and error
- Good principles + trial and error

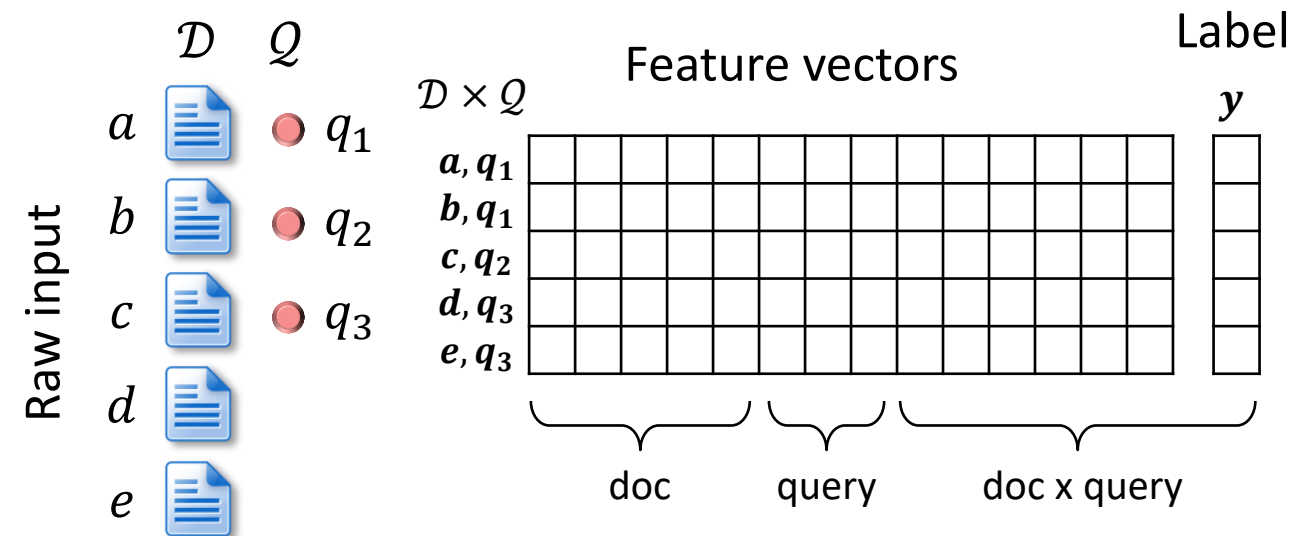


2. Supervised IR (Since 2000)

Given query q , sort $d \in \mathcal{D}$ by decreasing score $s = f(d, q)$

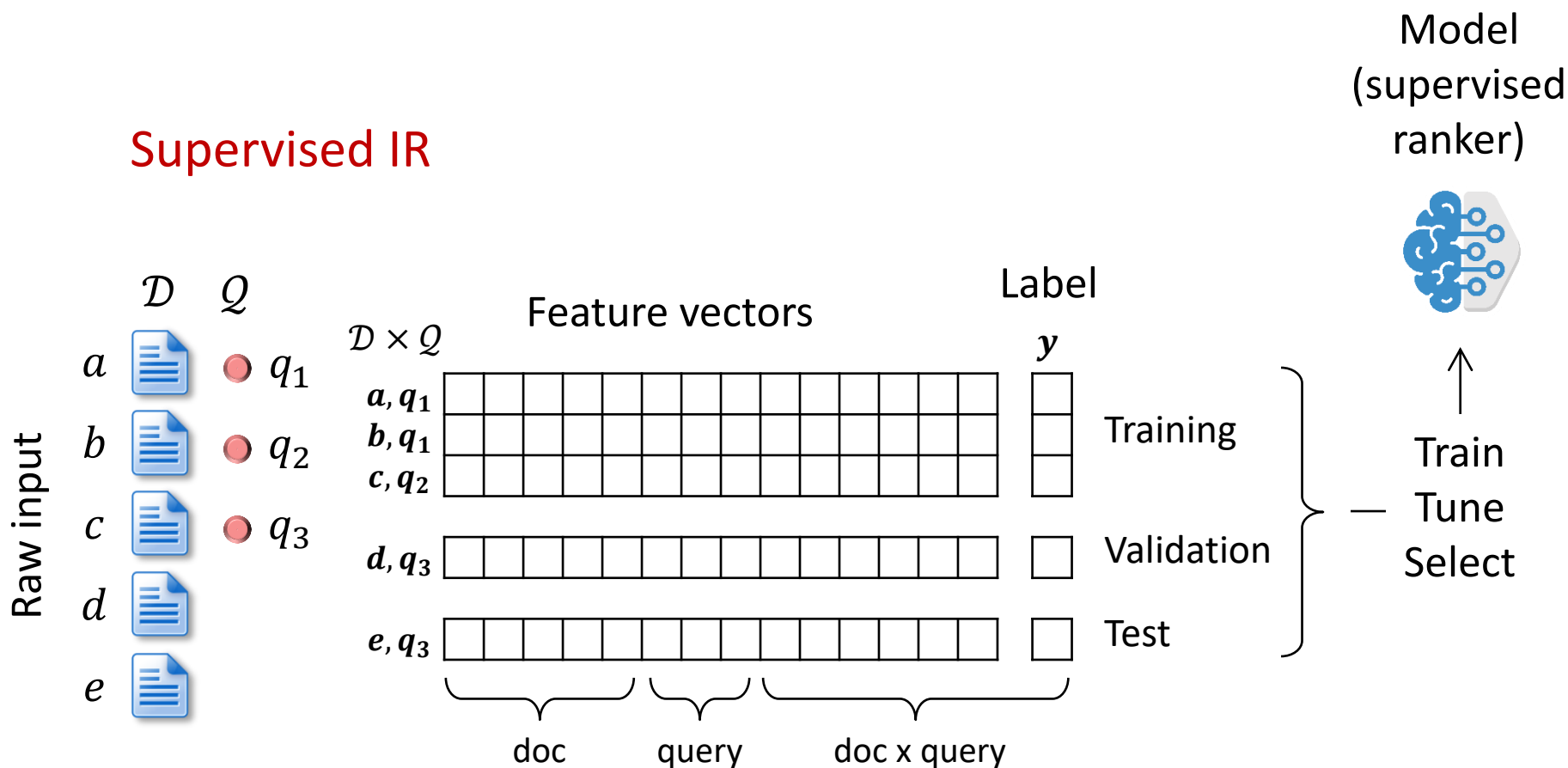
Aprendiendo a recuperar información (búsqueda)

