Explainable Information Retrieval

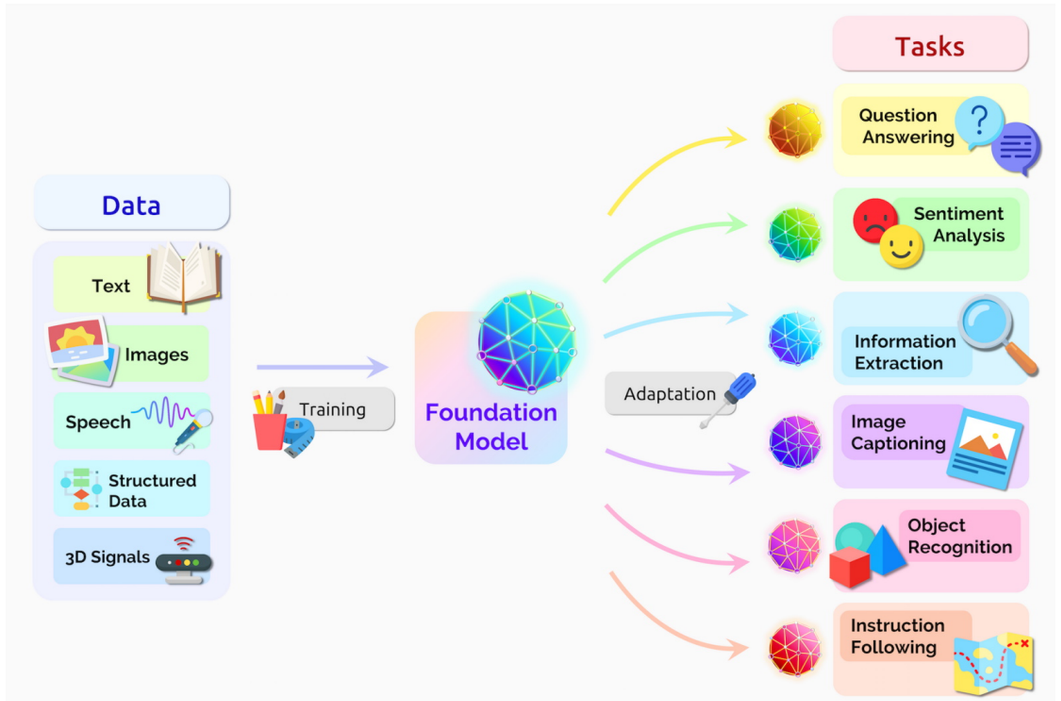
Avishek Anand, Procheta Sen, Manisha Verma, Sourav Saha, Mandar Mitra

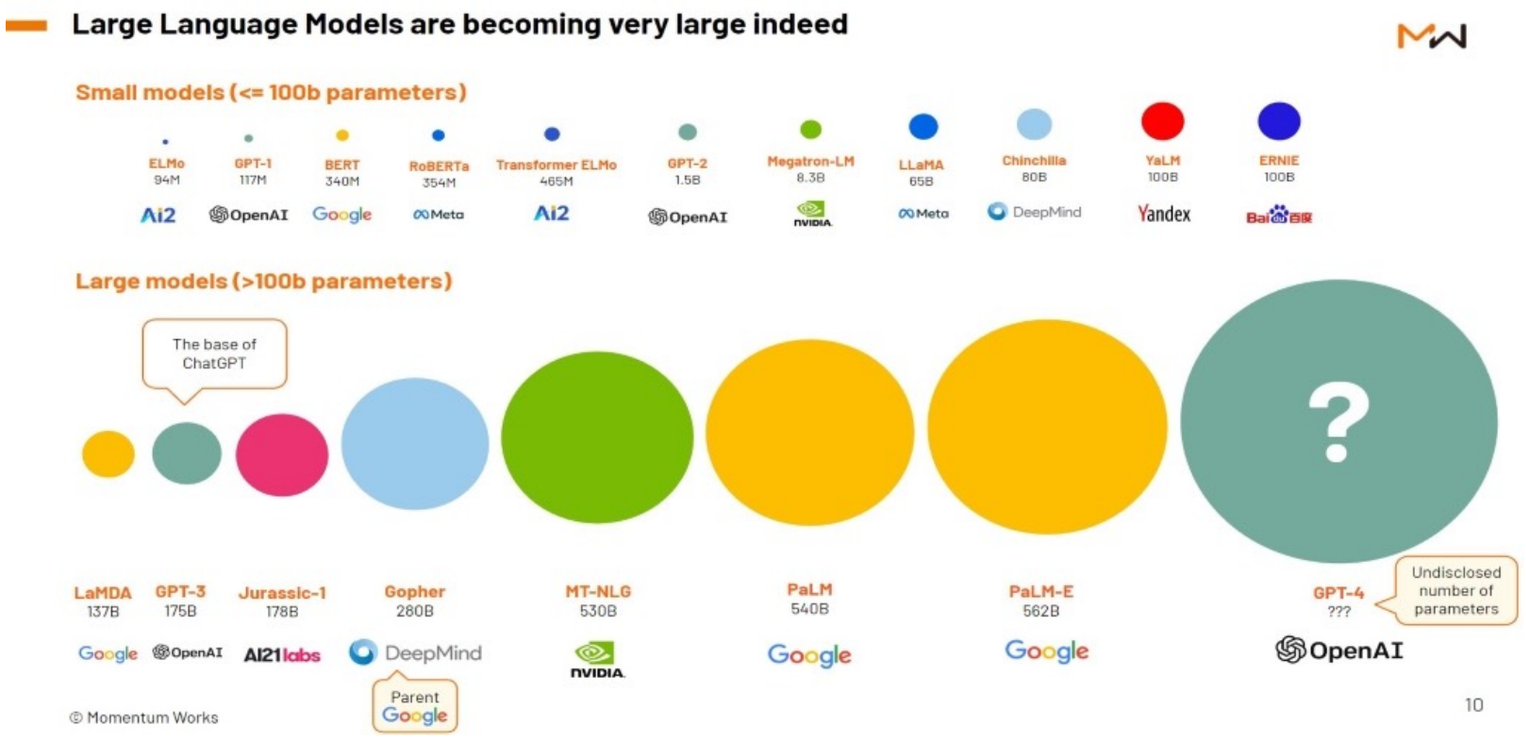
avishek.anand@tudelft.nl procheta.sen@liverpool.ac.uk mandar.mitra@gmail.com

