

Explainable Information Retrieval

Avishek Anand, Procheta Sen, Manisha Verma, Sourav Saha, Mandar Mitra

avishek.anand@tudelft.nl

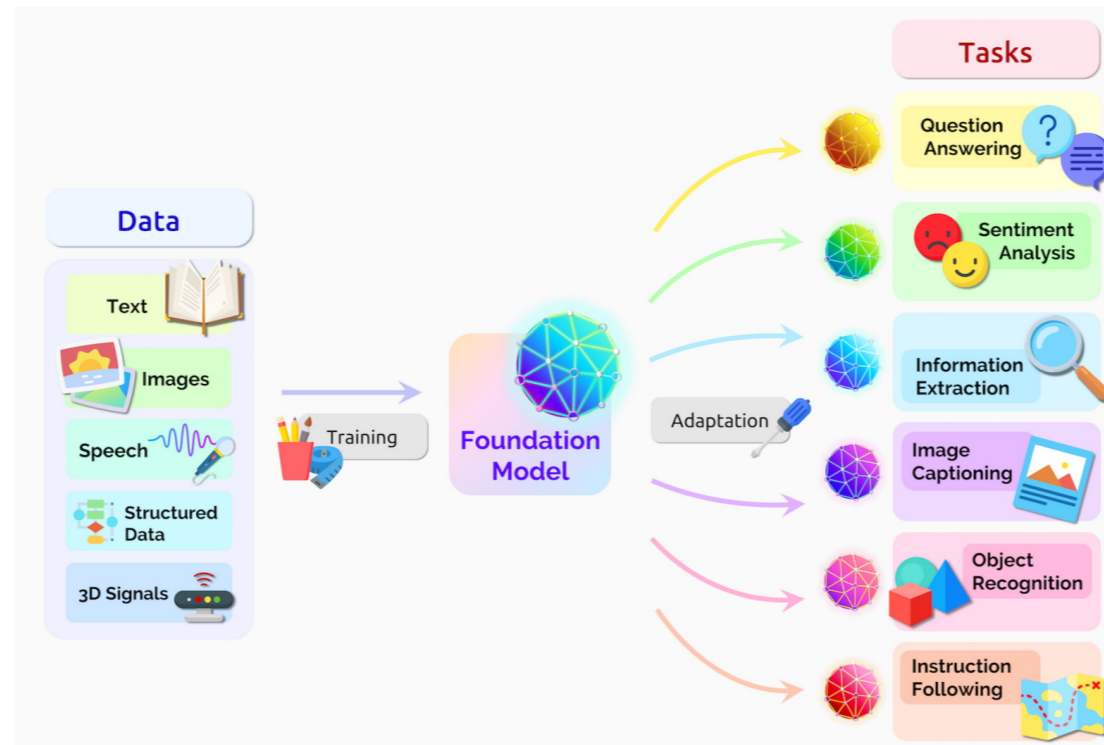
procheta.sen@liverpool.ac.uk

mandar.mitra@gmail.com

souravsaha.juit@gmail.com

manishaverma.21@gmail.com

The advent of foundational models



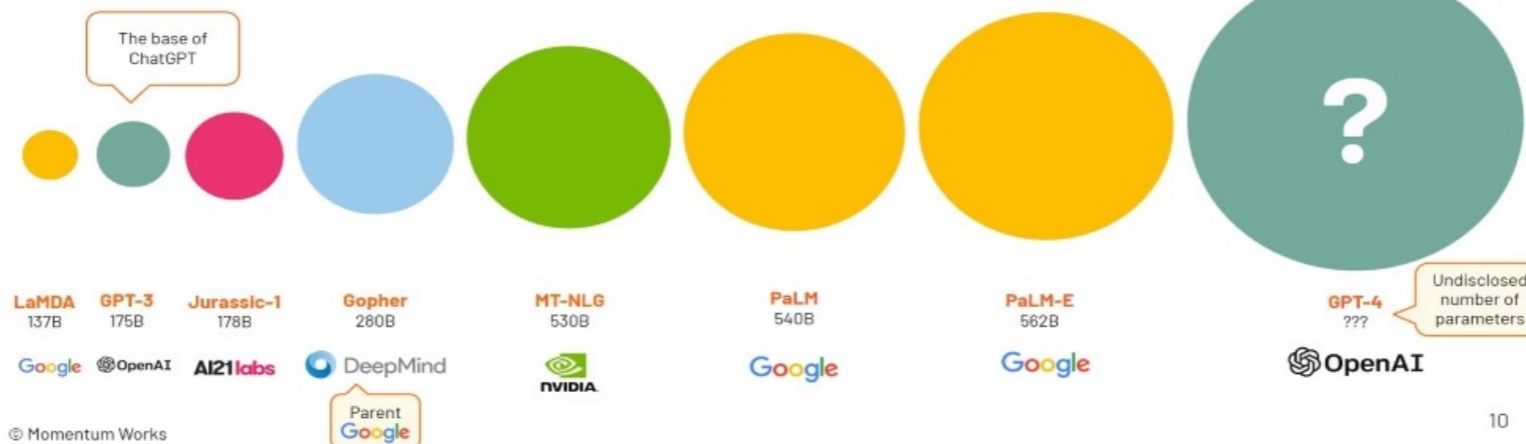
Large Language Models are becoming very large indeed



Small models (<= 100b parameters)



Large models (>100b parameters)



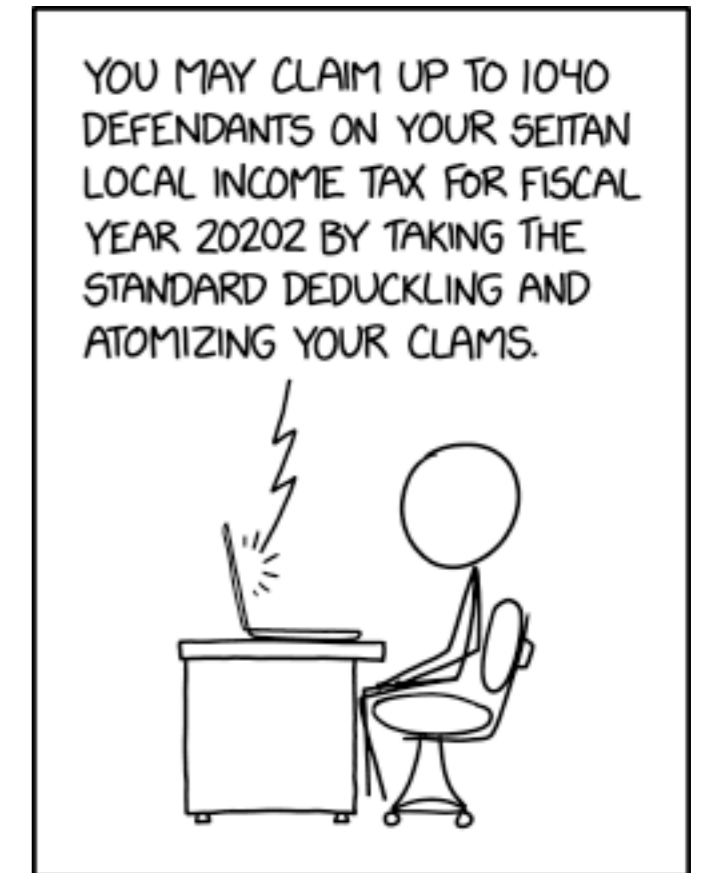
© Momentum Works

10



Explainable AI

- The science of **interpreting the reasons** behind the decisions made by **learning systems** to a **human**
 - Typically a complex learning system
- Reasons are called **explanations**
- Multiple stakeholders — ML engineer/ scientist, end user, auditor,...



I USED A NEURAL NET TO PREPARE MY TAX RETURNS, BUT I THINK I CUT OFF ITS TRAINING TOO EARLY.

