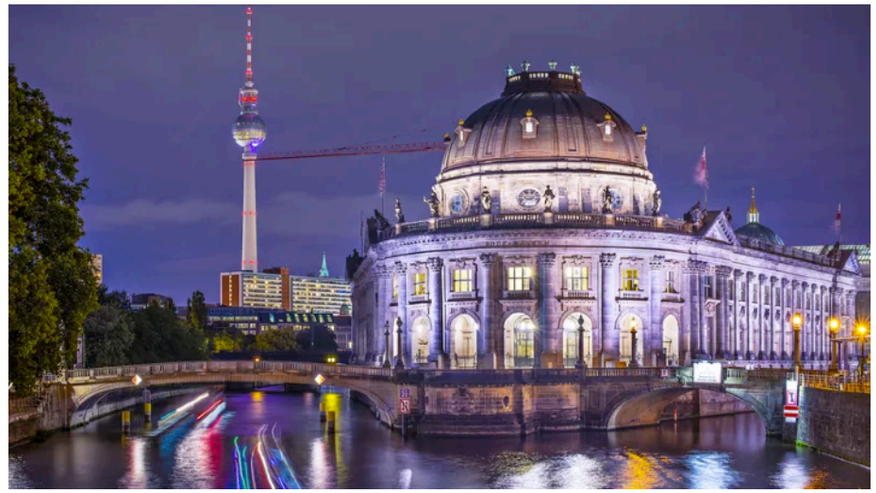
XAI for IR

Jaspreet Singh

Applied Scientist, Search Science & AI
Amazon

About Me

- Research Background:
  - IR, Deep Learning, XAI
- I live in Berlin
- I work on:
  - Semantic Search
  - Productizing LLMs
  - Representation Learning for Amazon Entities

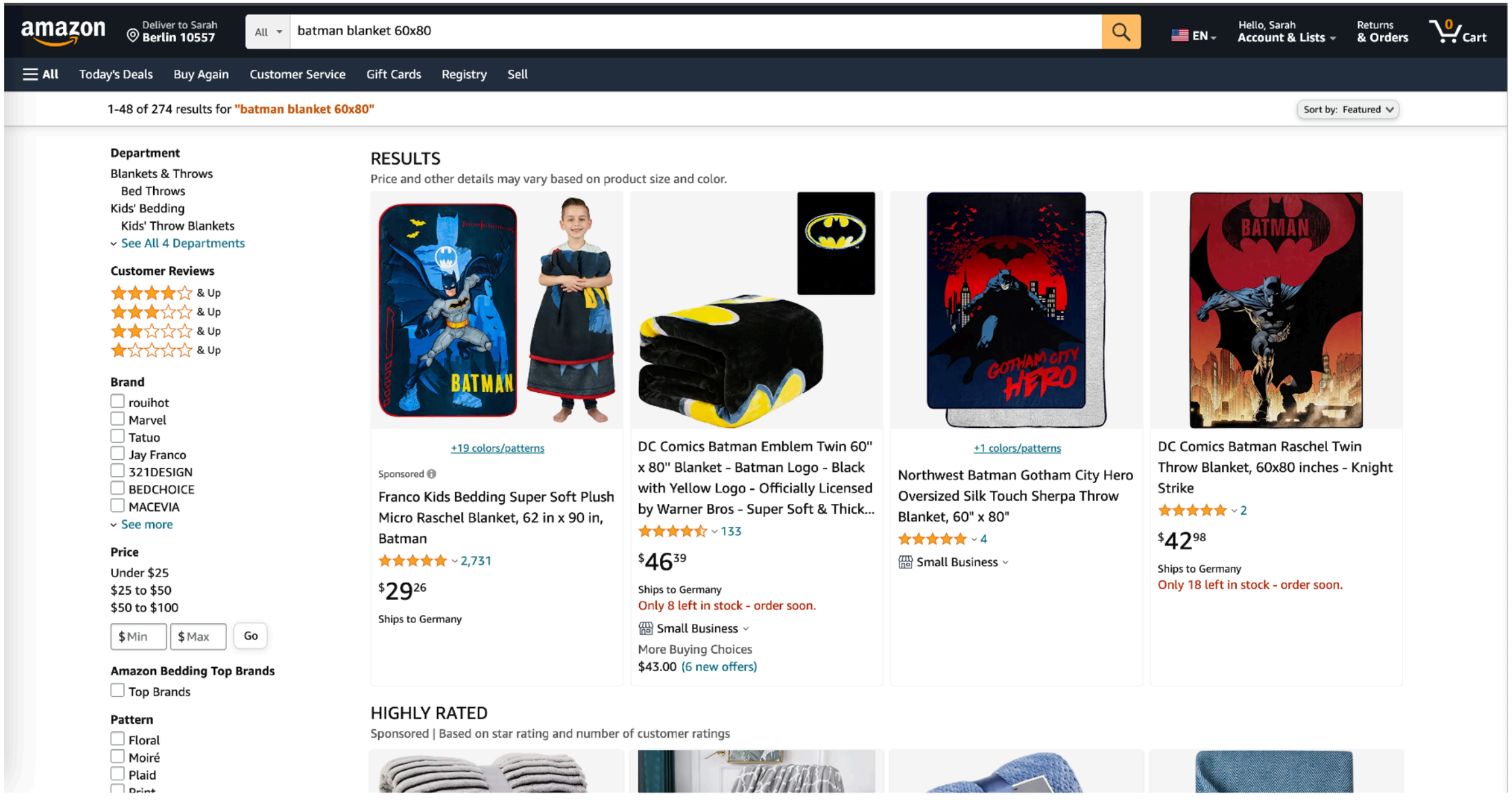


Outline

- ML in Search [10 min]
- Global Interpretability of ranking models [5 min]
- Local Model Agnostic Interpretability of ranking models [30 min]
- Model-specific interpretability of Neural Rankers [30 min]
- Code Demo! [5 min]