souravsaha.juit@gmail.com manishaverma.21@gmail.com

TU Delft

The advent of foundational models



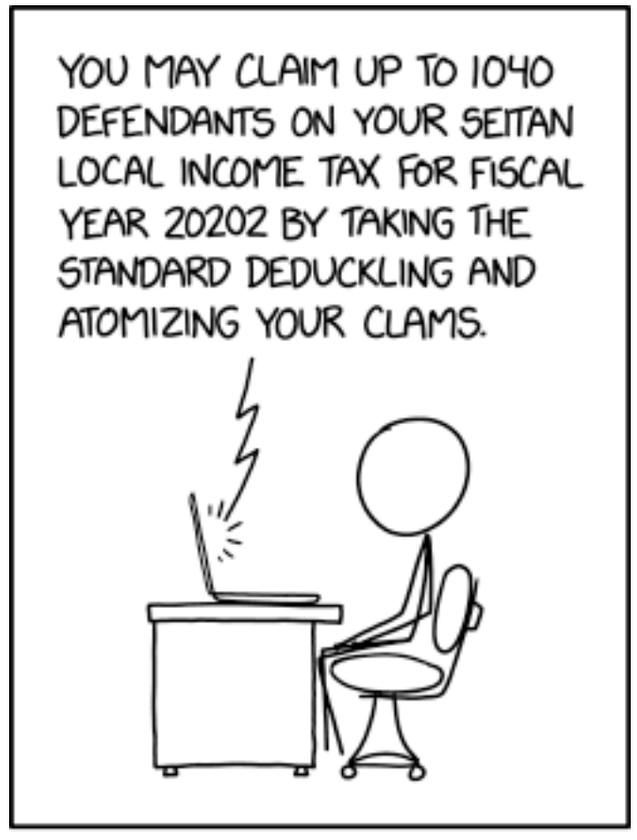


https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/


BUT HOW DO I KNOW
I CAN TRUST YOU?

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,..



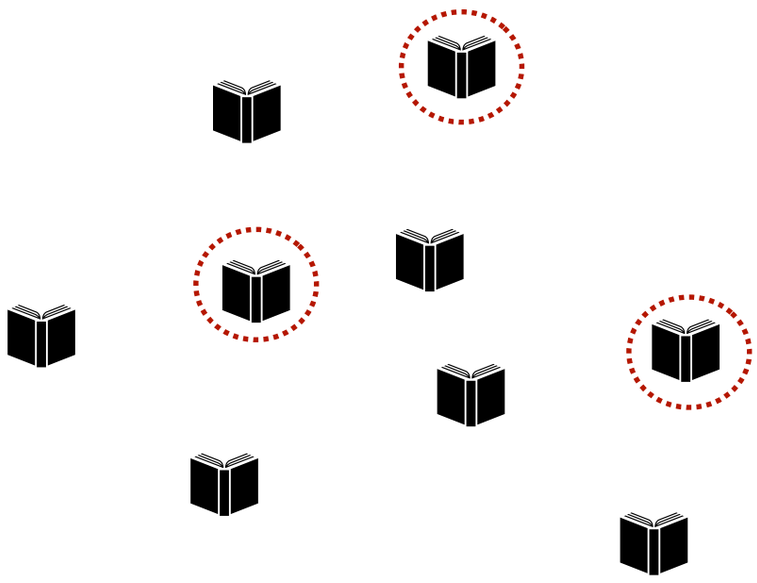
Credit: https://xkcd.com/2265/

Why, When, and What

Explainable Information Retrieval

Information Prioritisation in IR

The Document Ranking Task



Millions or Billions of documents

Ranking

Deep Models for Ranking

The Document Ranking Task

Health hazards | Search

Ranking Model

Deep Learning based
Contextual Models

Millions or Billions of documents

Ranking

Why interpretability ?