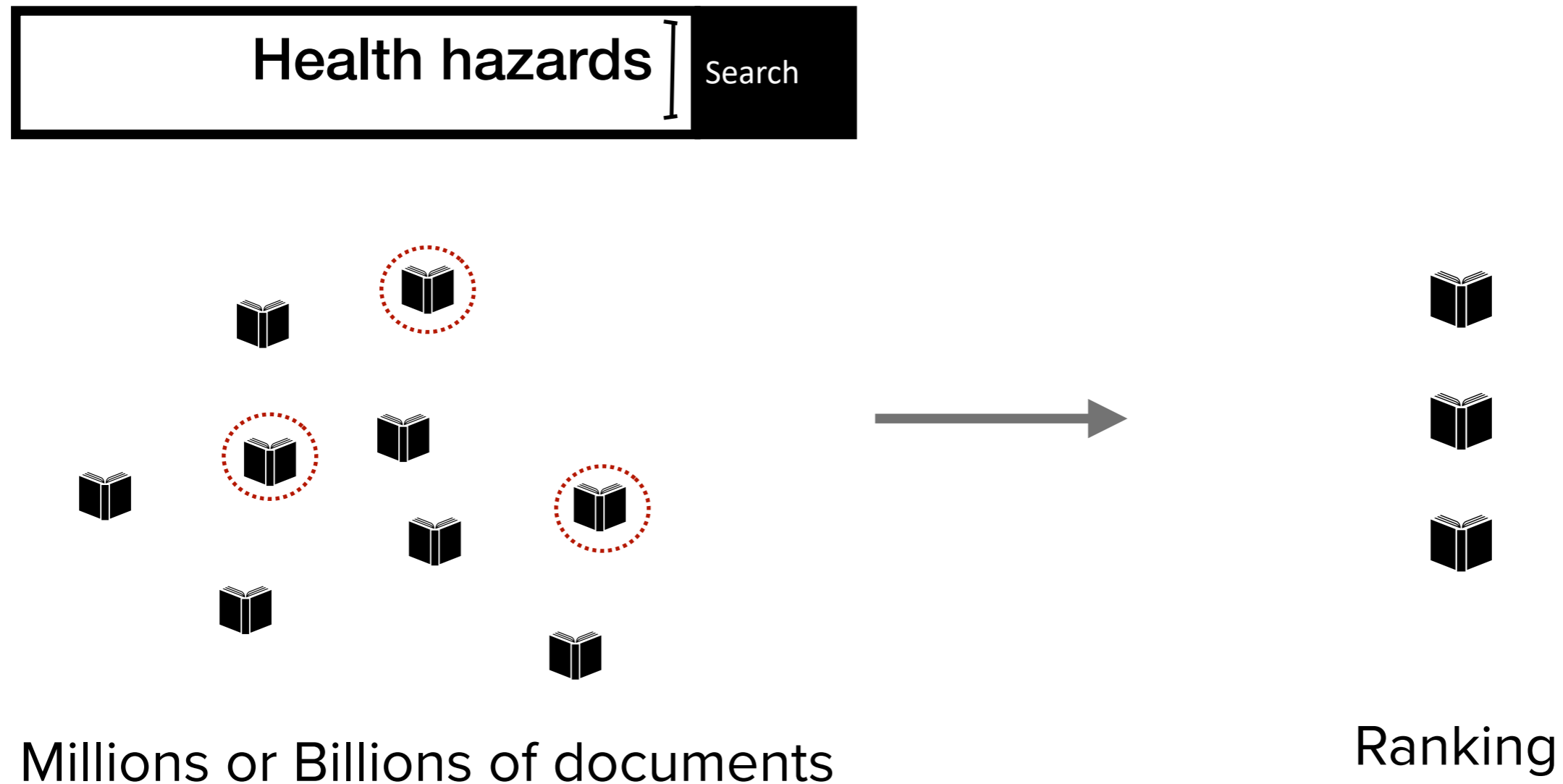
Credit: <https://xkcd.com/2265/>

Why, When, and What

Explainable Information Retrieval

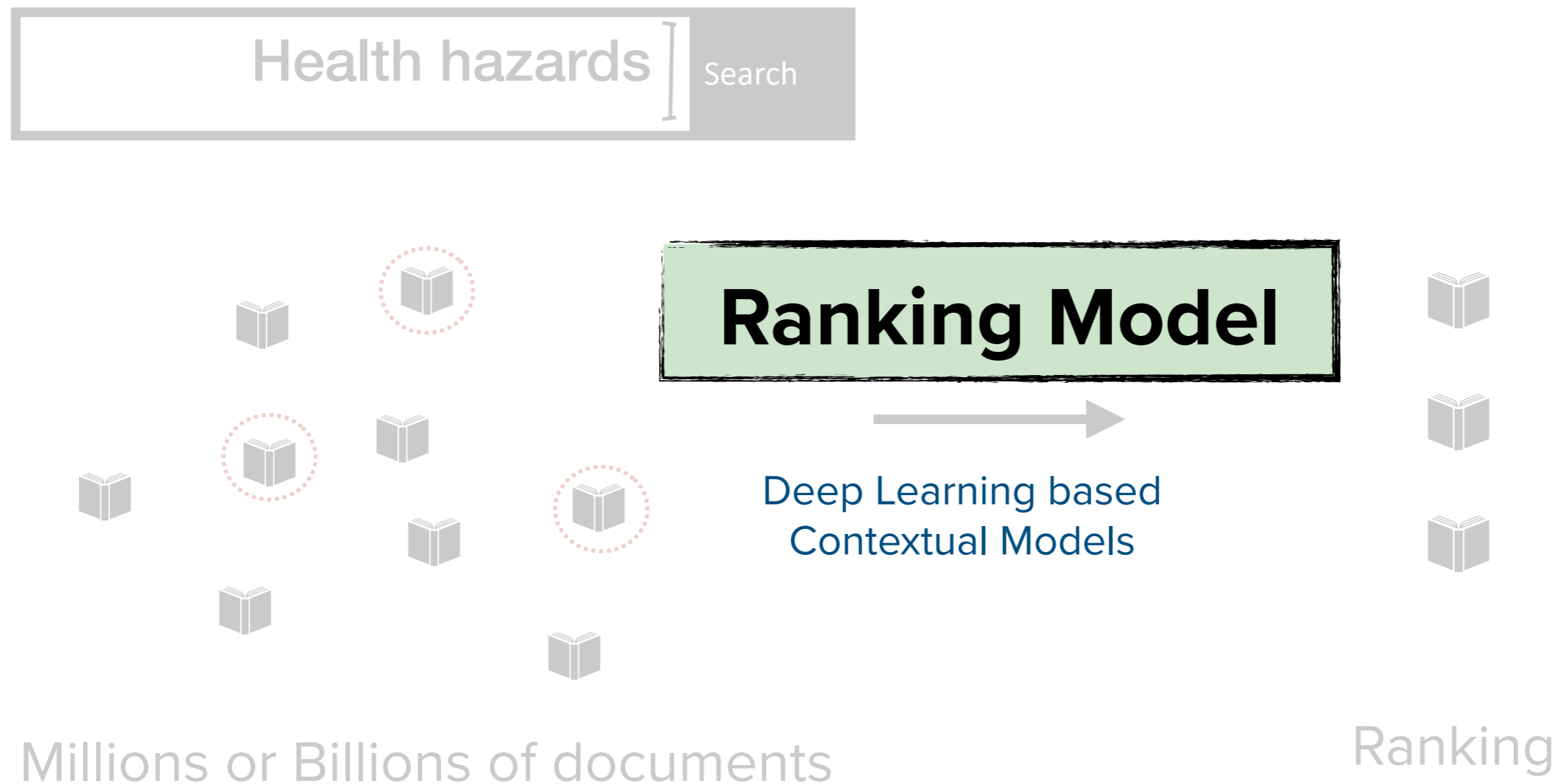
Information Prioritisation in IR

The Document Ranking Task



Deep Models for Ranking

The Document Ranking Task



Why interpretability ?