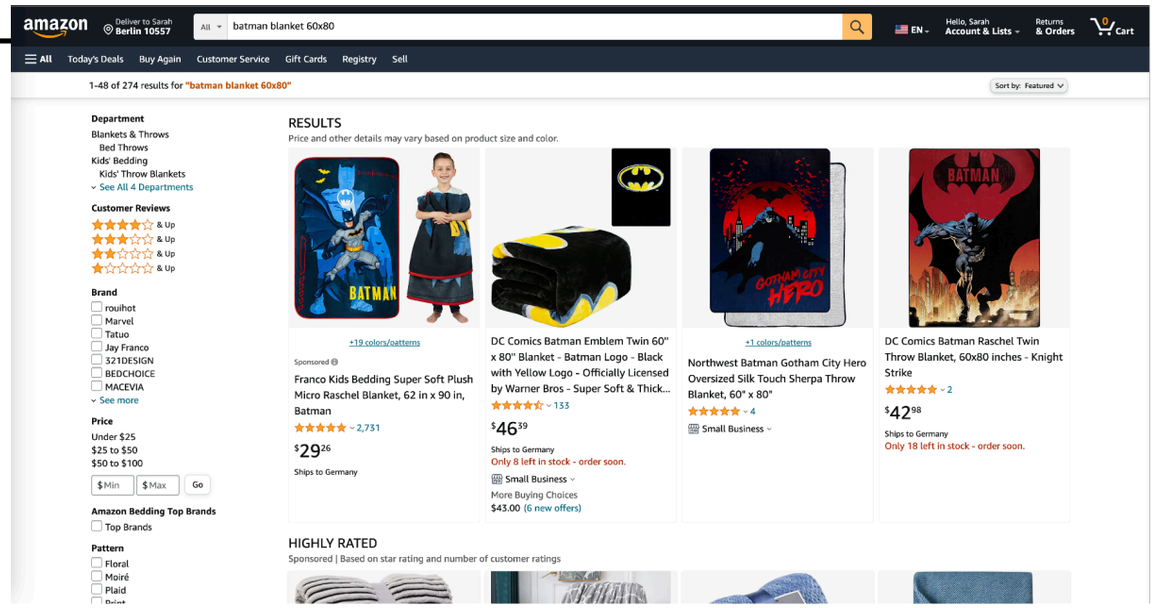
ML for Search

Anatomy of a Search Engine

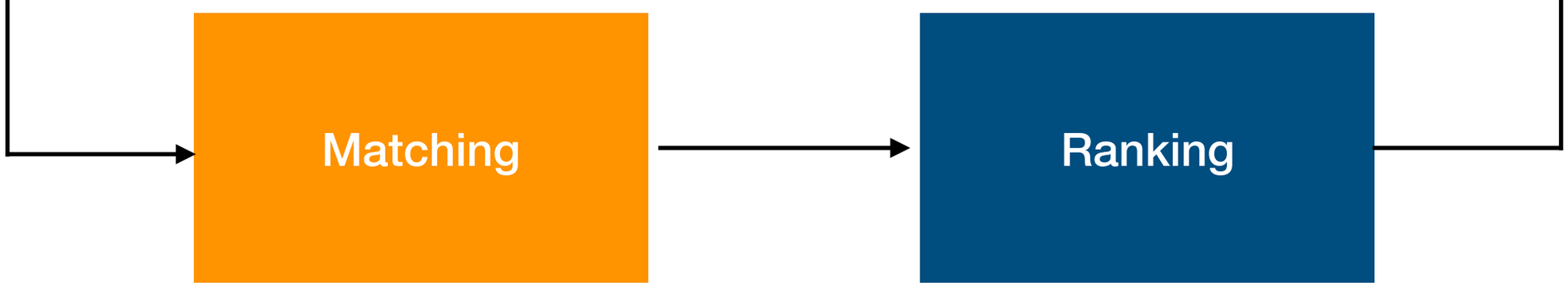


Anatomy of a Search Engine

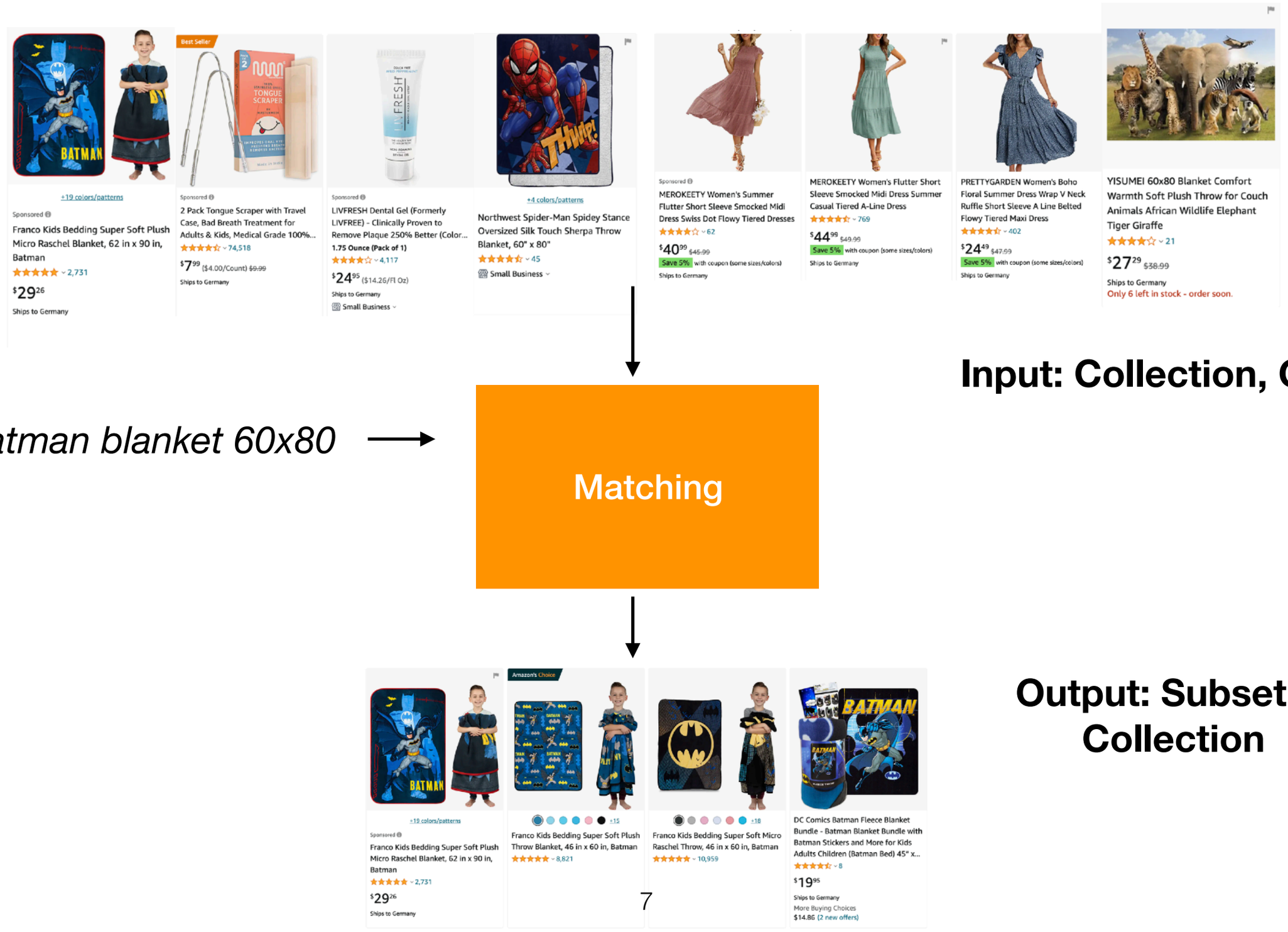
batman blanket 60x80



Most <u>Relevant</u> batman blankets



Matching



batman blanket 60x80 →

Matching

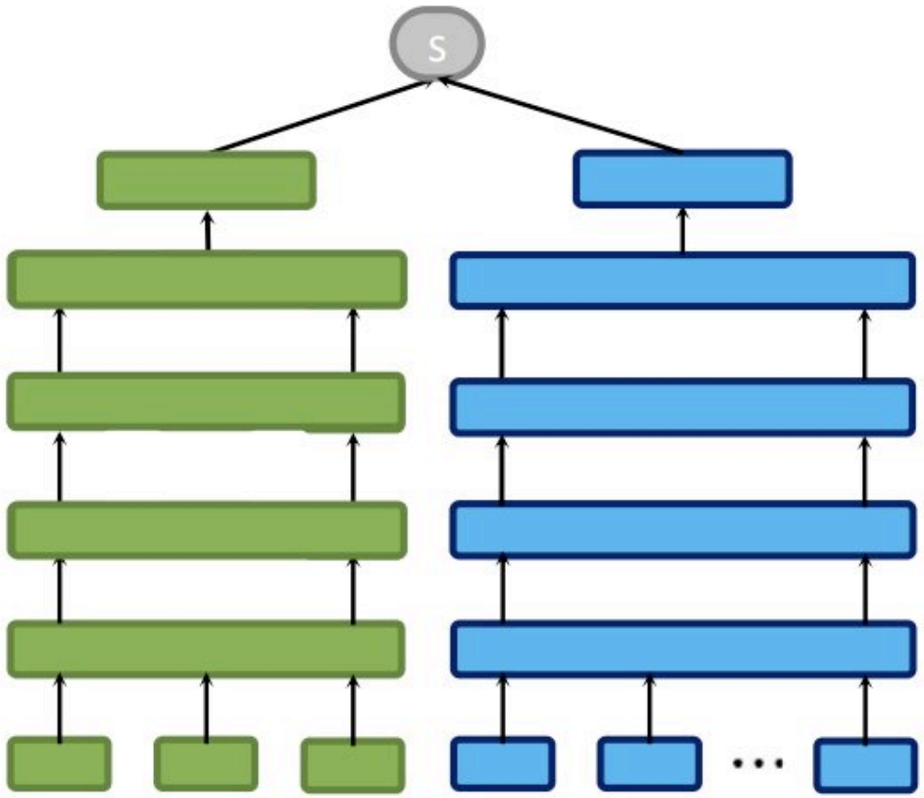
Input: Collection, Query

Output: Subset of Collection

Matching

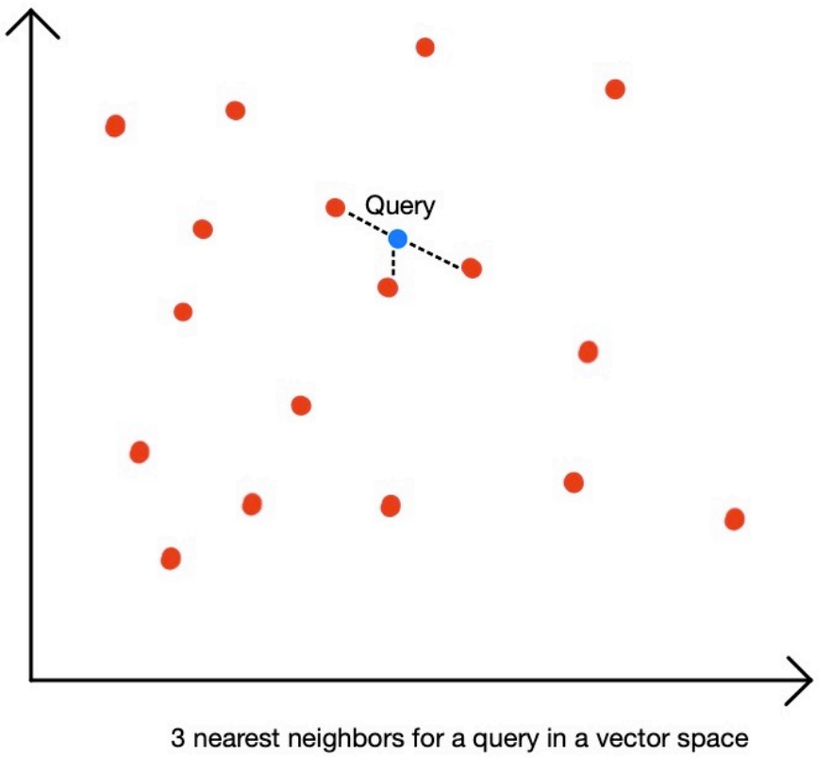
- ML to address the vocabulary mismatch problem
- Standard Inverted Indexes are lexical matchers
- Deep learning models are used for semantic matching, query expansion, document expansion
- Improve recall

Two Tower Matching Models