Rank 1 Logitech light speed is nothing but the best keyboard acc. to rankings

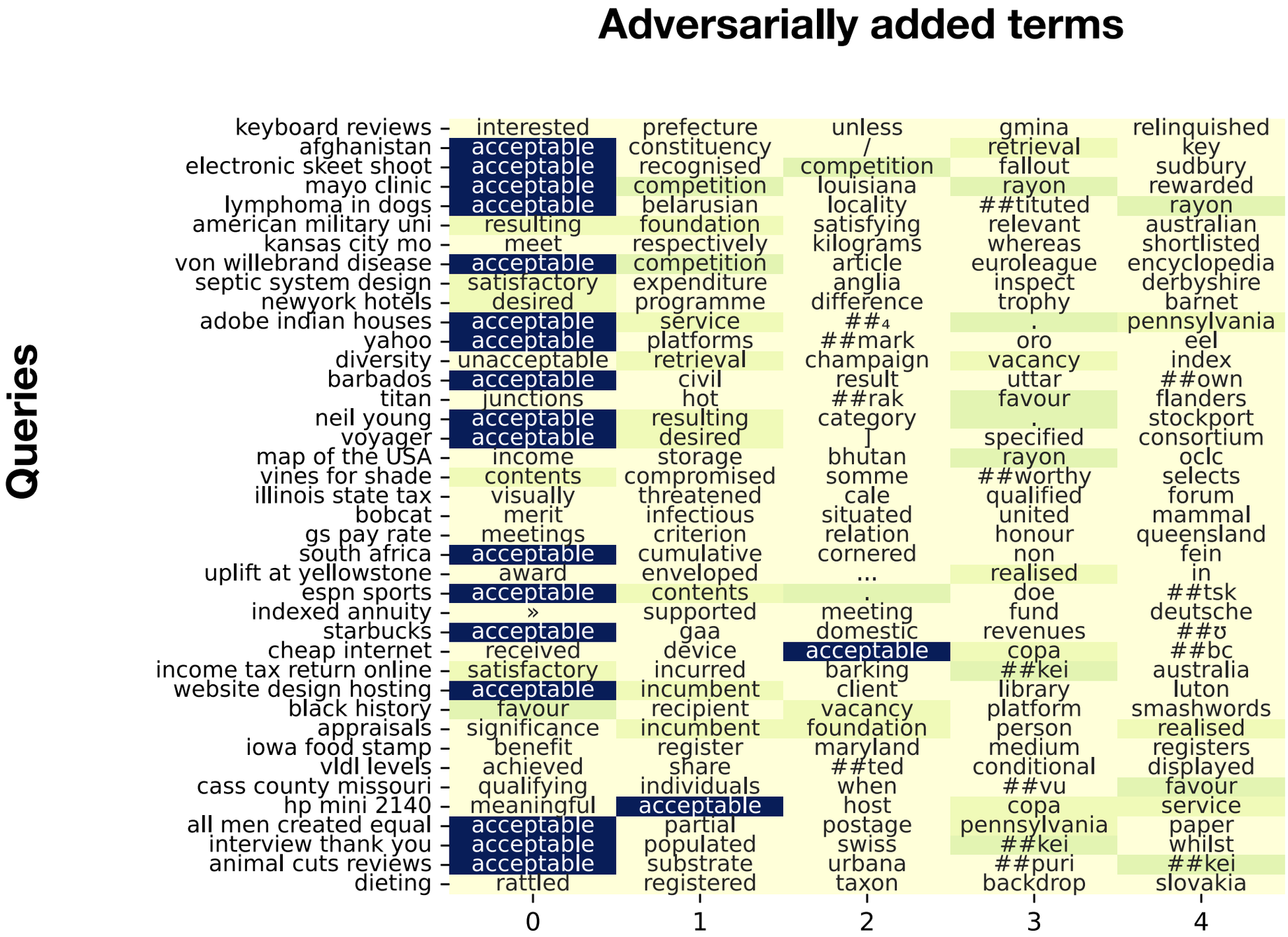
Big drop

Add the term “acceptable”

Rank 20 acceptable Logitech light speed is nothing but the best keyboard acc. to rankings

Bert rankers are brittle, Wang, Lyu & Anand, [ICTIR 2022]

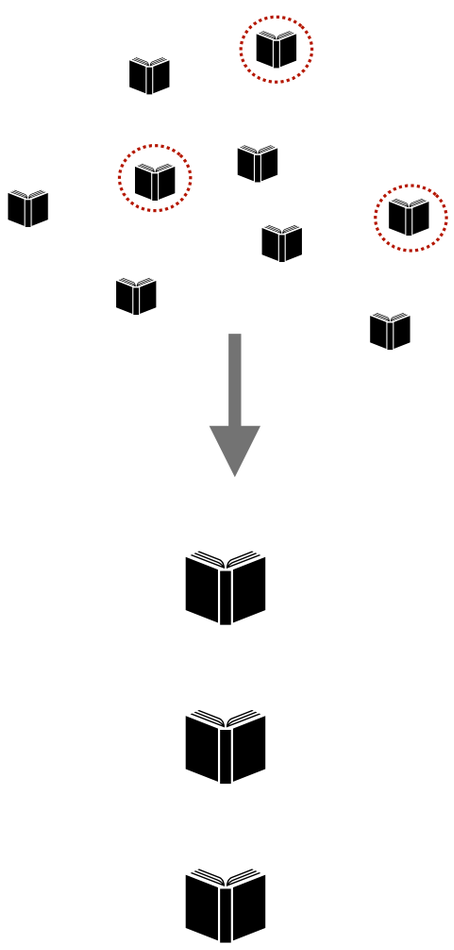
Why interpretability ?



Bert rankers are brittle, Wang, Lyu & Anand, [ICTIR 2022]