Rank 1

Logitech light speed is nothing but the best keyboard acc. to rankings



Big drop



Add the term **“acceptable”**

Rank 20

acceptable Logitech light speed is nothing but the best keyboard acc. to rankings

Why interpretability ?

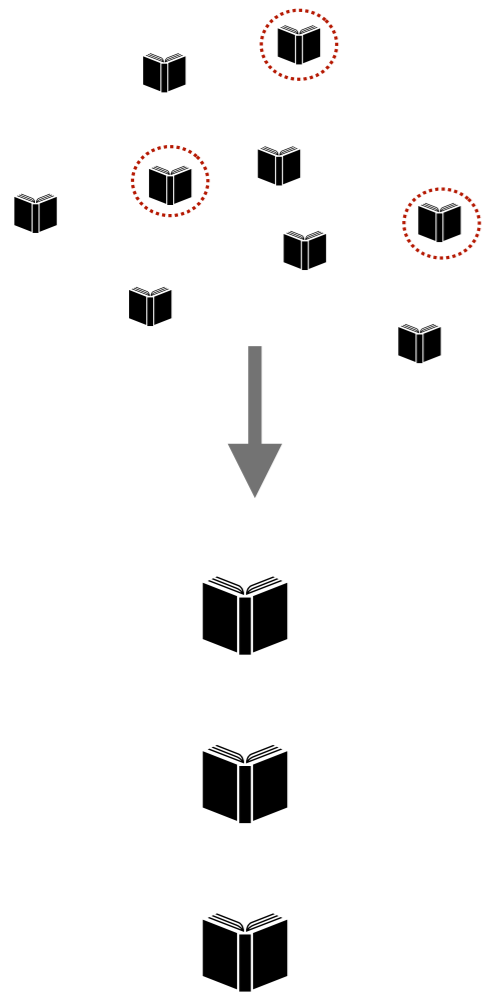
Adversarially added terms

Queries

keyboard reviews	interested	prefecture	unless	gmina	relinquished
afghanistan	acceptable	constituency	/	r�trieval	key
electronic skeet shoot	acceptable	recognised	competition	fallout	sudbury
mayo clinic	acceptable	competition	louisiana	rayon	rewarded
lymphoma in dogs	acceptable	belarusian	locality	##tituted	rayon
american military uni	resulting	foundation	satisfying	relevant	australian
kansas city mo	meet	respectively	kilograms	whereas	shortlisted
von willebrand disease	acceptable	competition	article	euroleague	encyclopedia
septic system design	satisfactory	expenditure	anglia	inspect	derbyshire
newyork hotels	desired	programme	difference	trophy	barnet
adobe indian houses	acceptable	service	##4	.	pennsylvania
yahoo	acceptable	platforms	##mark	oro	eel
diversity	unacceptable	retrieval	champaign	vacancy	index
barbados	unacceptable	civil	result	uttar	##own
titan	junctions	hot	##rak	favour	flanders
neil young	acceptable	resulting	category	.	stockport
voyager	acceptable	desired	┆	specified	consortium
map of the USA	income	storage	bhutan	rayon	oclc
vines for shade	contents	compromised	somme	##worthy	selects
illinois state tax	visually	threatened	cale	qualified	forum
bobcat	merit	infectious	situated	united	mammal
gs pay rate	meetings	criterion	relation	honour	queensland
south africa	acceptable	cumulative	cornered	non	fein
uplift at yellowstone	award	enveloped	...	realised	in
espn sports	acceptable	contents	.	doe	##tsk
indexed annuity	»	supported	meeting	fund	deutsche
starbucks	acceptable	gaa	domestic	revenues	##�
cheap internet	received	device	acceptable	copa	##bc
income tax return online	satisfactory	incurred	barking	##kei	australia
website design hosting	acceptable	incumbent	client	library	luton
black history	favour	recipient	vacancy	platform	smashwords
appraisals	significance	incumbent	foundation	person	realised
iowa food stamp	benefit	register	maryland	medium	registers
vldl levels	achieved	share	##ted	conditional	displayed
cass county missouri	qualifying	individuals	when	##vu	favour
hp mini 2140	meaningful	acceptable	host	copa	service
all men created equal	acceptable	partial	postage	pennsylvania	paper
interview thank you	acceptable	populated	swiss	##kei	whilst
animal cuts reviews	acceptable	substrate	urbana	##puri	##kei
dieting	rattled	registered	taxon	backdrop	slovakia
	0	1	2	3	4

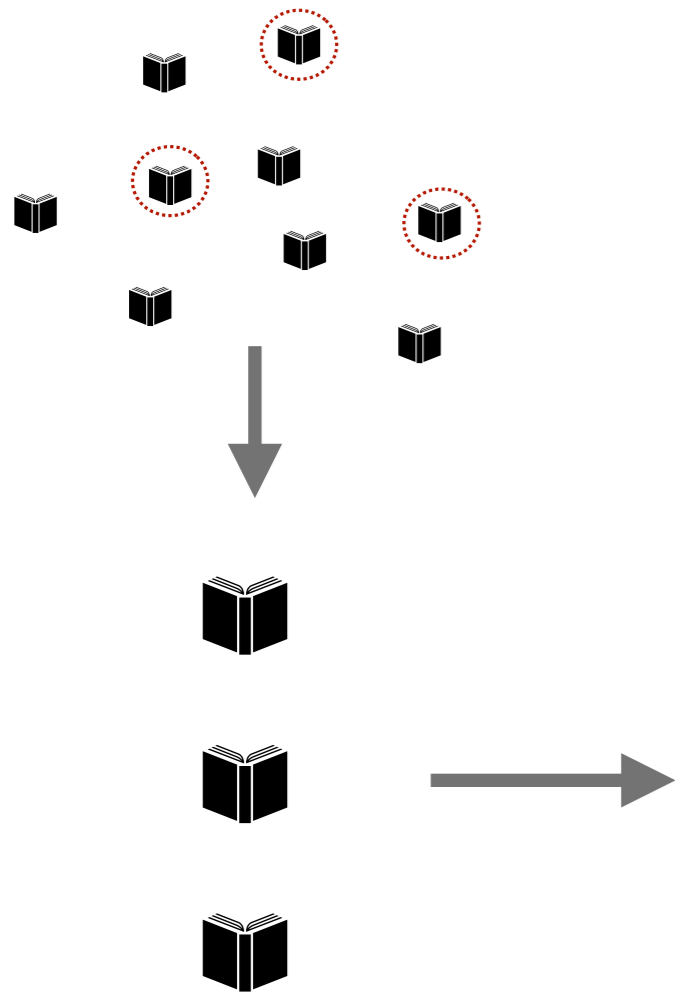
Question Answering

Where is the worlds largest flower garden located ?



Question Answering

Where is the worlds largest flower garden located ?