Supervised IR



Aprendiendo a recuperar información (búsqueda)

Supervised IR



Aprendiendo a recuperar información (búsqueda)

Supervised IR

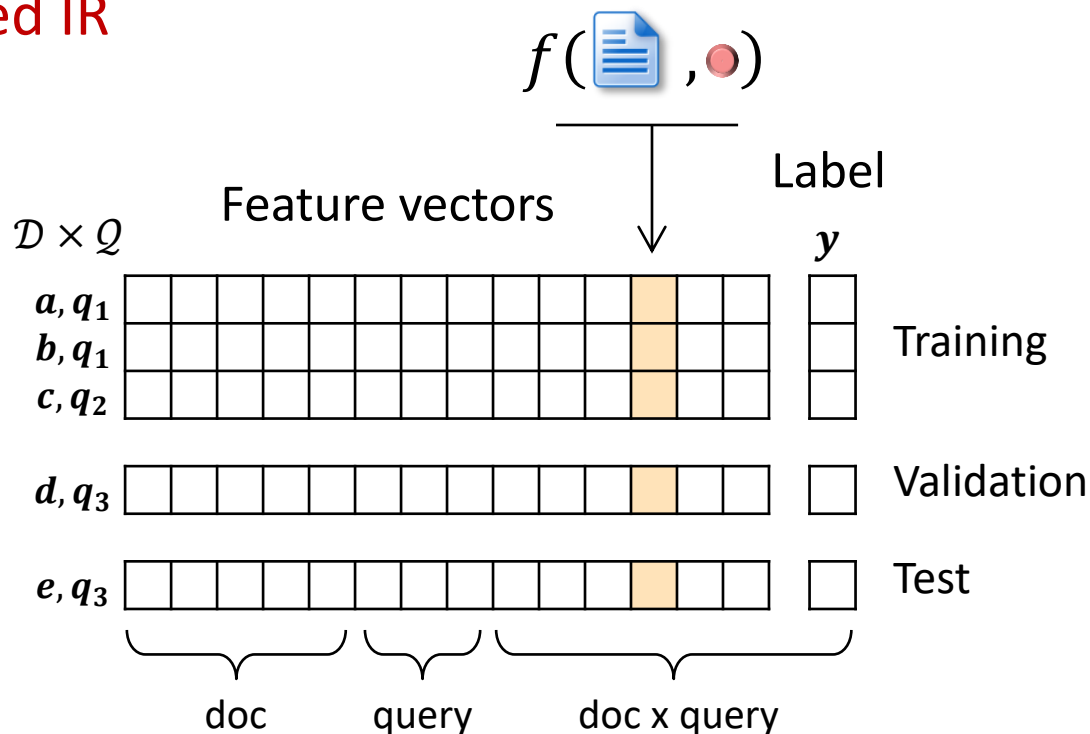
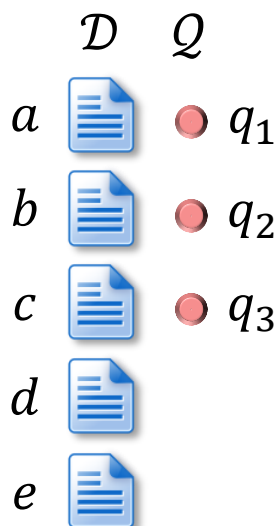
Unsupervised IR