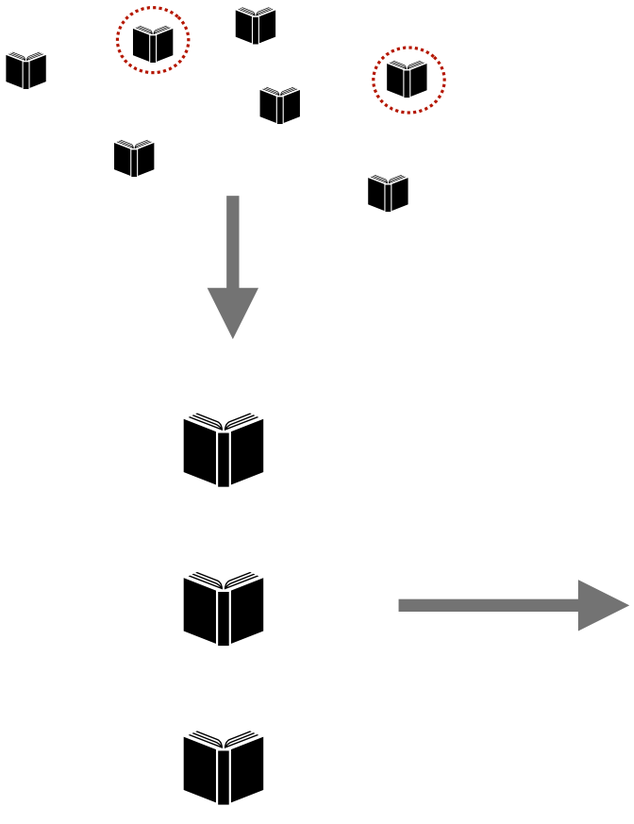
Question Answering

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

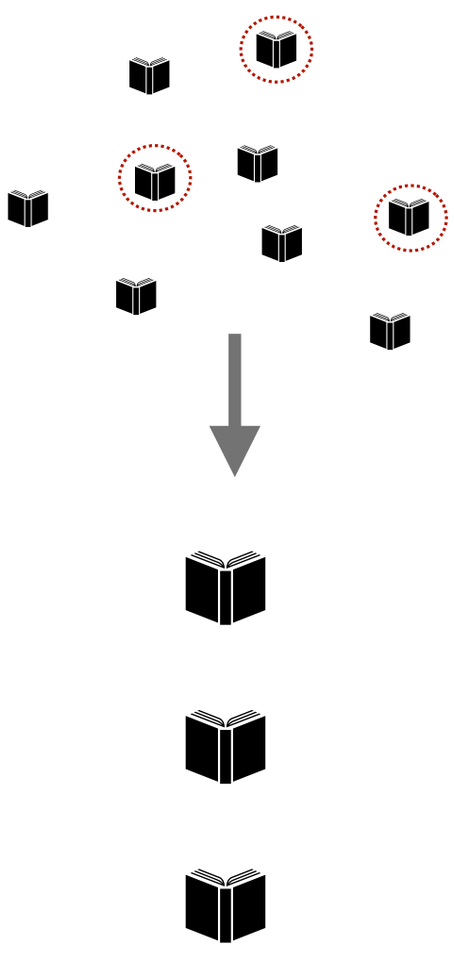
Dubai Miracle Garden: The power of the flower

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

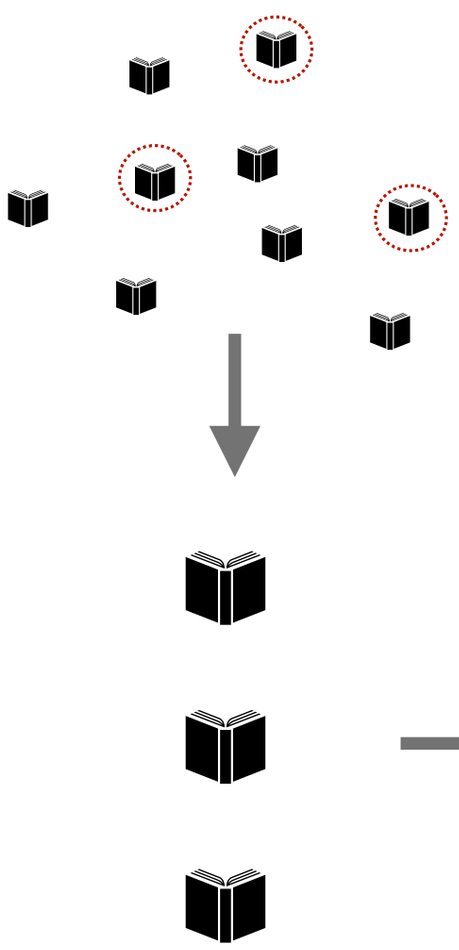
Query: **san francisco bay area contains zero towns**



Explain and Predict, and then Predict Again, Zhang, Rudra & Anand, [WSDM 2021]

Fact Checking

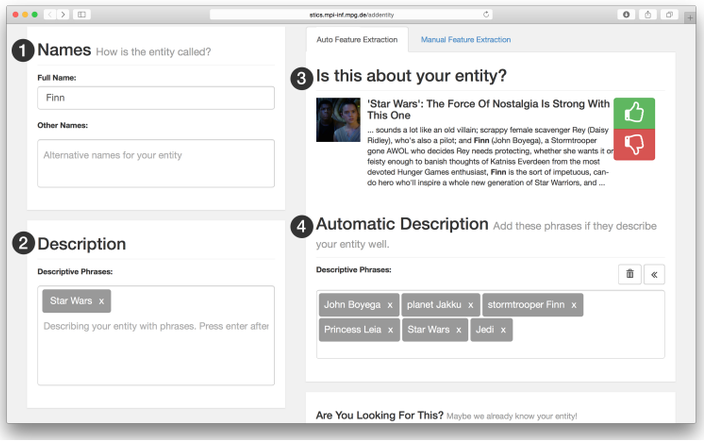
Query: **san francisco bay area contains zero towns** ✗



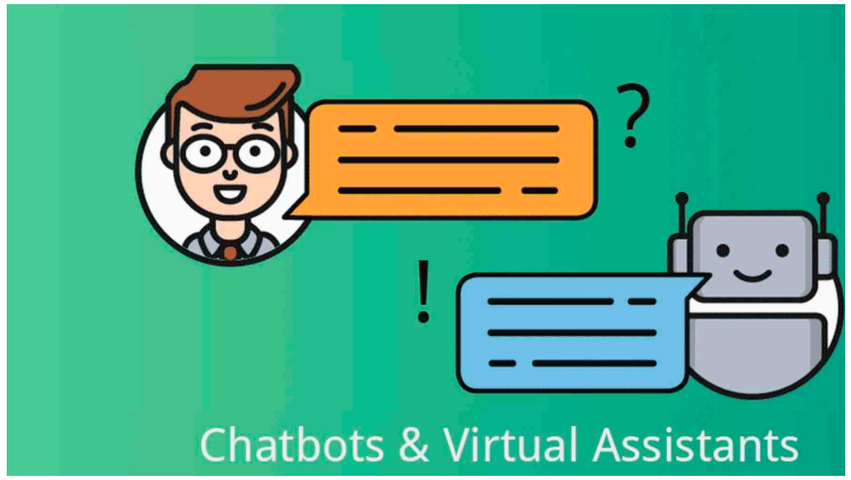
Retrieved Document: the san francisco bay area, referred to locally as the bay area is a populous region surrounding the san francisco and san pablo estuaries in northern california. The region encompasses the major cities and metropolitan areas of san jose, san francisco, and Oakland, along with smaller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