(a) Representation-based Similarity

(e.g., DSSM, SNRM)



3 nearest neighbors for a query in a vector space

https://www.pinecone.io/learn/what-is-similarity-search/

batman blanket 60x80

Why?

YISUMEI 60x80 Blanket Comfort Warmth Soft Plush Throw for Couch Animals African Wildlife Elephant Tiger Giraffe

4.4 out of 5 stars 21

$27.29 ~~$38.99~~

Ships to Germany

Only 6 left in stock - order soon.

+19 colors/patterns

Sponsored

Franco Kids Bedding Super Soft Plush Micro Raschel Blanket, 62 in x 90 in, Batman

4.7 out of 5 stars 2,731

$29.26

Ships to Germany

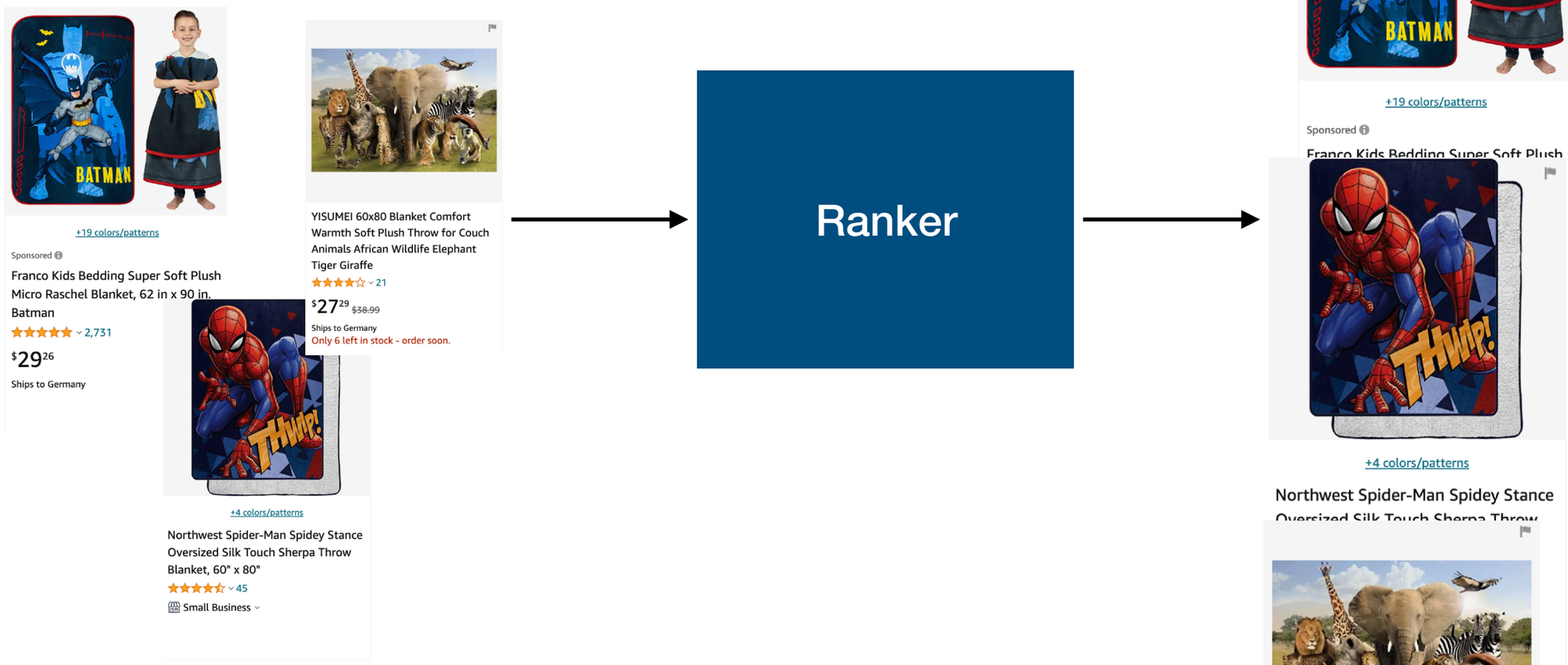
+4 colors/patterns

Northwest Spider-Man Spidey Stance Oversized Silk Touch Sherpa Throw Blanket, 60" x 80"