Model
(supervised
ranker)











Train
Tune
Select

Raw input



Raw input

	\mathcal{D}	\mathcal{Q}
a		 q_1
b		 q_2
c		 q_3
d		
e		

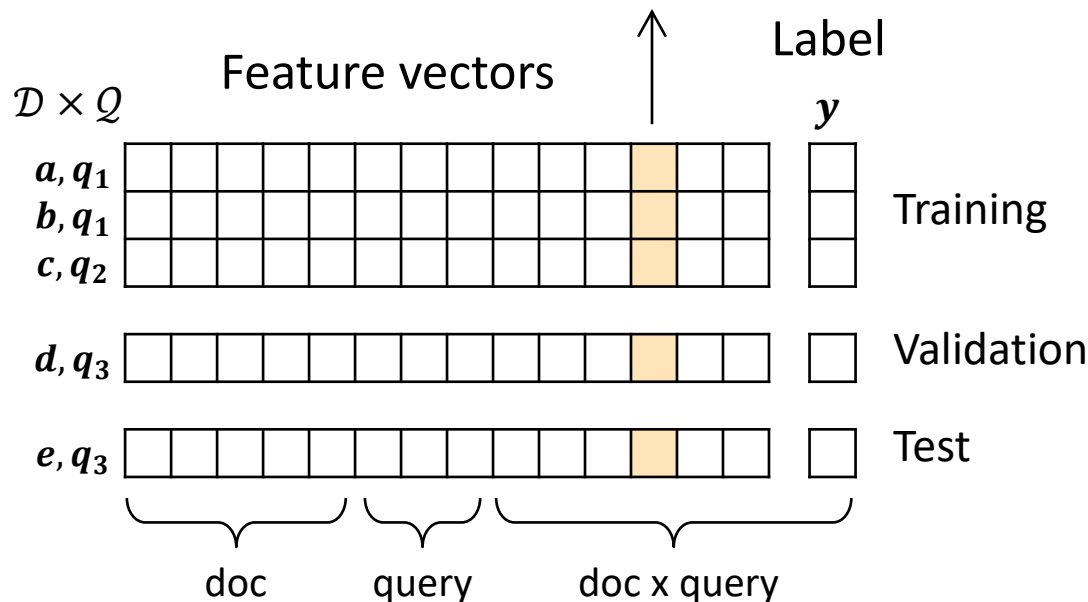
$$f(\text{document icon}, \text{red dot})$$

(top 1,000s
or so)

Supervised
(re)ranking
model
(2nd stage)



Train
Tune
Select



Aprendiendo a recuperar información

- ♦ ¿Cómo se extraen características?
 - De documento
 - De consulta
 - De documento \times consulta
- ♦ ¿De dónde salen las etiquetas?
- ♦ ¿Cómo se definen los modelos?
 - Espacios de hipótesis
 - Funciones de predicción (a.k.a. scoring, ranking), pérdida, coste
 - Entrenamiento
- ♦ ¿Qué métricas offline?
 - Ajuste de parámetros, selección de modelos
- ♦ ¿Cómo sabemos si el sistema funciona?
- ♦ ¿Y si no hay consulta?