Many Others

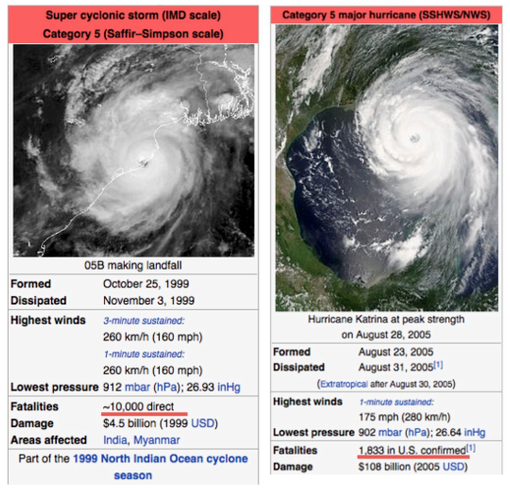
Knowledge-base Construction



Conversational AI



Citation Discovery



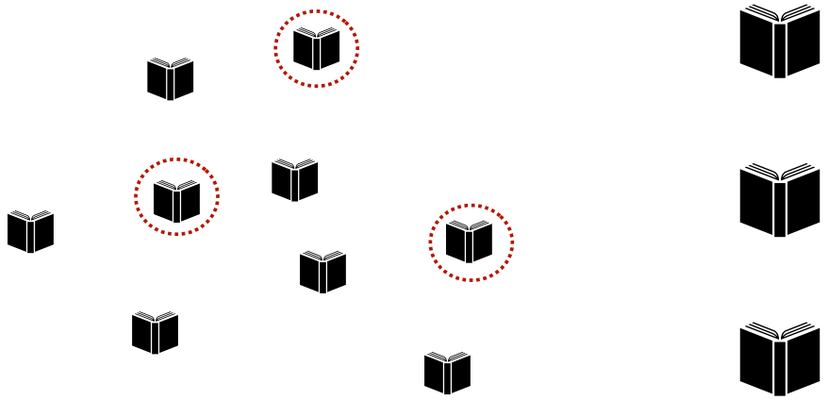
Wide variety of tasks that can be solved using intelligent algorithms with access to world knowledge