4.5 out of 5 stars 45

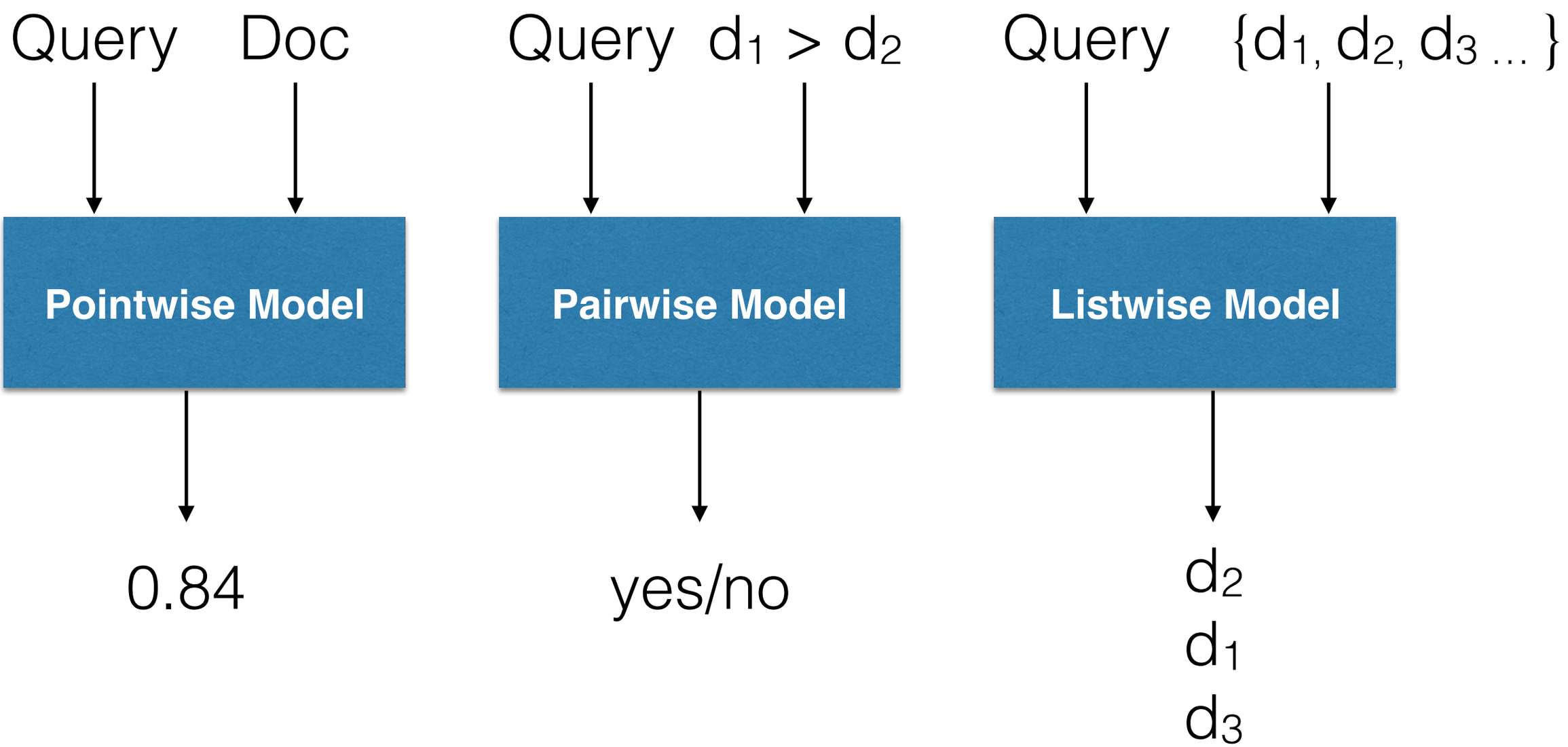
Small Business

Ranking Models

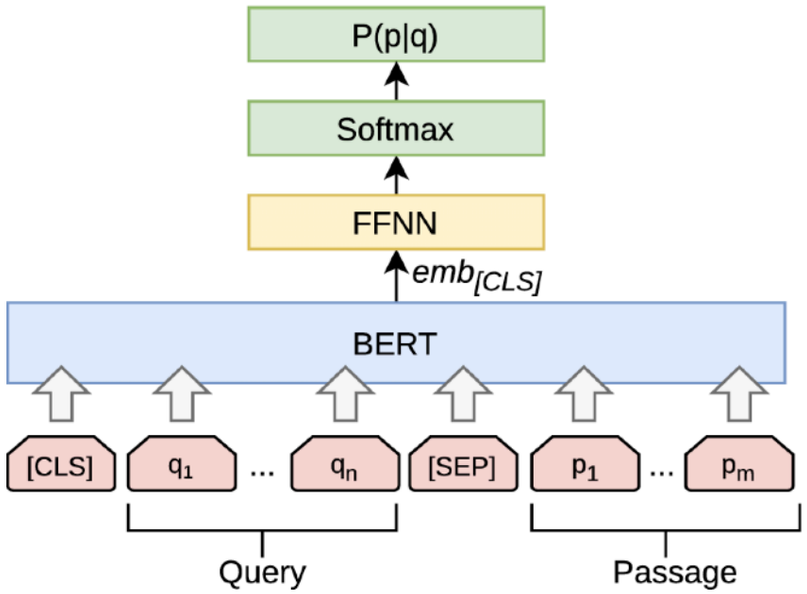


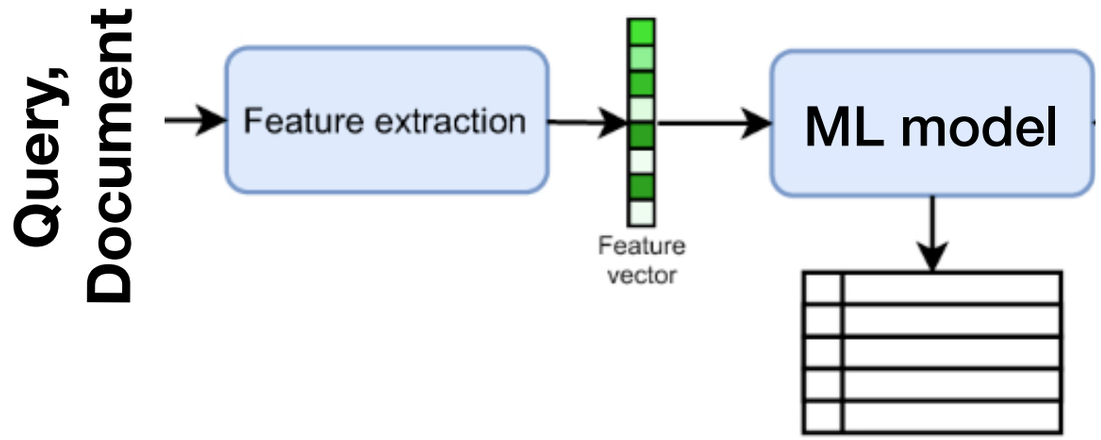
Input: Set of Items

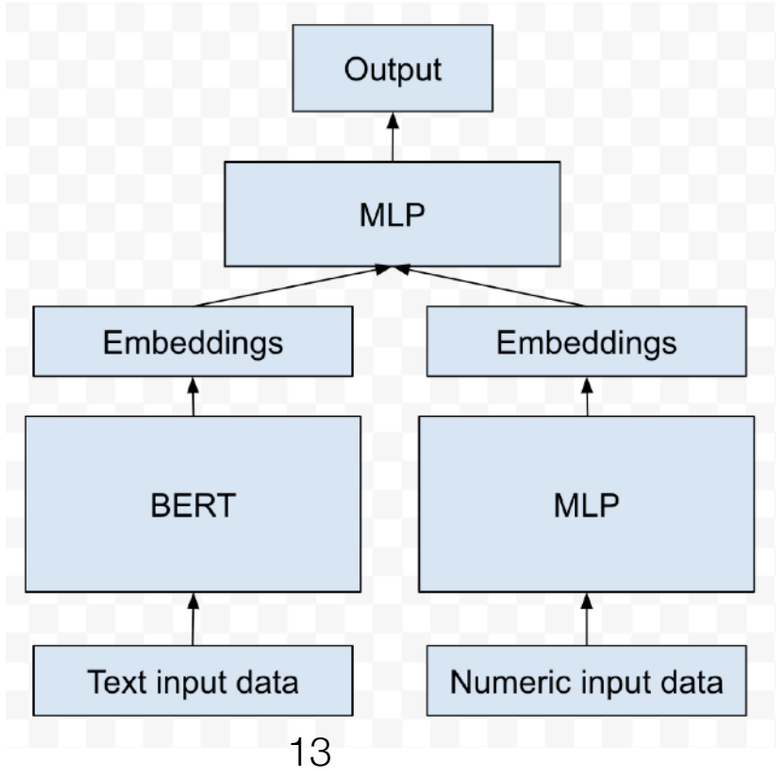
Output: Ordered List of Items


Query
Doc
Query
$d_1 > d_2$
Query
$\{d_1, d_2, d_3 \ldots\}$
Pointwise Model
Pairwise Model
Listwise Model
0.84
yes/no
d_2
d_1
d_3

Ranking Models







batman blanket 60x80

Why?

+19 colors/patterns