Búsqueda sin consulta: recomendación

1. Observamos las acciones de los usuarios
2. Predecimos qué les puede interesar
3. Ofrecemos sugerencias a los usuarios

Recomendación como machine learning

◆ Aprendizaje automático supervisado

Input (features)

- Interacción usuarios – ítems
- Atributos de ítems y de usuarios

Output (labels)

- Interacción usuarios – ítems



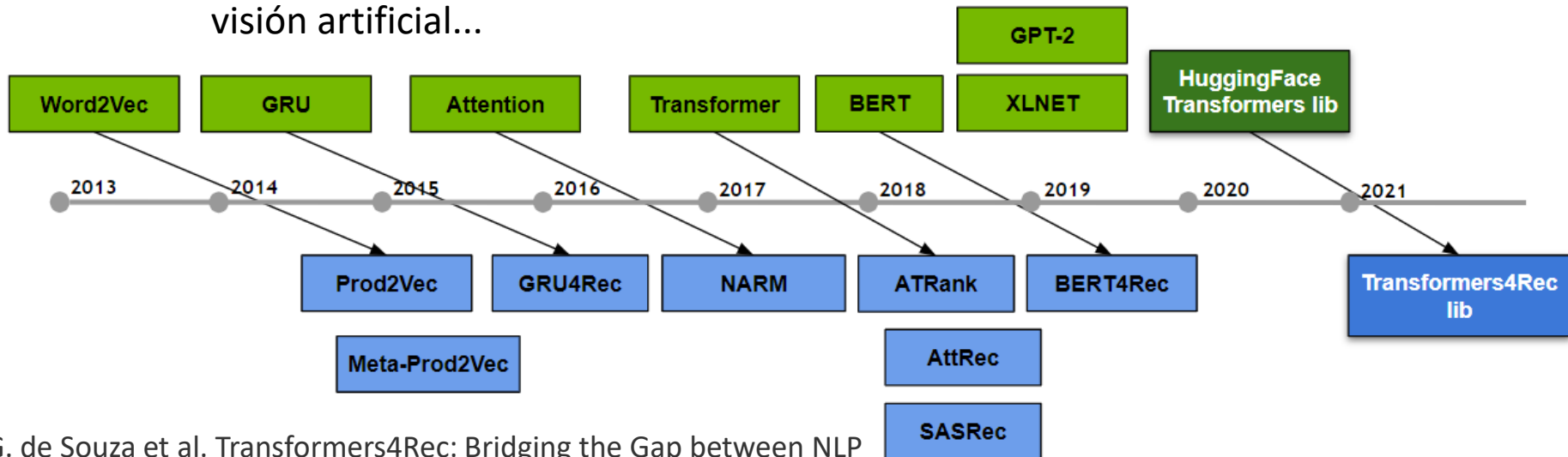
		?	?	
	?	?	?	
		?		?
	?			?

◆ Nuevas perspectivas: aprendizaje de secuencias



Recomendación: algoritmos

- ◆ Si te gustó *Narcos* te gustará *Drug Lords*
 - Recomendación basada en contenido
- ◆ Usuarios que vieron *Narcos* vieron *Kaleidoscope*
 - Vecinos próximos
- ◆ Espacio de factores latentes
 - Factorización de matriz de interacción usuarios / ítems
- ◆ Deep learning models
 - Convergencia de dominios: NLP, visión artificial...