Solution Framework

Knowledge Intensive Tasks

Question Answering | Fact Verification | Knowledge Enrichments

Query: san francisco bay area contains zero towns

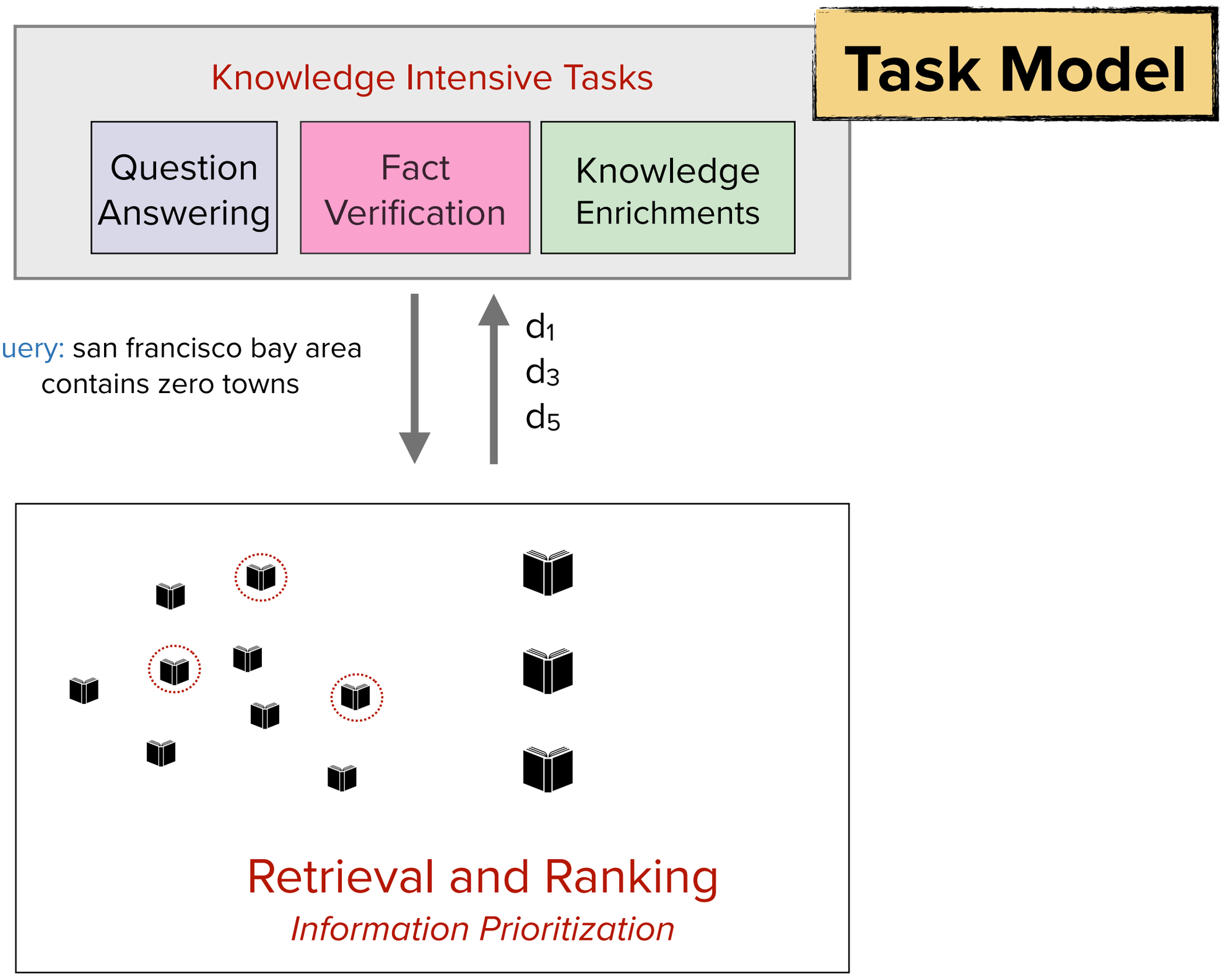


Retrieval and Ranking

Information Prioritisation

Ranking Model

Solution Framework



Solution Framework

Knowledge Intensive Tasks

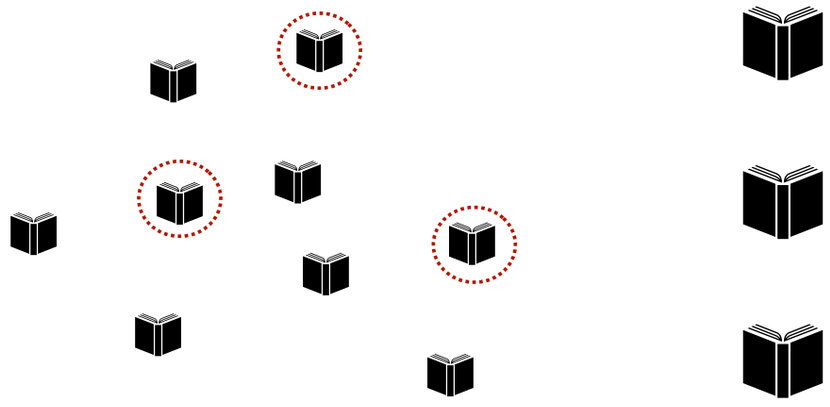
Question Answering | Fact Verification | Knowledge Enrichments

Task Model

Deep Learning based Contextual Models

Query: san francisco bay area contains zero towns

d_1
d_3
d_5



Retrieval and Ranking

Information Prioritization

Ranking Model

Deep Learning based Contextual Models

Solution Framework

Knowledge Intensive Tasks

Question Answering

Fact Verification

Knowledge Enrichments

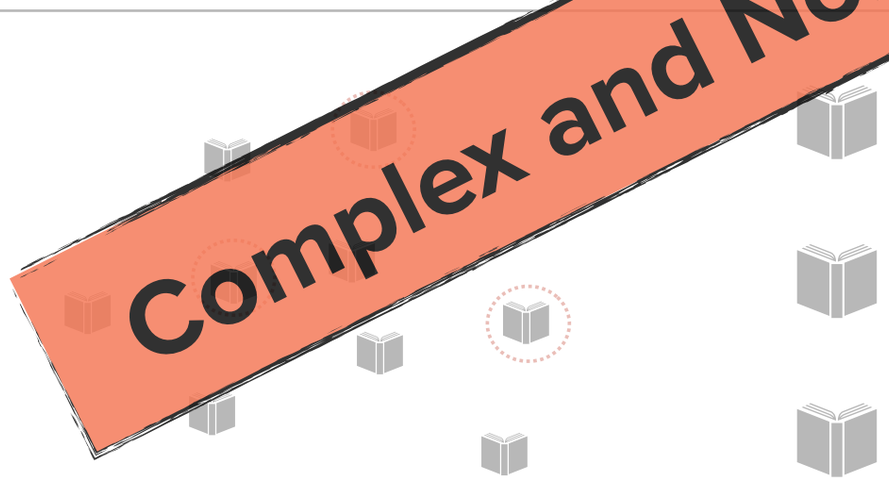
Task Model

Deep Learning based Contextual Models

Query: san francisco bay area contains zero towns

d_1
d_3
d_5

Complex and Non Interpretable



Retrieval and Ranking

Information Prioritization

Ranking Model

Deep Learning based Contextual Models

Why interpretability ?

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

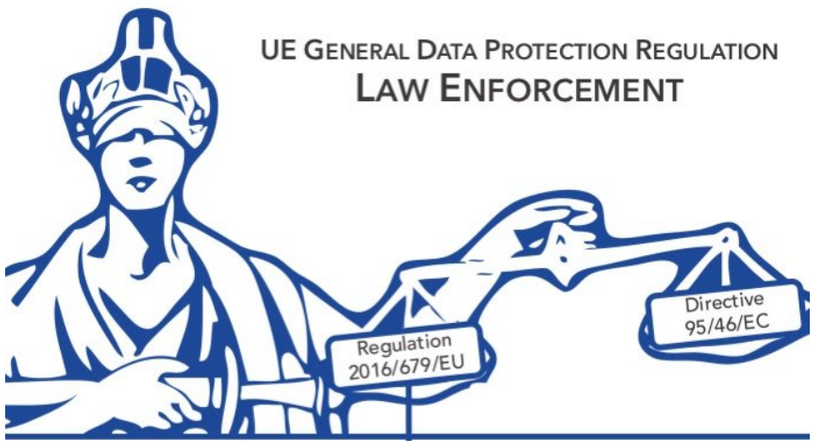
Evidence Document: the **Boston area**, encompasses the major cities and metropolitan areas. The Boston area is a bustling region for economic and commercial growth. This is one of the oldest developed metro areas in the united states of America.

Right for the Right Reason

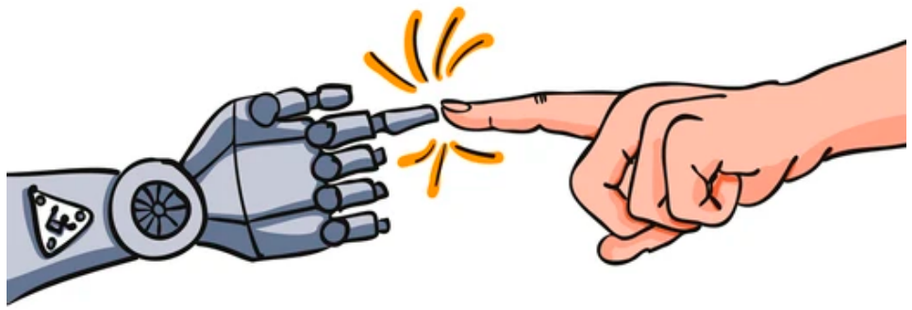
Why interpretability ?