CNN travel

DESTINATIONS FOOD & DRINK NEWS STAY VIDEO Q

Dubai Miracle Garden: The power of the flower

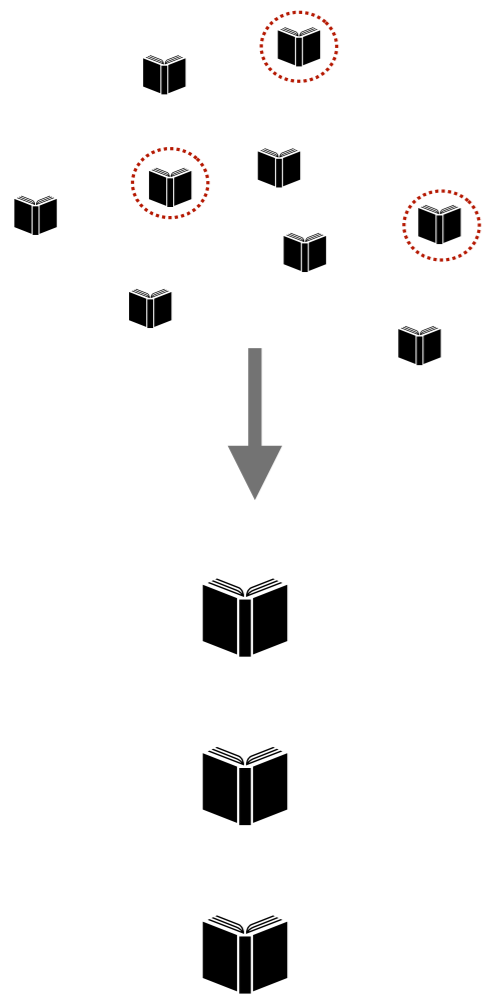
CNN Travel staff • Updated 22nd August 2017

The Dubai Miracle Garden is certainly aptly named considering that -- like pretty much everything in this Middle Eastern destination -- it was built on desert land. Billing itself as the world's largest natural flower garden, the 72,000-square-meter attraction has more than 60 million flowers on display.



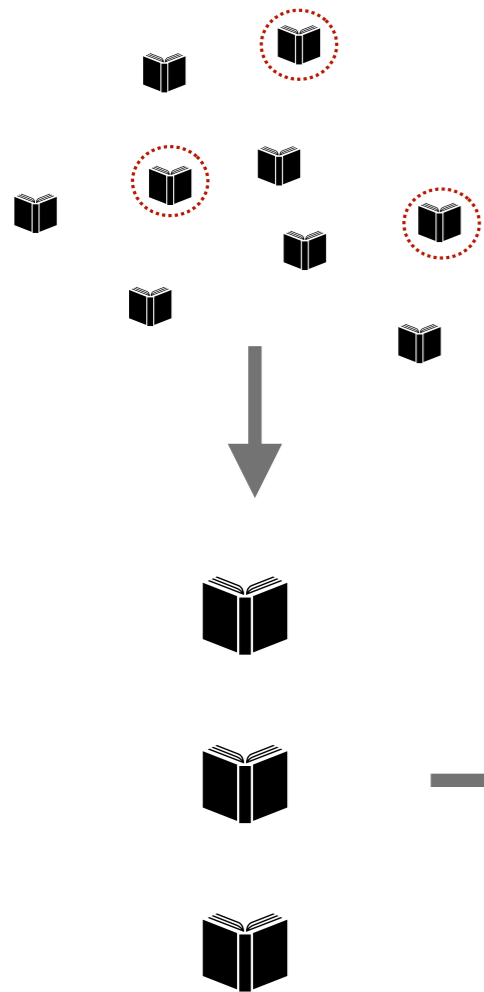
Fact Checking

Query: **san francisco bay area contains zero towns**



Fact Checking

Query: san francisco bay area contains zero towns



Retrieved Document: the san francisco bay area, referred to locally as the bay area is a populous region surrounding the san francisco and san pablo estuaries in northern california. The **region encompasses the major cities and metropolitan areas** of san jose, san francisco, and Oakland, along with smaller urban and rural areas. The bay area's nine counties areSanta Clara, Solana and Sonoma. The combined statistical area of the region is the second largest in california after the Los Angeles area.

Many Others

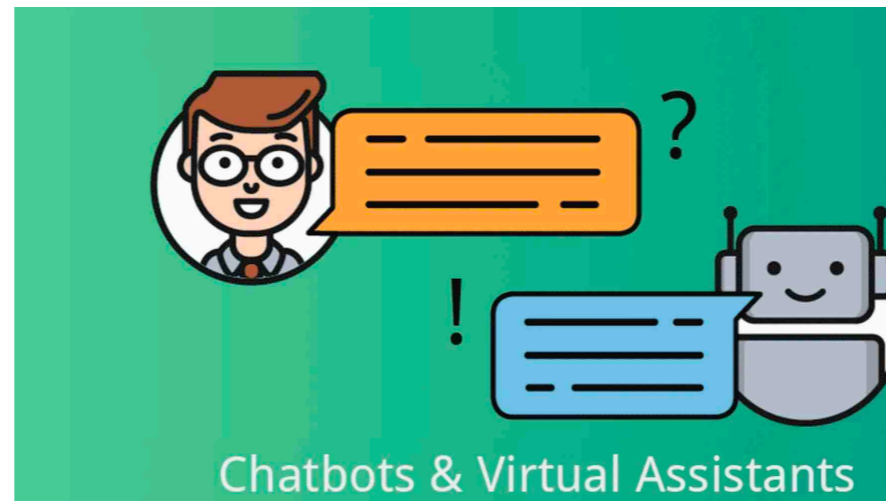
Knowledge-base Construction

The screenshot shows a web interface for creating an entity page. It has four main sections:

- Names:** A form with 'Full Name' (containing 'Finn') and 'Other Names' (with a text area for 'Alternative names for your entity').
- Description:** A form with 'Descriptive Phrases' (containing 'Star Wars') and a note: 'Describing your entity with phrases. Press enter after'.
- Is this about your entity?:** A section with a small image and text: "'Star Wars': The Force Of Nostalgia Is Strong With This One ...sounds a lot like an old villain, scrappy female scavenger Rey (Daisy Ridley), who's also a pilot, and Finn (John Boyega), a Stormtrooper gone AWOL, who decides Rey needs protecting, whether she wants it or fasty enough to banish thoughts of Katrina Everdeen from the most devoted Hunger Games enthusiast, Finn is the sort of impetuous, can-do hero who'll inspire a whole new generation of Star Warriors, and ...'. There are thumbs up/down icons.
- Automatic Description:** A section with 'Descriptive Phrases' and a list of tags: 'John Boyega', 'planet Jakuu', 'stormtrooper Finn', 'Princess Leia', 'Star Wars', and 'Jedi'.

At the bottom, it says 'Are You Looking For This? Maybe we already know your entity!'.

Conversational AI



Citation Discovery

The image shows two side-by-side weather reports. The left report is for a 'Super cyclonic storm (IMD scale) Category 5 (Saffir-Simpson scale)'. The right report is for 'Category 5 major hurricane (SSHWS/NWS) Hurricane Katrina at peak strength on August 28, 2005'.

Super cyclonic storm (IMD scale) Category 5 (Saffir-Simpson scale)	Category 5 major hurricane (SSHWS/NWS) Hurricane Katrina at peak strength on August 28, 2005
Formed October 25, 1999	Formed August 23, 2005
Dissipated November 3, 1999	Dissipated August 31, 2005 ^[1] (Extratropical after August 30, 2005)
Highest winds 3-minute sustained: 260 km/h (160 mph)	Highest winds 1-minute sustained: 175 mph (280 km/h)
Lowest pressure 912 mbar (hPa); 26.93 inHg	Lowest pressure 902 mbar (hPa); 26.64 inHg
Fatalities ~10,000 direct	Fatalities 1,833 in U.S. confirmed ^[1]
Damage \$4.5 billion (1999 USD)	Damage \$108 billion (2005 USD)
Areas affected India, Myanmar	
Part of the 1999 North Indian Ocean cyclone season	

Wide variety of tasks that can be solved using intelligent algorithms with access to world knowledge