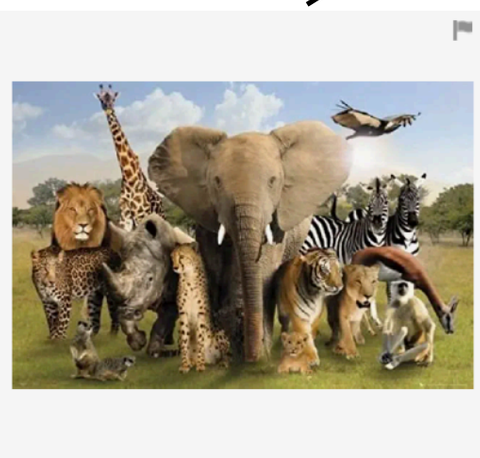
Sponsored

Franco Kids Bedding Super Soft Plush Micro Raschel Blanket, 62 in x 90 in, Batman

★★★★★ 2,731

$29.26

Ships to Germany



YISUMEI 60x80 Blanket Comfort Warmth Soft Plush Throw for Couch Animals African Wildlife Elephant Tiger Giraffe

★★★★☆ 21

$27.29 $38.99

Ships to Germany

Only 6 left in stock - order soon.

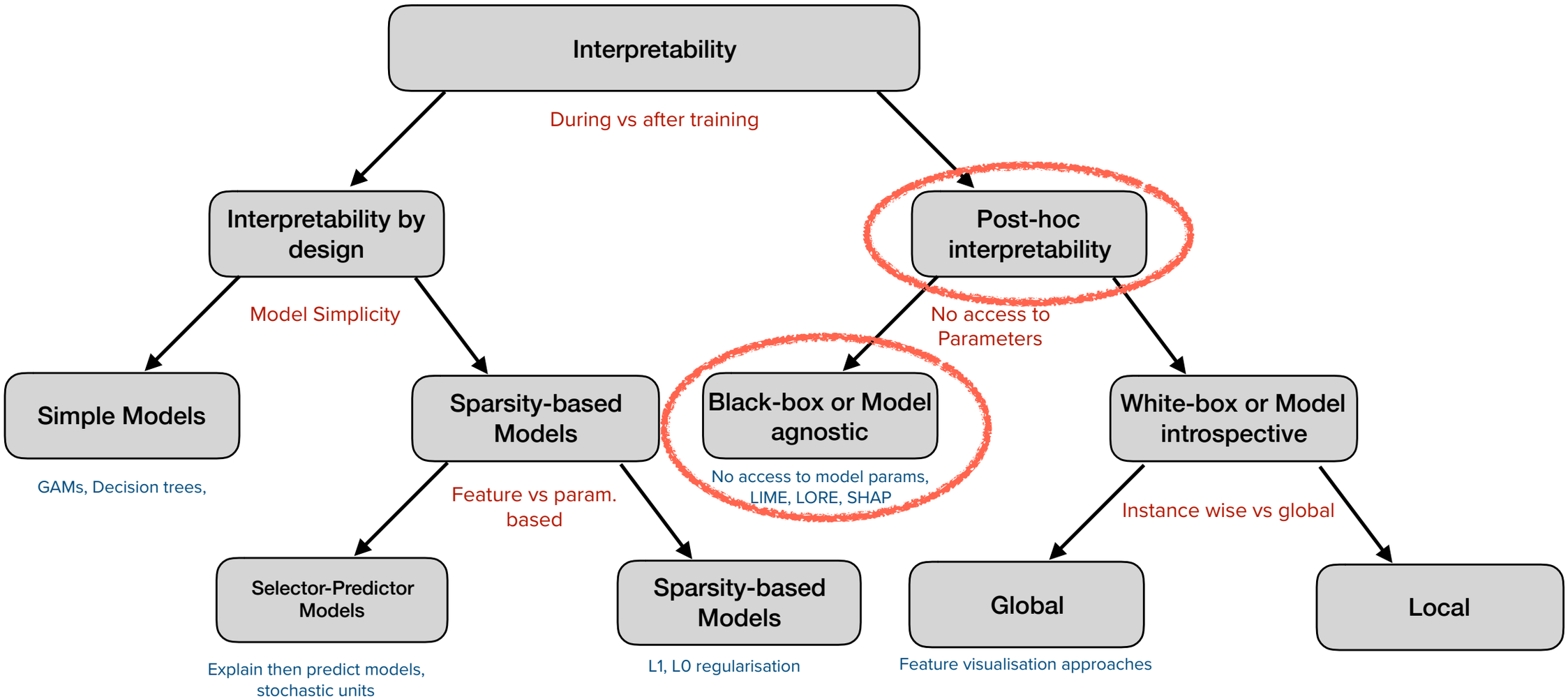
+4 colors/patterns

Northwest Spider-Man Spidey Stance Oversized Silk Touch Sherpa Throw Blanket, 60" x 80"