Interpretable algorithms for Knowledge Intensive Tasks

Utilize insights to improve models



Legal recourse



Improve trust

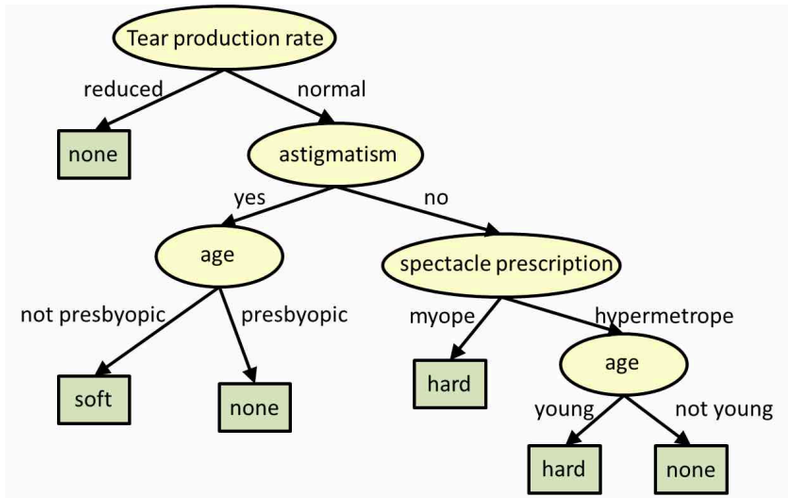
Notions in ExIR

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

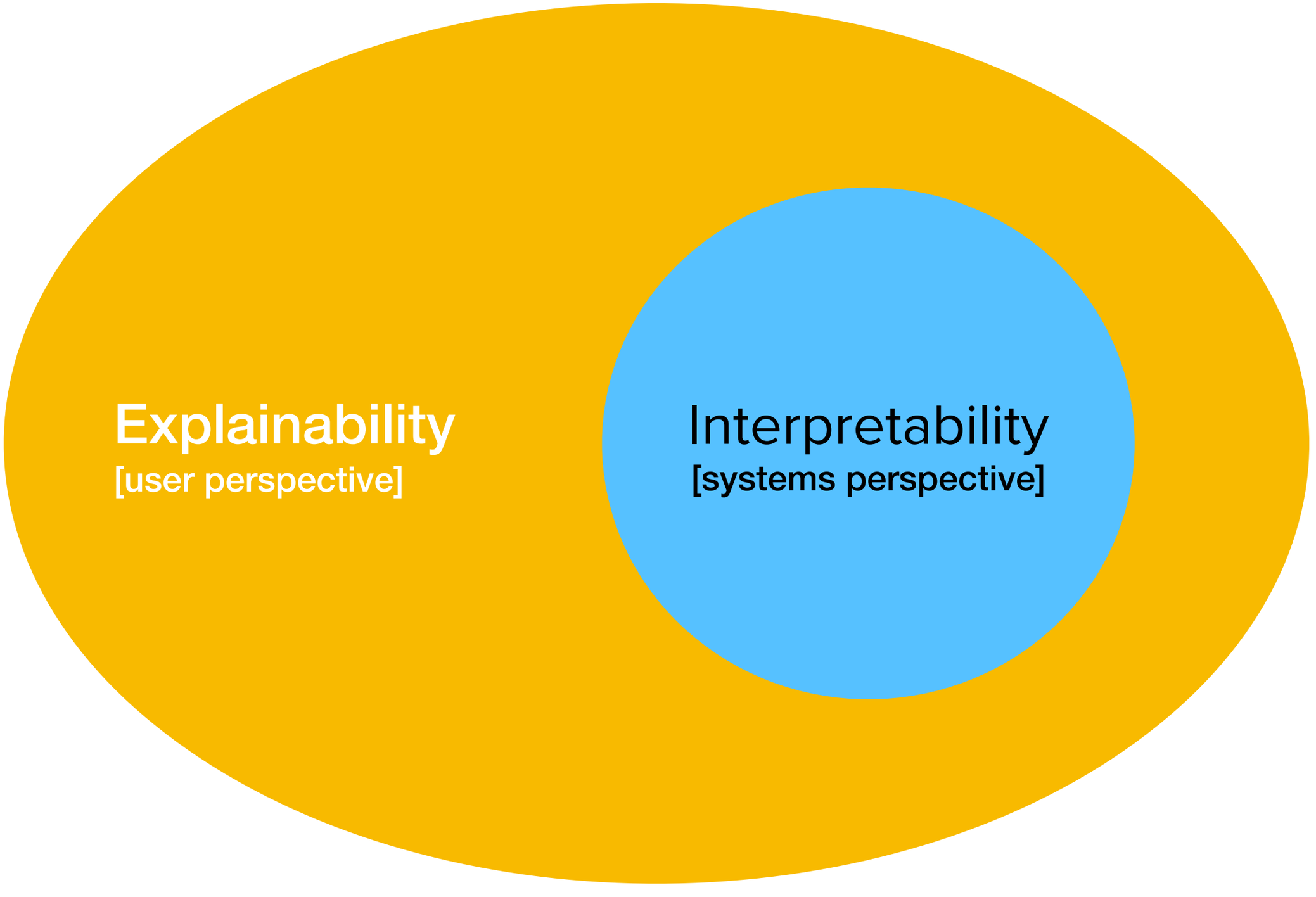
Explainable methods aim to produce human-understandable explanations in natural language or other interpretable forms. [users perspective]