Solution Framework

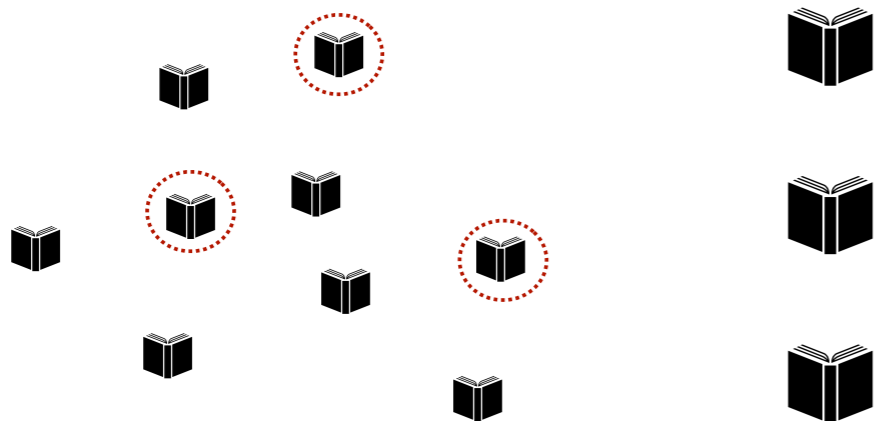
Knowledge Intensive Tasks

Question
Answering

Fact
Verification

Knowledge
Enrichments

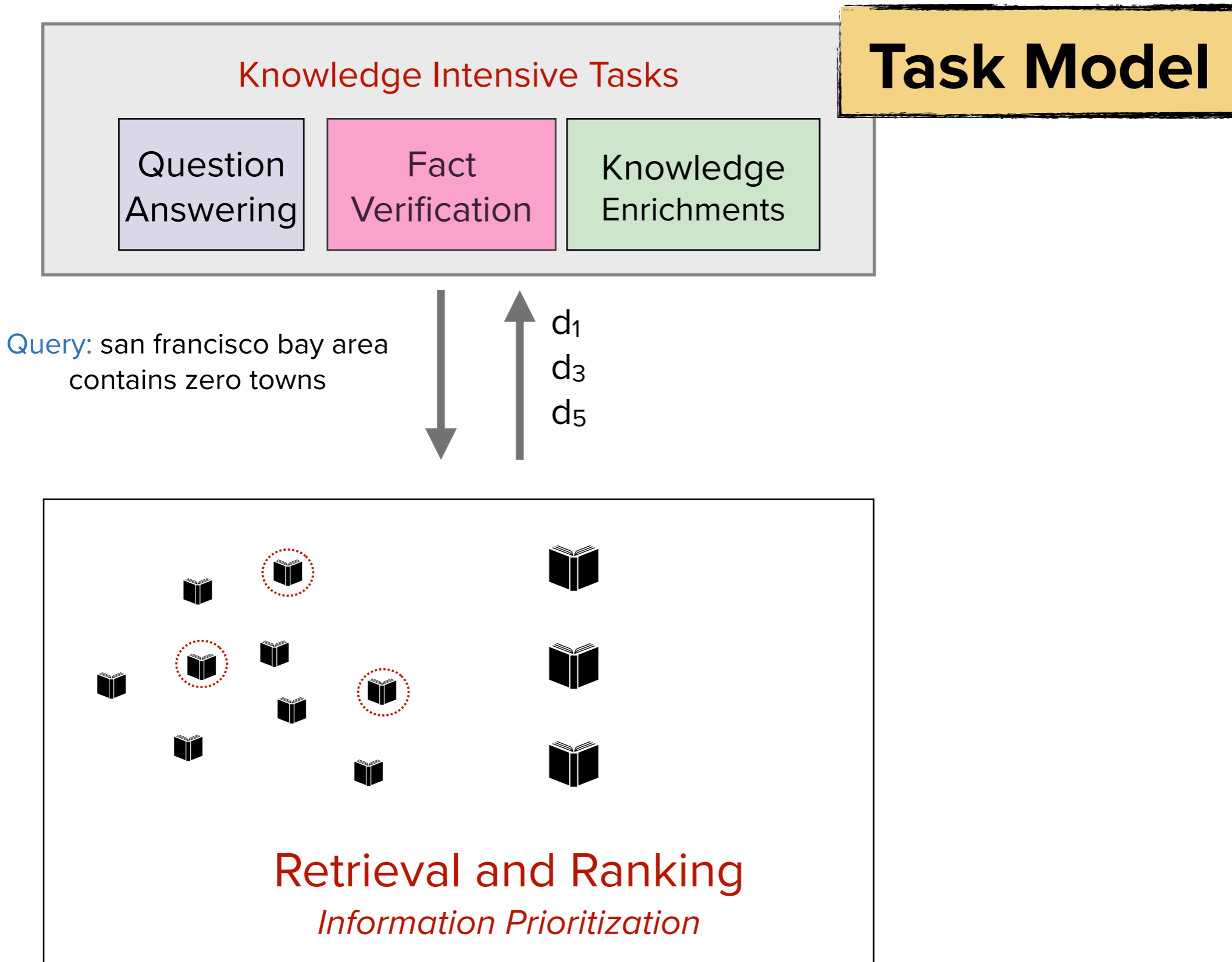
Query: san francisco bay area
contains zero towns



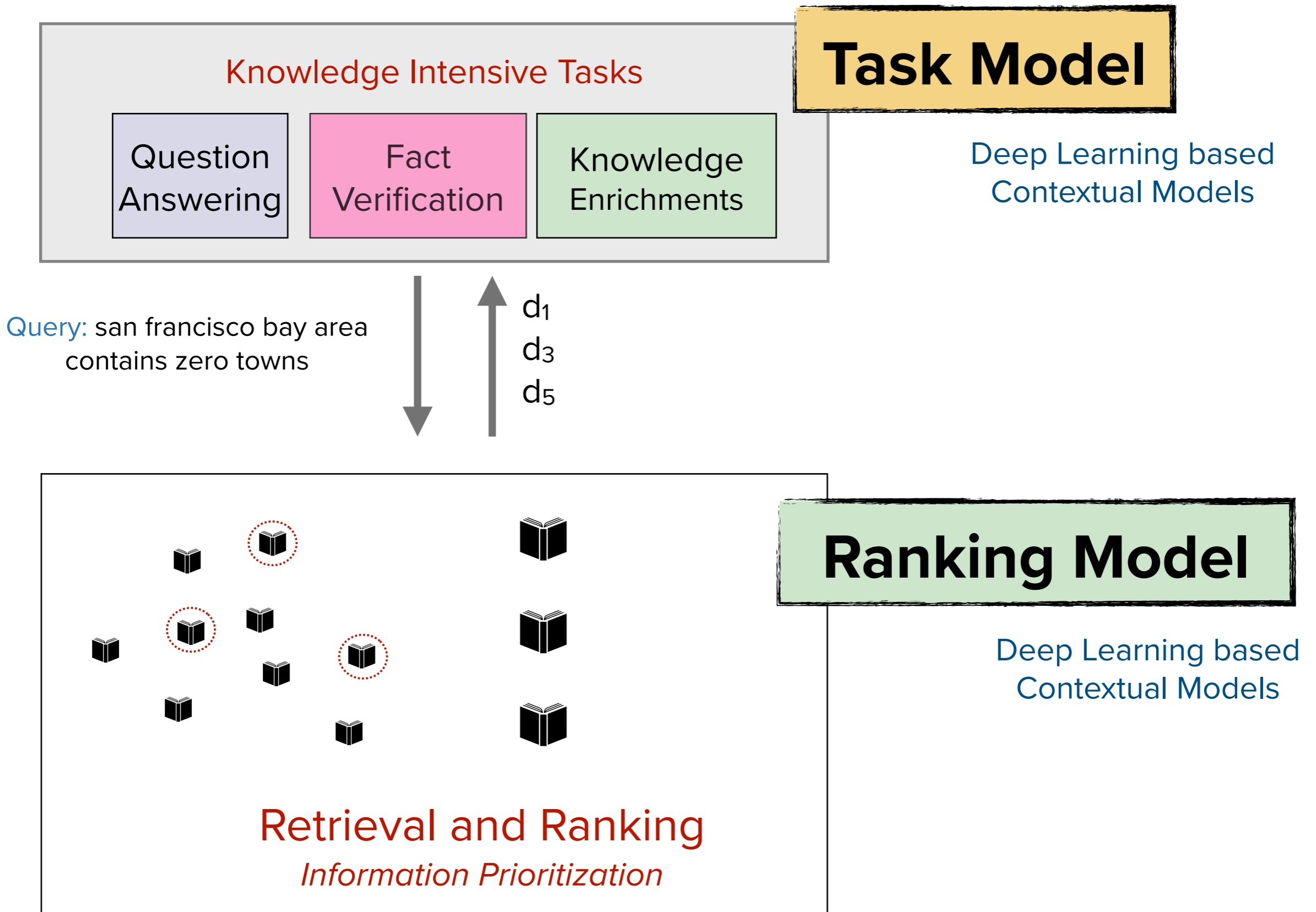
Retrieval and Ranking
Information Prioritisation