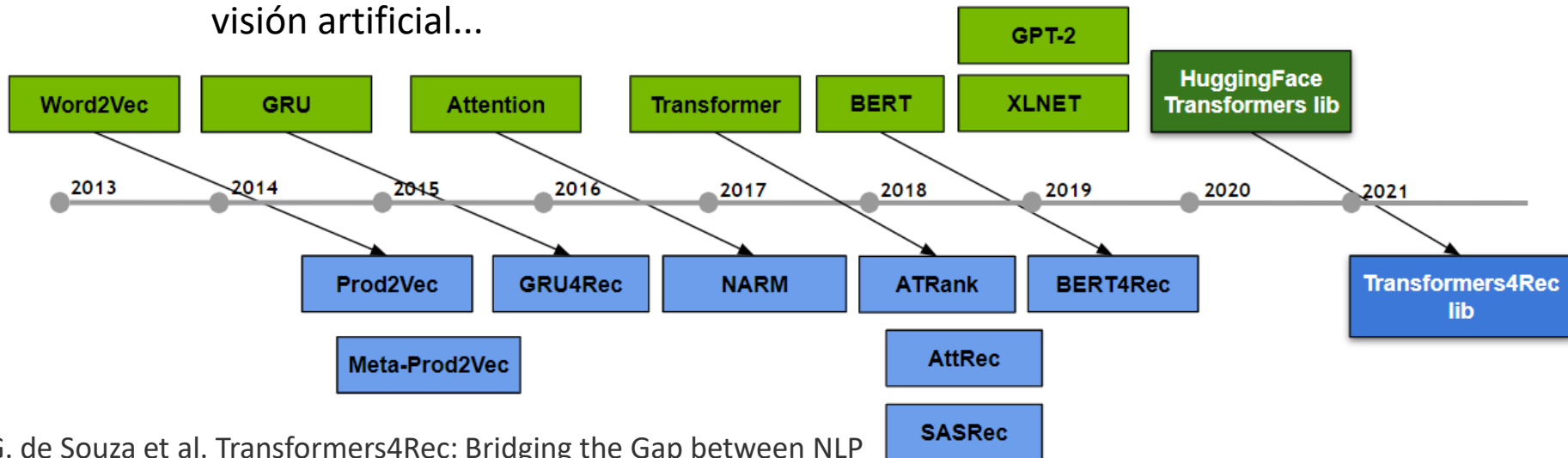


Aprendiendo a recuperar información (búsqueda)

No supervisado	50's - 80's	<ul style="list-style-type: none">• Búsqueda en bibliotecas• Teoría, evaluación, métricas• Bag of words, modelo vectorial, relevance feedback• Modelos probabilísticos• Stemming, tratamiento morfológico
	90's	<ul style="list-style-type: none">• Datasets (TREC)• BM25• La Web!! Arranca la explosión industrial IR• Google, PageRank, sistemas de recomendación (Amazon)
		<ul style="list-style-type: none">• Modelos de lenguaje
Supervisado	00's	<ul style="list-style-type: none">• Aprendizaje automático (investigación)• Evaluación online: A/B testing, interleaving
	10's	<ul style="list-style-type: none">• Aprendizaje automático (comercial)• Tratamiento NLP• Deep learning

Recomendación: algoritmos

- ◆ Si te gustó *Narcos* te gustará *Drug Lords*
 - Recomendación basada en contenido
- ◆ Usuarios que vieron *Narcos* vieron *Kaleidoscope*
 - Vecinos próximos
- ◆ Espacio de factores latentes
 - Factorización de matriz de interacción usuarios / ítems
- ◆ Deep learning models
 - Convergencia de dominios: NLP, visión artificial...



¿Qué tiene que decir
ChatGPT a todo esto?



IR & LLMs

- ◆ Los LLMs están resolviendo problemas de IR que parecían imposibles
- ◆ Han cambiado completamente la experiencia del usuario



Marca

<https://www.marca.com> › 2024/01/24 · [Translate this page](#)

Carlos Alcaraz se queda a media remontada ante Zverev ...

4 days ago — **Open de Australia 2024**: Carlos Alcaraz se queda a media remontada ante Zverev y dice adiós en cuartos del Open de Australia. Marca.

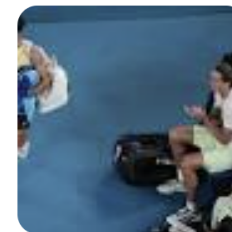


Mundo Deportivo

<https://www.mundodeportivo.com> › ... · [Translate this page](#)

Australian Open 2024 | La cruda declaración de Alcaraz ...

2 days ago — El mundo del tenis quedó shockeado este miércoles por la sorpresiva eliminación de Carlos **Alcaraz** en el **Australian Open**.



Australian Open

<https://ausopen.com> › players › spain › carlos-alcaraz

Carlos Alcaraz [ESP] | AO

Assured **Alcaraz** relishing debut second week at AO Carlos **Alcaraz** is through to the **Australian Open** second week for the first time after teenage wildcard . Jerry



ChatGPT 4

You
alcaraz australia open

ChatGPT
I did a quick search for more information and here's what I found.