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

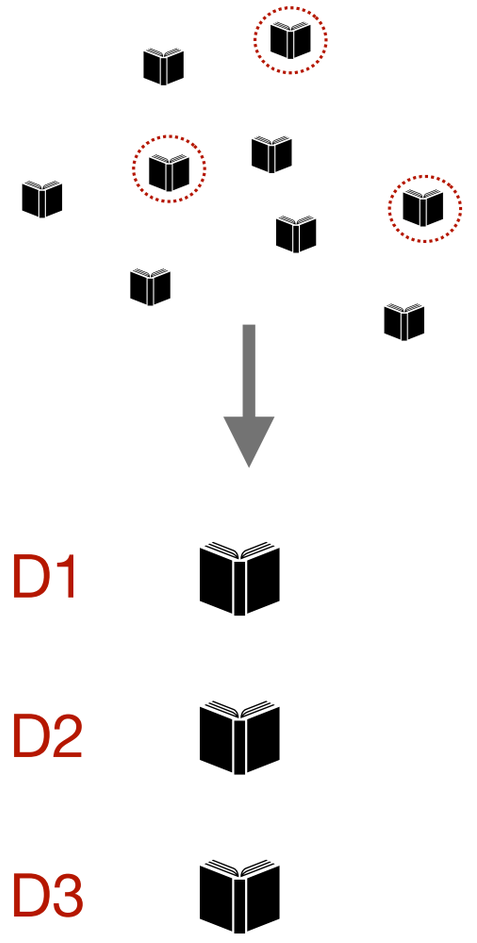
The scope of explanations for ranking tasks

How to find the Mean ? Search

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

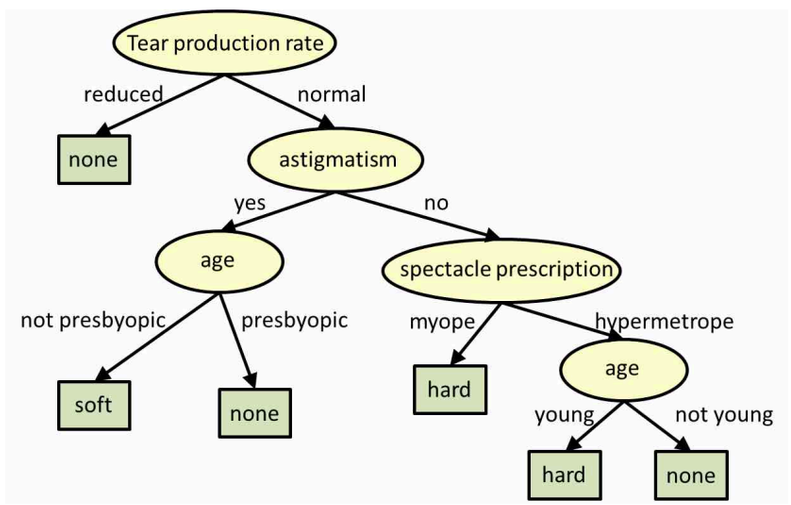
Approach Families

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

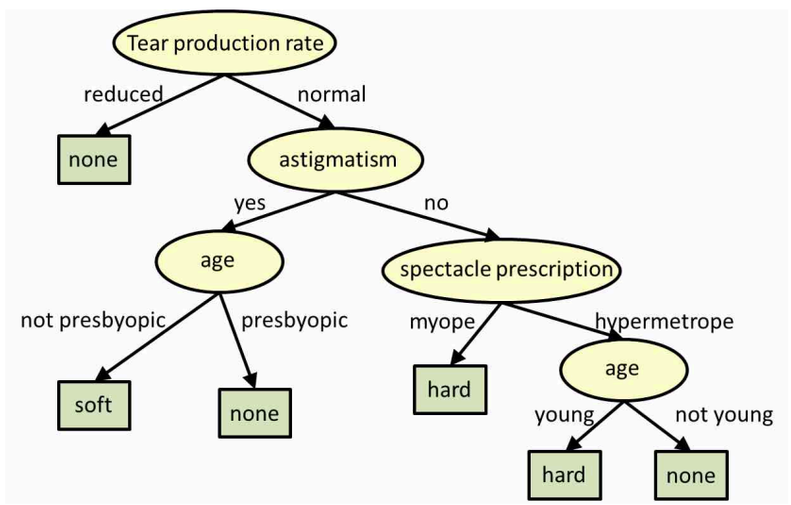
Interpretable algorithms for Knowledge Intensive Tasks

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 ?

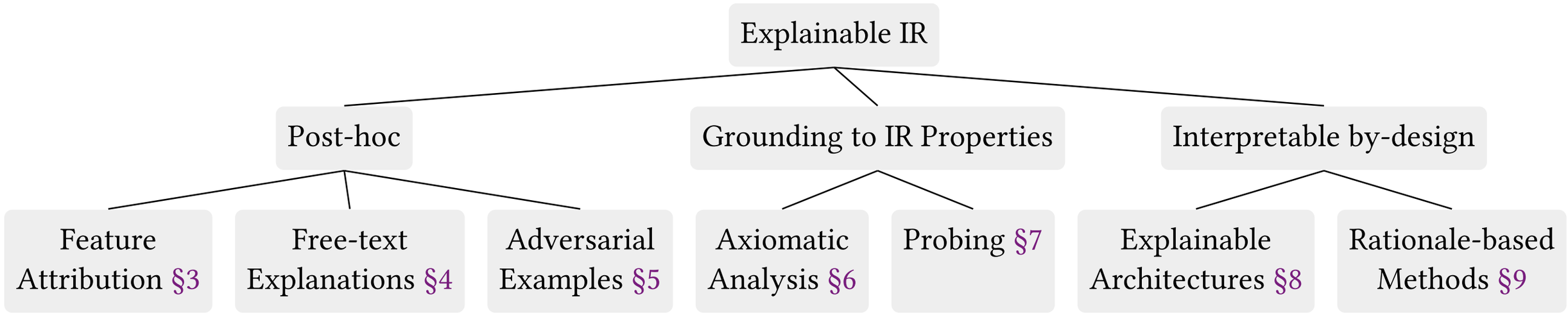
Does Ranking models understand matching, BM25, entity matching ?

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