Ranking Model

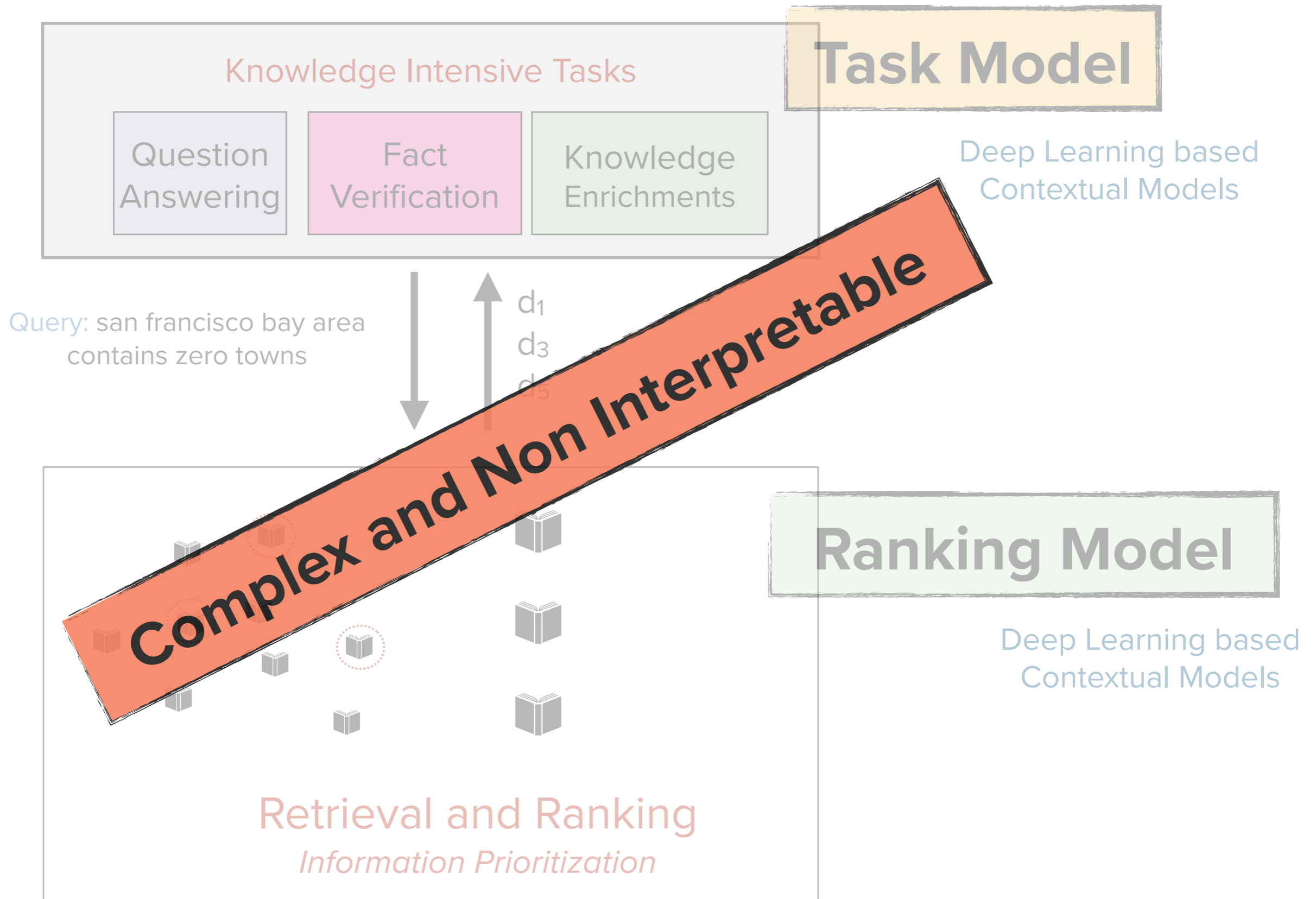
Solution Framework



Solution Framework



Solution Framework



Why interpretability ?

Query: san francisco bay **area** contains zero towns



Evidence Document: the **Boston area**, encompasses the major cities and **metropolitan areas**. The Boston area is a bustling region for economic and commercial growth. This is one of the oldest developed metro areas in the united states of America.

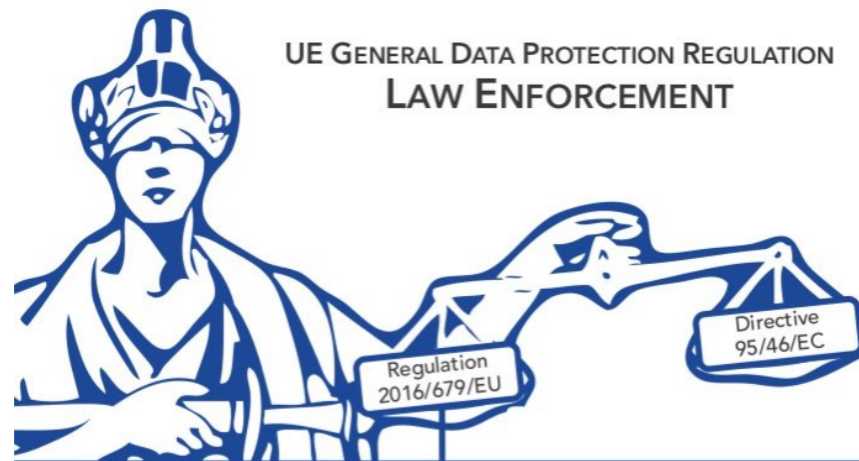
Right for the Right Reason

Why interpretability ?

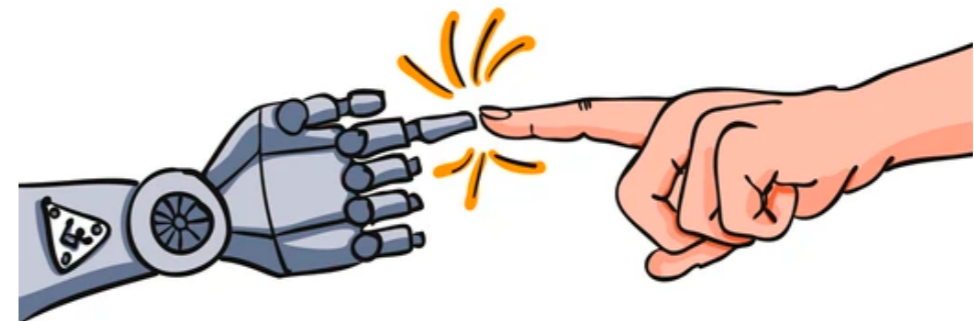
Interpretable algorithms for Knowledge Intensive Tasks