Retrieval-Augmented Generation (RAG)

Carlos Alcaraz had a notable journey in the 2024 Australian Open but was ultimately eliminated in the quarter-finals. He comfortably reached the fourth round for the first time in this tournament after his opponent Shang Juncheng retired due to injury, with Alcaraz leading 6-1, 6-1, 1-0 at that point [1]. In the fourth round, he faced Miomir Kecmanovic and secured a decisive victory with scores of 6-4, 6-4, 6-0, marking his first quarter-final berth at the Australian Open [1].

However, Alcaraz's run came to an end in the quarter-finals where he faced Alexander Zverev. Despite a valiant effort, including a comeback in the third set, Alcaraz was defeated by Zverev with scores of 6-1, 6-3, 6-7(2), 6-4. Alcaraz acknowledged his performance in this match, reflecting on the ups and downs he experienced and his inability to maintain a consistent level of play against Zverev [1] [1]. This quarter-final appearance represents his best showing in three appearances at the season's opening Grand Slam, and it's an experience he intends to build tournaments.

English (US)

Send a message or hold space bar to talk

ChatGPT

https://chat.ope...

New Chrome available

ChatGPT 4

You

I enjoyed "The Creator" - what other movies can you recommend to me?

ChatGPT

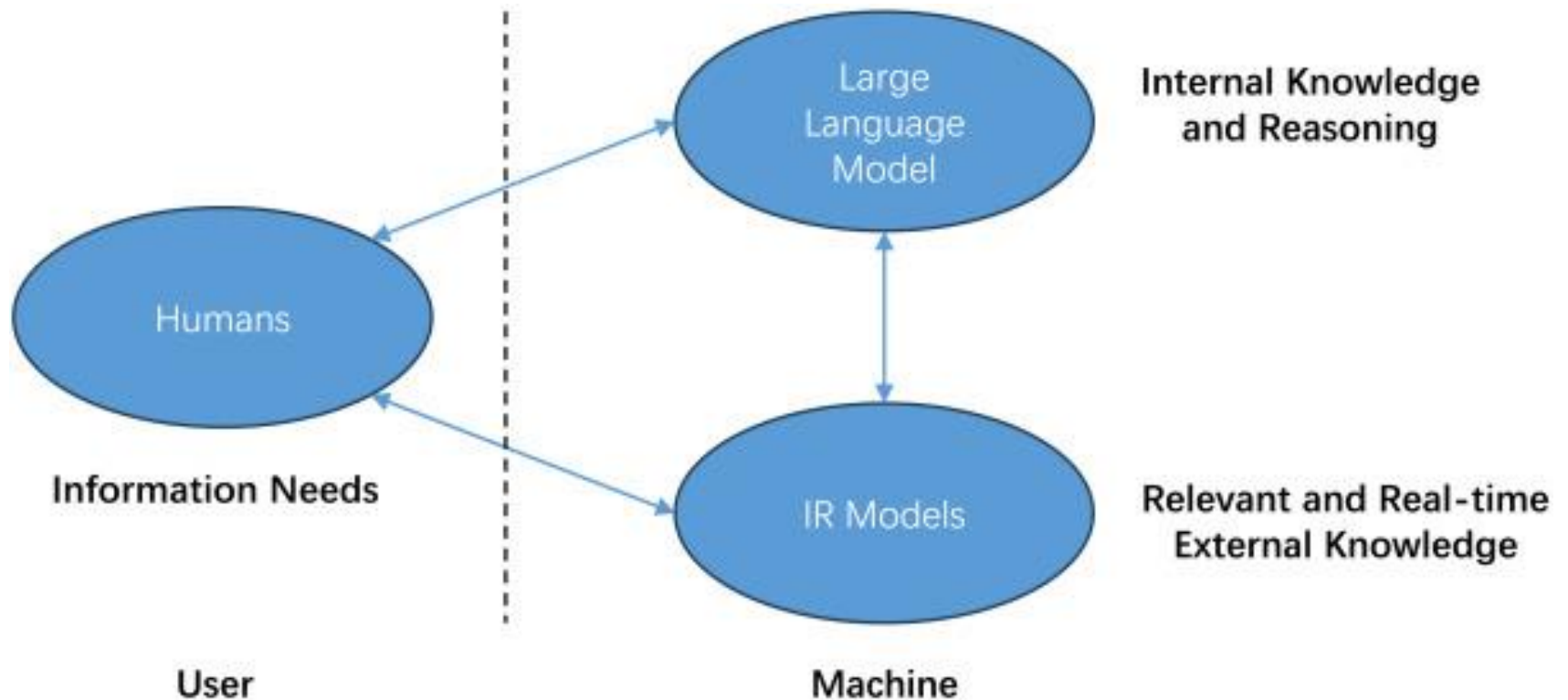
"The Creator" is a unique film with its own distinctive themes and style. To recommend similar movies, I'll consider films that blend elements of science, philosophy, and deep character exploration, often with a focus on creativity or invention. Here are some recommendations:

- "The Imitation Game" (2014):** This historical drama focuses on Alan Turing and his team's efforts to crack the Nazi Enigma code during World War II. It delves into themes of genius, creativity, and the ethical dilemmas of technology and war.
- "A Beautiful Mind" (2001):** A biographical film about John Nash, a brilliant but asocial mathematician, and his journey through the challenges of mental illness. It beautifully portrays the intersection of genius, creativity, and the human mind.
- "Ex Machina" (2014):** A science fiction thriller that explores the themes of artificial intelligence, consciousness, and the moral implications of creating life-like AI. It's a thought-provoking film that delves into the nature of creativity and ethics in technology.
- "The Social Network" (2010):** Though more of a biographical drama, this film about the creation of Facebook explores themes of innovation, entrepreneurship, and the

Send a message or hold space bar to talk

ChatGPT can make mistakes. Consider checking important information.

IR & LLMs



IR & LLMs

IR complementa a un LLM

- ◆ Adquisición de información reciente que el LLM no ha visto
- ◆ Consistencia factual, mitigar alucinación
- ◆ Enlaces a fuentes documentales

LLM complementa IR: generative retrieval, generative recom.

- ◆ Comprensión de consultas, matices semánticos
- ◆ Elaboración del output
- ◆ Interacción y diálogo avanzado
- ◆ Conocimiento del mundo

Temario

1. Motores de búsqueda

- ♦ Búsqueda no supervisada (features)
 - Modelo vectorial
 - Modelos probabilísticos: Naïve Bayes, BM25, modelos de lenguaje
- ♦ Búsqueda supervisada (learning to rank)

2. Evaluación

- ♦ Evaluación offline
 - Datos de prueba
 - Métricas básicas
 - Métricas basadas en modelo
- ♦ Evaluación en producción: A/B testing

3. Búsqueda avanzada

- ♦ Búsqueda Web
 - Crawling e indexación
 - Ránking basado en enlaces: PageRank y HITS
- ♦ Búsqueda personalizada
 - PageRank personalizado
 - Personalización de consultas y ránking
 - Relevance Feedback

Temario (cont)

4. Sistemas de recomendación

- ♦ La tarea de recomendación
- ♦ Métodos basados en vecinos
- ♦ Factorización de matrices
- ♦ Modelos deep learning
- ♦ Evaluación de recomendaciones

5. Temas avanzados

- ♦ Diversidad en búsqueda y recomendación
- ♦ Sesgos en búsqueda y recomendación
- ♦ Recomendación sensible al tiempo

Bibliografía

- ◆ Information Retrieval – Implementing and Evaluating Search Engines
Stefan Büttcher, Charles L. A. Clarke, Gordon V. Cormack, 2010
- ◆ Introduction to Information Retrieval
C. D. Manning, P. Raghavan, H. Schütze, 2008
- ◆ Recommender Systems Handbook, 3rd ed
F. Ricci, L. Rokach, B. Shapira (eds.), 2022
- ◆ Recommender Systems: A primer
P. Castells, D. Jannach, 2023

Evaluación

- ◆ 50% prácticas (3 prácticas)
 - Principalmente Python
 - Algún ejercicio en hoja de cálculo
 - Baremo orientativo sobre 3 puntos
 - 0 – Práctica no entregada
 - 1 – Práctica con defectos, entregada con retraso, etc.
 - 2 – Práctica realizada correctamente sin parte opcional
 - 3 – Práctica realizada correctamente con parte opcional
 - > 3 – Práctica excepcional
- ◆ 50% examen final
 - Con libros y apuntes
 - ¡Traed calculadora!
 - Miércoles 22 de mayo, 16h