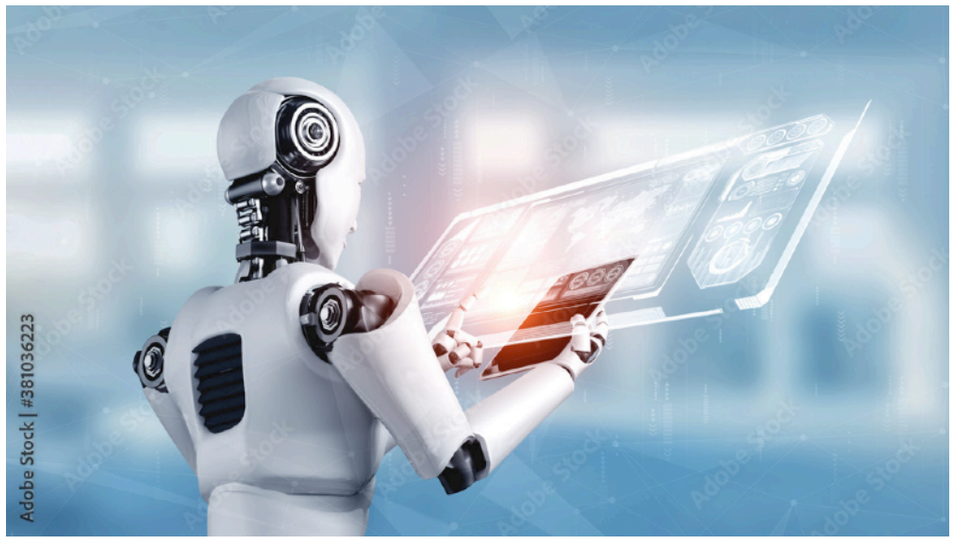
★★★★½ 45

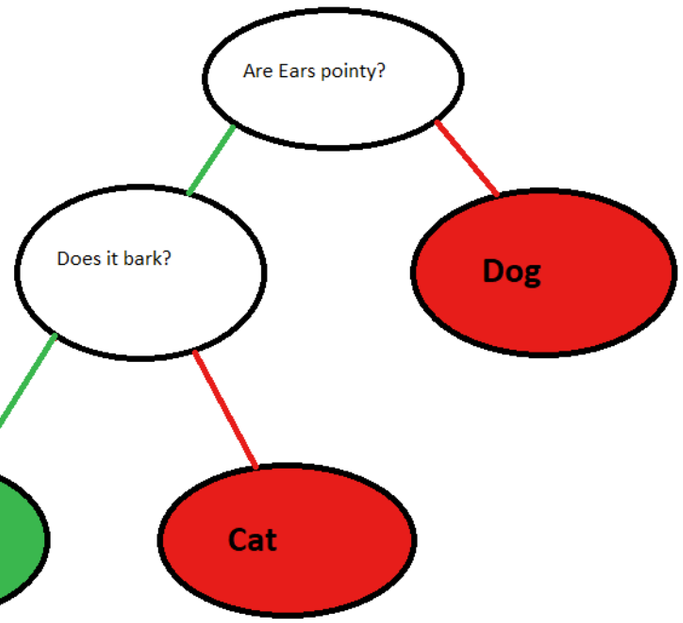
Small Business

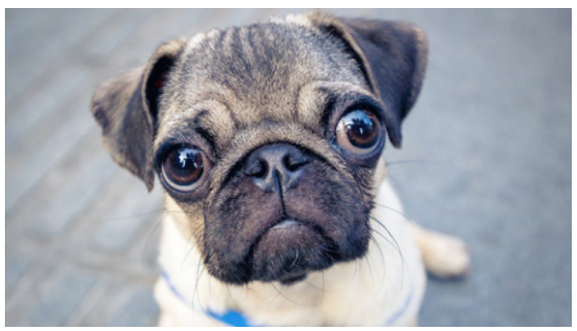
XAI for ~~Search~~ Ranking Models

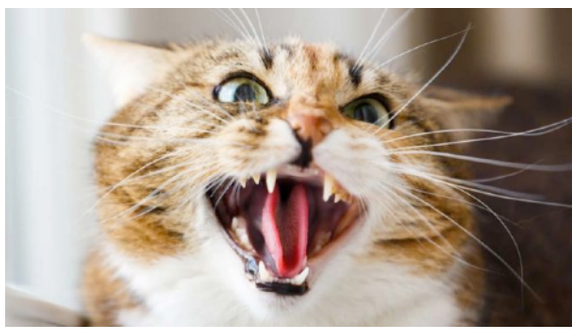

Interpretability
During vs after training
Interpretability by design
Post-hoc interpretability
Model Simplicity
No access to Parameters
Simple Models
GAMs, Decision trees,
Sparsity-based Models
Black-box or Model agnostic
No access to model params, LIME, LORE, SHAP
White-box or Model introspective
Feature vs param. based
Instance wise vs global
Selector-Predictor Models
Explain then predict models, stochastic units
Sparsity-based Models
L1, L0 regularisation
Global
Feature visualisation approaches
Local

Global Model Agnostic Explanations









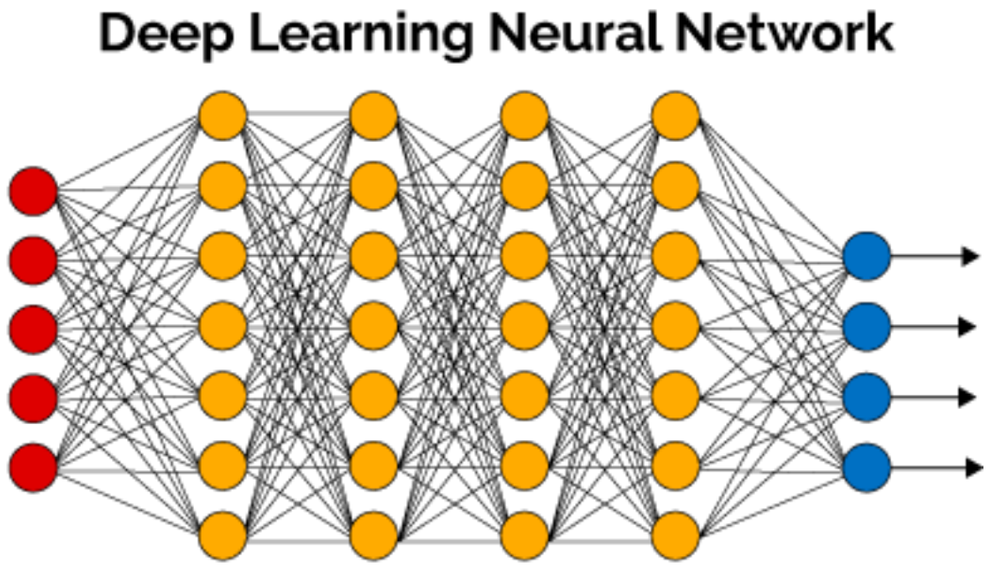
Mimic Models

- Generate infinite amounts of “secondary” training data
- Train a simpler model on a simpler feature space
  - Similar to LIME
- We are essentially distilling a smaller more interpretable model from a larger model

Posthoc interpretability of learning to rank models using secondary training data [SIGIR 2018]

- Mimic a more complex LtR model with a simple tree based model that operates on an interpretable feature subset

- How far can we mimic a ranker with such strict limitations?