Utilize insights to improve models



Legal recourse



Improve trust

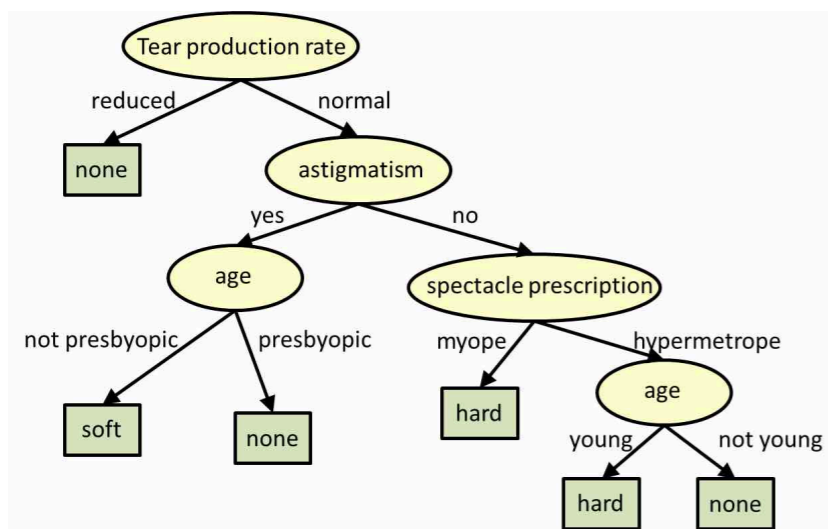
Notions in ExIR

Explainable Information Retrieval

Interpretable models

We say something is **interpretable** is if its capable to be understood by a human on its own

How does the AI system arrive at its decisions or predictions?



$$\sum_{i \in q} \log \frac{N}{df_i} \cdot \frac{(k_1 + 1)tf_i}{k_1((1 - b) + b \frac{dl}{avdl}) + tf_i}$$

Interpretable models are designed to have transparent and understandable structures, making it easier to trace and comprehend the factors that influence the system's outputs.

Explainability vs Interpretability

Explainable methods are additional methods to understand a complex model

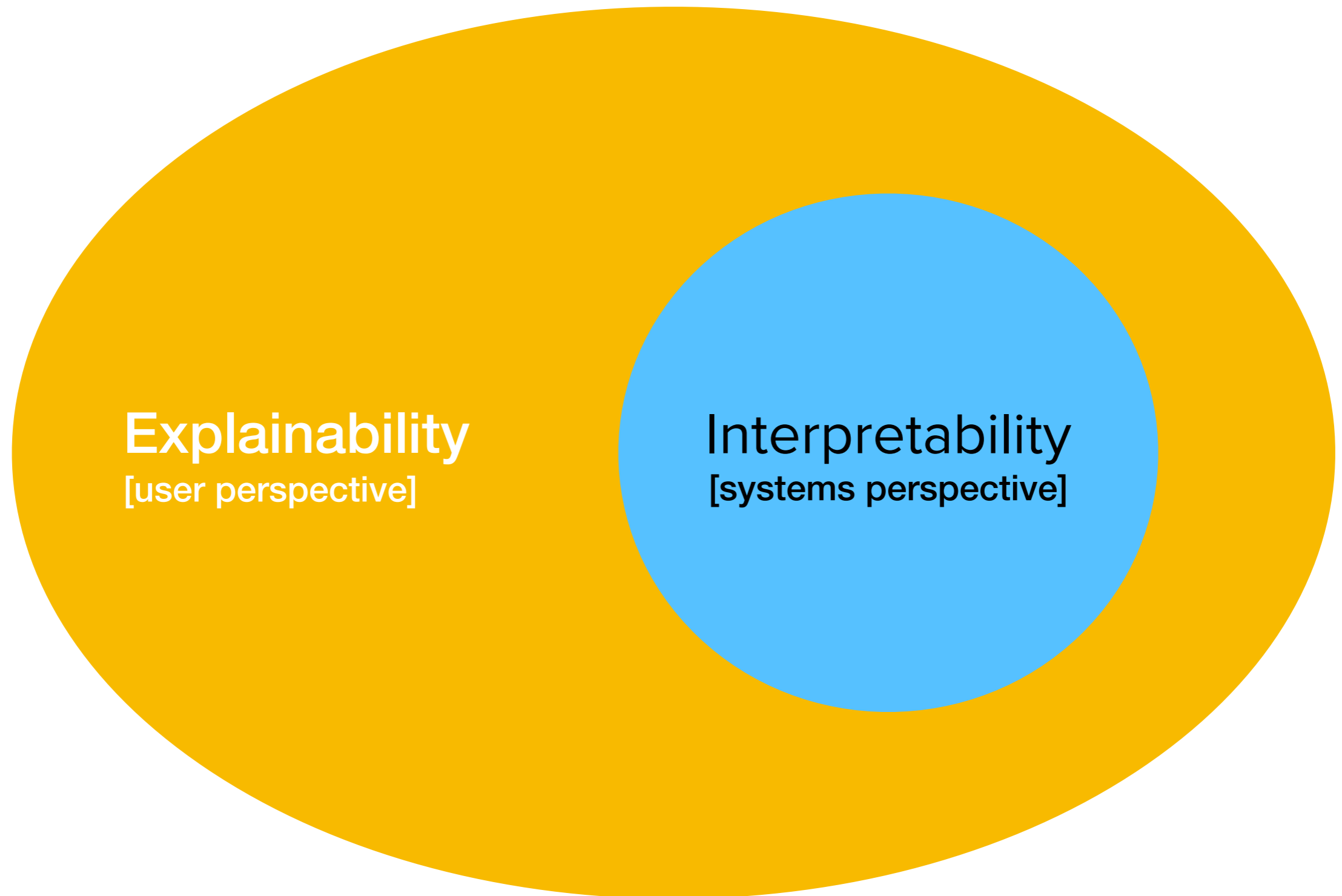
Explainable methods aim to produce human-understandable explanations in natural language or other interpretable forms. [users perspective]

Interpretability is the ability to explain or to present in understandable terms to a human

[Doshi-velez & Kim]

No Consensus

Explainability vs Interpretability



What questions are we interested in ?

- Which features are responsible for a document ranking ?
- Which data instance are responsible for a classification ?
- Which parameters are responsible ?
- Why is one document ranked higher than the other for my query ?
- What happens if we vary the query/document content`s ?
- ...

Is there ONE interpretability ?



Multiple stakeholders

Multiple Explanations

Multiple Methods

Feature attributions

Heatmaps, Saliency Maps, Attributions, soft masks

How to find the Mean ?	Search
------------------------	--------

Ranking model 1

Query Explanation

How to find the Mean ?

Doc Explanation

Mean is the average of the input arguments

Ranking model 2

Query Explanation

How to find the Mean ?

Doc Explanation

Judy was the meanest of the girls

Free-form text

How to find the Mean ?

Search

Model 1

Explanation terms

X, statistics, plus

Model 2

Explanation terms

Meaning, definition, dictionary

- Terms from the potentially relevant documents
- Topics mined from relevant documents

Extractive explanations

How to find the Mean ?

Search

Heatmaps

Free-text Explanation

Model 1

How to find the Mean ?

X, statistics, plus

Extractive Explanation: The **mean** is the average of the numbers and it is easy to **calculate**: add up all the numbers, then divide by how many numbers there are. In other words it is the sum divided by the count. How do you handle negative numbers? Adding a negative number is the same as subtracting the number (without the negative). For example $3 + (-2) = 3 - 2 = 1$.