Trained Ranker

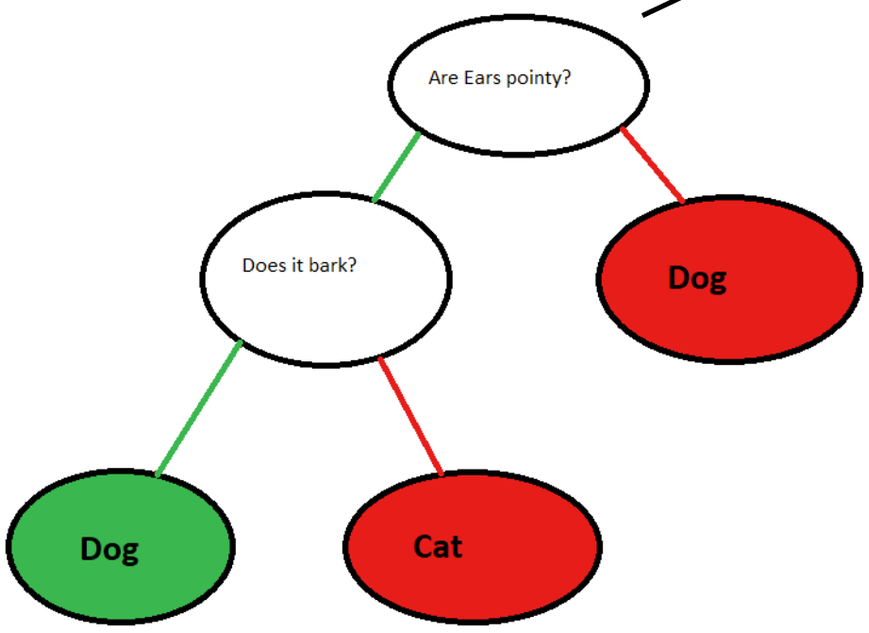
1. Create secondary training data

Training corpus

q_1	q_2	q_3	
d_4	d_6	d_9	.....
d_3	d_3	d_4	
d_1	d_5	d_5	
d_2	d_2	d_1	

2. Select interpretable features (manual) — 24/132

3. Train tree model



Results

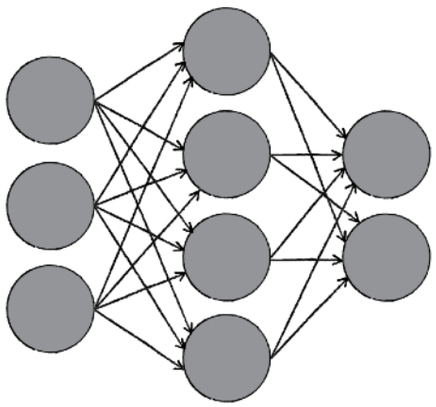
Training Size	All Features (AM)				Interpretable Features (IM)			
	NDCG@10	Prec.@10	τ	τ@10	NDCG@10	Prec.@10	τ	τ@10
100	0.3280	0.574	0.8664	0.7440	0.2819	0.5482	0.4363	0.2004
200	0.3397	0.5415	0.7328	0.6592	0.2837	0.5507	0.4442	0.2692
300	0.3373	0.5932	0.7290	0.6736	0.2859	0.5535	0.4400	0.2810
400	0.3396	0.594	0.7488	0.7108	0.2849	0.5522	0.4632	0.3394
500	0.3398	0.5945	0.7197	0.7183	0.2870	0.5535	0.3987	0.1806
1K	0.3397	0.5947	0.6960	0.6960	0.2893	0.5593	0.4341	0.2483
2.5K	0.3401	0.5941	0.6394	0.6489	0.2886	0.5593	0.4088	0.1392
5K	0.3396	0.5928	0.7188	0.6709	0.2879	0.5574	0.4216	0.1434
7.5K	0.3417	0.5948	0.7390	0.7144	0.2877	0.5572	0.4127	0.2206
15K	0.3405	0.5936	0.7420	0.7173	0.2881	0.5571	0.4275	0.2180
M-P	0.3430	0.5569	NA	NA	0.3430	0.5569	NA	NA

Table 1: Models trained on increasing secondary training data from M-P

Local Model Agnostic Explanations

- Interpret a **single decision** made by a complex **ranking model**

Query **q**



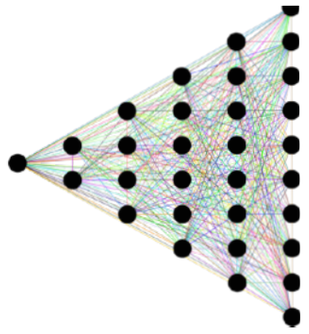
Blackbox ranker

→

d_1
d_3
d_5
d_2
d_4

- Why is a document relevant to the query?
- Why is a document d ranked higher than another for the query?
- **What is the intent of the query according to the ranker?**

This talk is great! I need a **quick introduction** to learning to rank so that i can understand this better!



learning to rank

[PDF] A Short Introduction to Learning to Rank
times.cs.uiuc.edu/course/598f16/l2r.pdf
by H LI - 2011 - Cited by 191 - Related articles
Benchmark data sets on **learning to rank** have also been released [4]. The evaluation on the performance of a **ranking** model is carried out by comparison between the **ranking** lists output by the model and the **ranking** lists given as the ground truth. Several evaluation measures are widely used in IR and other fields.

Learning to Rank 101 – Pere Urbon-Bayes – Medium
https://medium.com/@purbon/learning-to-rank-101-5755f2797a3a
Dec 9, 2017 - As we saw earlier LTR or **Learning to Rank** is introduced to improve the accuracy of your search by using Machine **Learning**. The ML model will be applied after the initial **ranking** done by the search engine, however ML models can be pretty expensive to run so we will only rescore using the model the top N results.

Learning to Rank: A Key Information Retrieval Tool for Machine ...
https://thenewstack.io/letor-machine-learning-web-search-technique-thats-turned-key-i...
Jan 9, 2018 - The **Learning To Rank** (LETOR or **LTR**) machine learning algorithms — pioneered first by Yahoo and then Microsoft Research for Bing — are ...

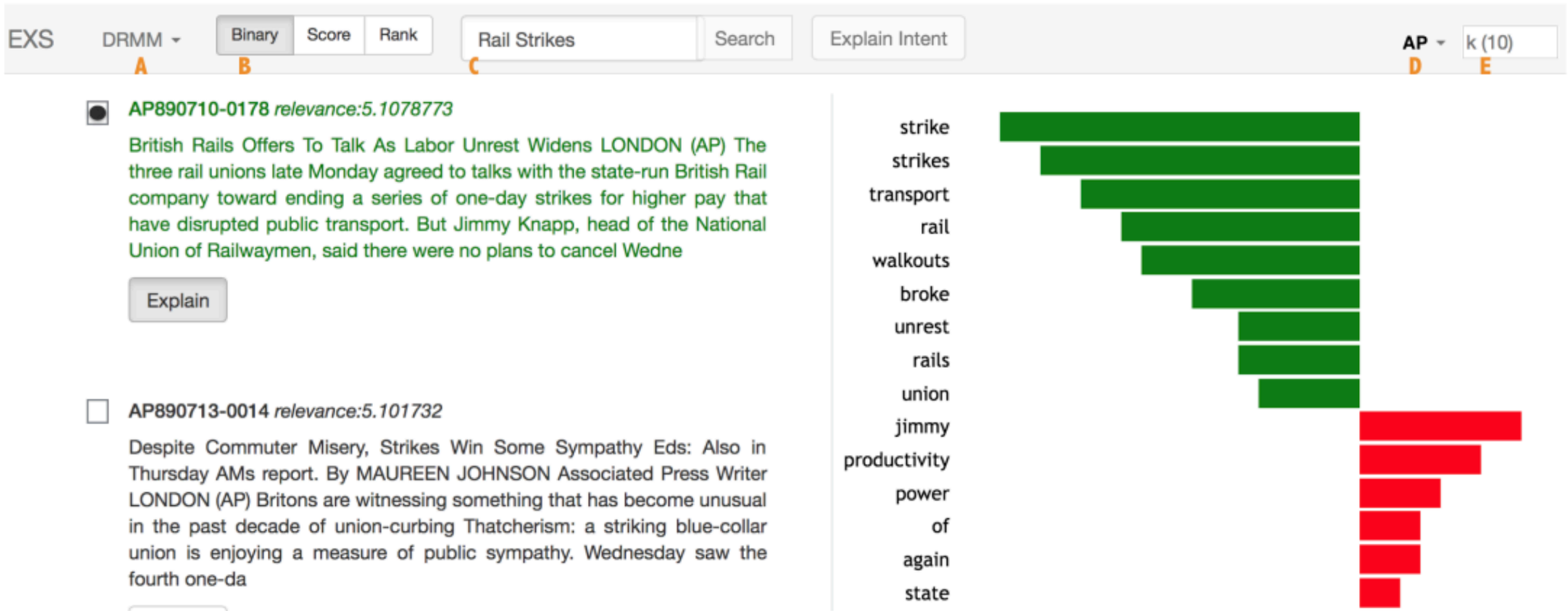
Explaining Rankings

- Documents are ranked according to relevance
- Relevance is computed against a perceived **information need / intent** encoded by a query
- To understand differences in relevance we must first understand the intent of the ranking model for a given query

Let’s try LIME?