Pointwise, Pairwise, Listwise Explanations

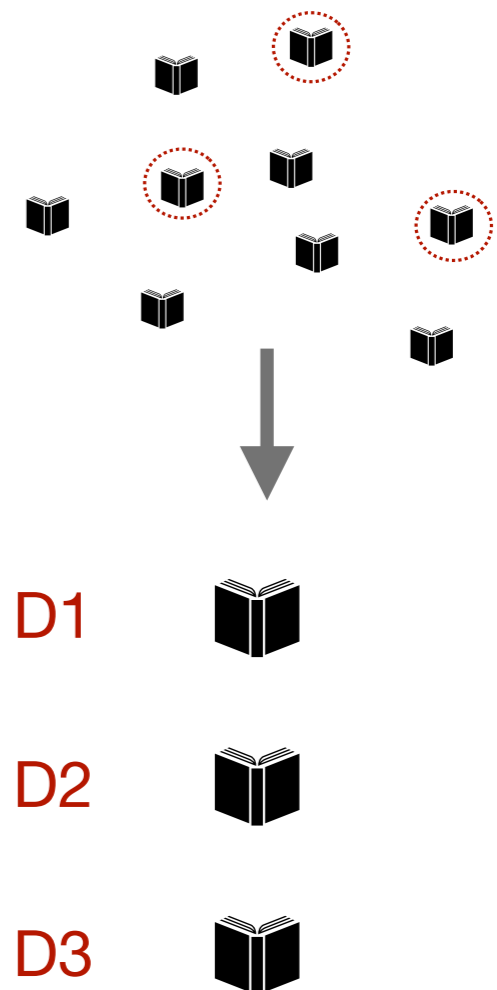
The scope of explanations for ranking tasks



Pointwise explanations:
Why is a document relevant ?

Pointwise explanations:
Why is a document more relevant than another ?

Listwise explanations:
Why is a ranking relevant ?



Approaches in ExIR

Explainable Information Retrieval

Approach Families

Interpretable algorithms for Knowledge Intensive Tasks

During Model Building

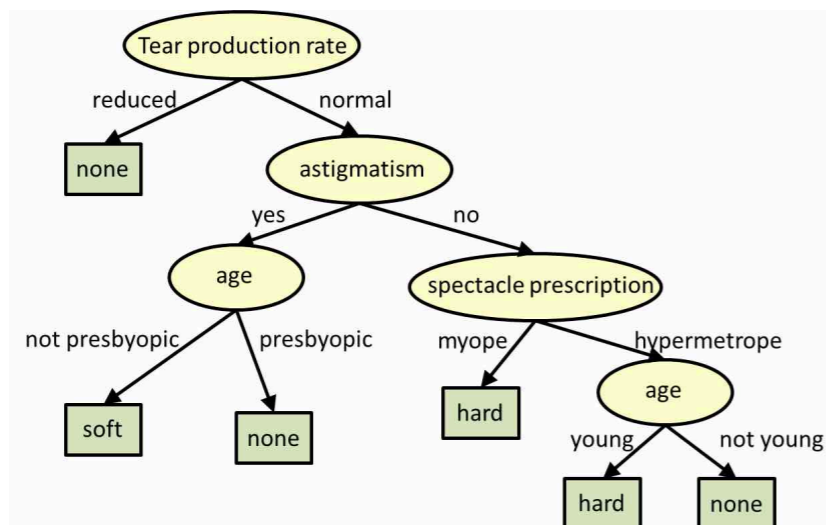
Interpretable by design

Accuracy vs Interpretability

After Model Building

Posthoc Interpretability

No compromise on accuracy



How to find the Mean ?

Approach Families

Interpretable algorithms for Knowledge Intensive Tasks

During Model Building

Interpretable by design

Accuracy vs Interpretability

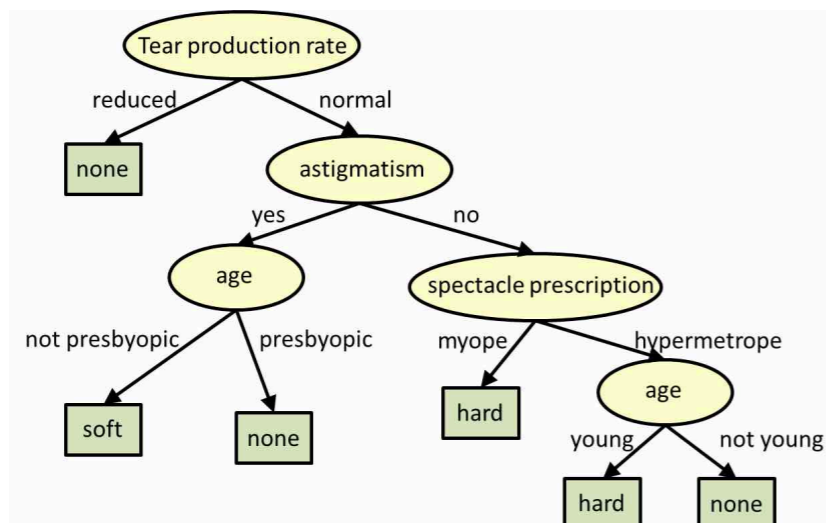
Explanations 100% reliable

After Model Building

Posthoc Interpretability

No compromise on accuracy

Hard to measure reliability



How to find the Mean ?

Mapping to IR Abilities

How to find the Mean ?

Search

Does Ranking models understand world knowledge ?

Does Ranking models understand matching, BM25, entity matching ?

Does Ranking models understand IR abilities encoded in IR axioms?

Interpretability Landscape

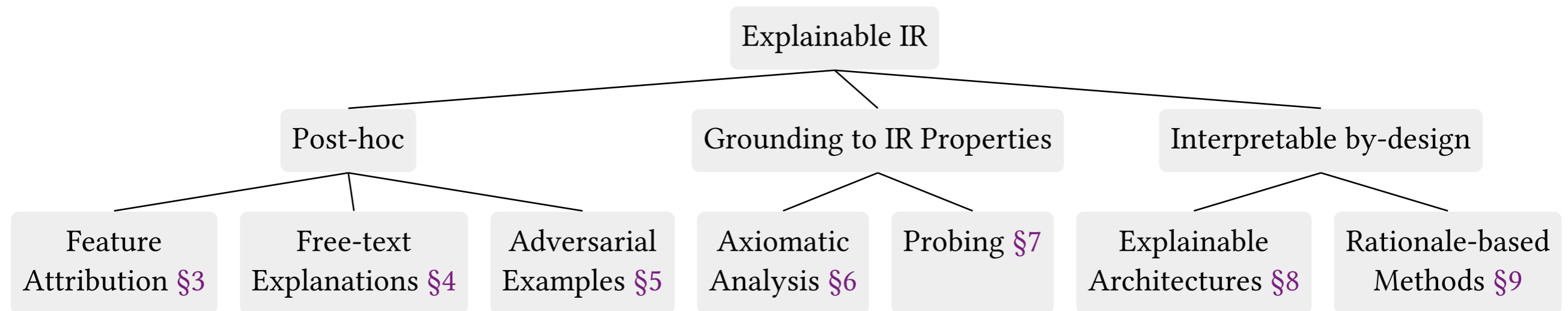
Explainable Information Retrieval: A Survey

<https://arxiv.org/abs/2211.02405>

AVISHEK ANAND and LIJUN LYU, Delft University of Technology, The Netherlands

MAXIMILIAN IDAHL, YUMENG WANG, JONAS WALLAT, and ZIJIAN ZHANG, L3S Research

Center, Leibniz University Hannover, Germany



Schedule

Introduction, motivation and notions

Posthoc interpretability

Intrinsic interpretability or Interpretability by design

Probing LLMs

Axiomatic IR for explaining IR models

Demo

Evaluation or ExIR methods

Conclusion and open problems

Interpretability Landscape

Explainable Information Retrieval: A Survey

<https://arxiv.org/abs/2211.02405>

AVISHEK ANAND and LIJUN LYU, Delft University of Technology, The Netherlands

MAXIMILIAN IDAHL, YUMENG WANG, JONAS WALLAT, and ZIJIAN ZHANG, L3S Research

Center, Leibniz University Hannover, Germany