- Cast a ranking model as a classifier
  - Pointwise model —> Next Slide [Most commonly used]
  - Pairwise model —> already a binary classifier but can you explain a whole list?
  - List-wise model —> create pairwise training data from lists to treat it as a ranking model? Is this enough?

EXS: Explainable search using local model agnostic interpretability [WSDM'19]



- Using LIME to explain point wise ranking models
  - What is the best way to convert point wise scores to interpret the ranked list?
  - Given a perturbed document document, how should we assign labels to it?

Top-k binary	**Score based**	**Rank based**
P(X=relevant | q, d', R) =1	$1 - \frac{\mathcal{R}(q,d_1) - \mathcal{R}(q,d')}{\mathcal{R}(q,d_1)}$	$1 - \frac{rank(d')}{k}$
If R(q,d') > R(q,d_k)	P(X=relevant) =1	P(X=relevant) = 0
Else 0	If R(q,d') >= R(q,d_1)	If R(q,d') <= R(q,d_k)

- Each method produces slightly different results. How do we know which one is better? More on this later

- Does EXS actually help explain rankings?

  - *Why is a document relevant to the query?* Yes we can get explanations similar to LIME with word attributions

  - *Why is a document d ranked higher than another for the query?* We can adjust the value of k instead of assigning it arbitrarily

- *What is the intent of the query according to the ranker?* Sort of. Aggregate individual predictions

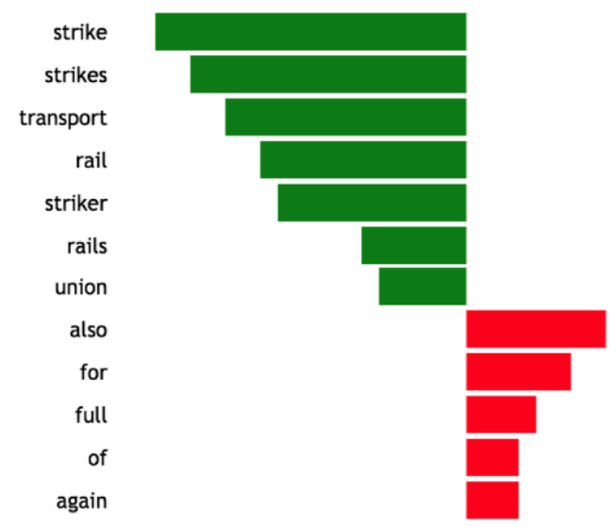


Figure 2: Intent explanation for the query 'Rail Strikes' when using DRMM to rank documents from a news collection.

LIRME [SIGIR'19]

- Locally Interpretable Ranking Model Explanation

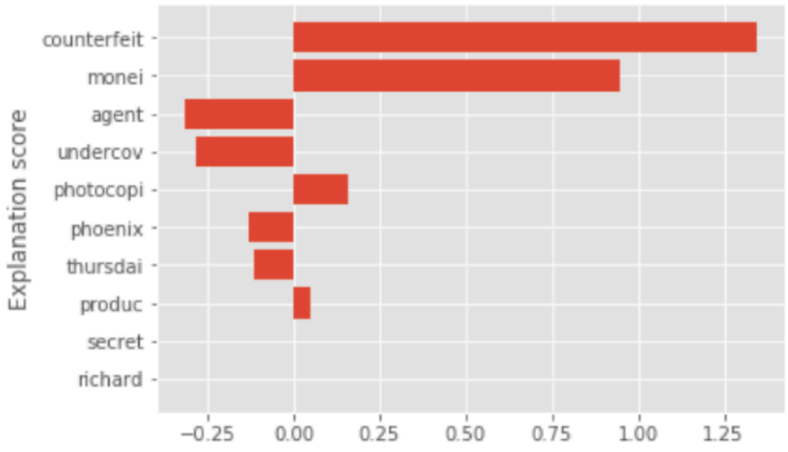
$$\mathcal{L}(D, Q, \sigma; \Theta) = \sum_{i=1}^{M} \rho(D, D'_i) \left(S(D, Q) - S_\Theta(D'_i, Q)\right)^2 + \alpha|\Theta|$$

Surrogate

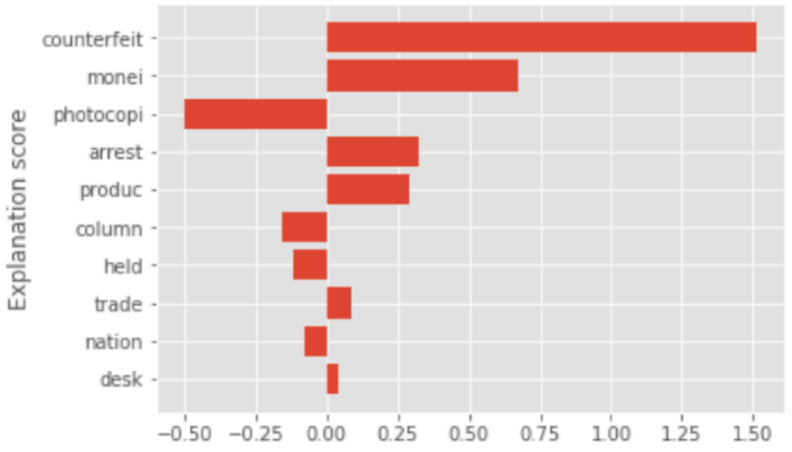
$$= \sum_{i=1}^{M} \rho(D, D'_i) \left(S(D, Q) - \sum_{j=1}^{p} \theta_j w(t_j, D'_i)\right)^2 + \alpha|\Theta|.$$

$$\rho(D, D') = \exp(-\frac{x^2}{h}), \;\; x = \arccos(D, D')$$

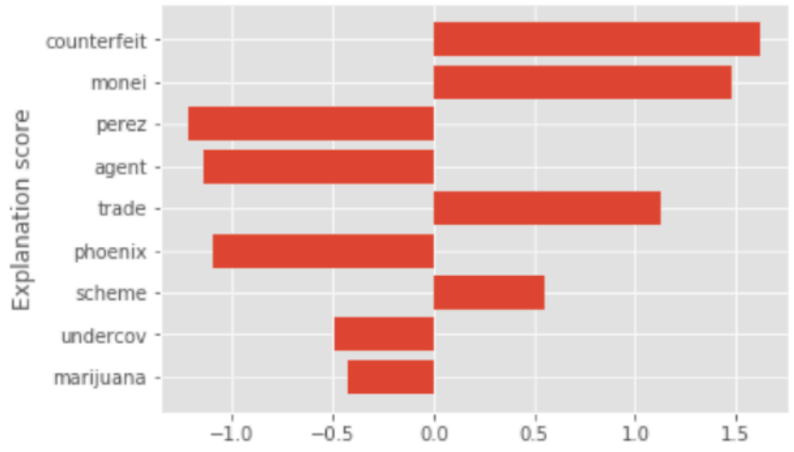
- They explored different sampling techniques for D'. Can you think of some good techniques that are better suited to rankers?



(a) Uniform sampling



(b) Tf-idf sampling

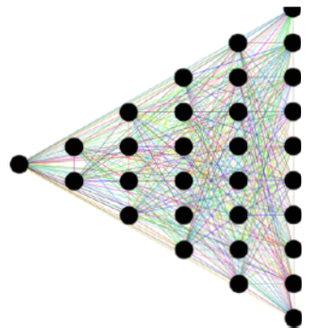


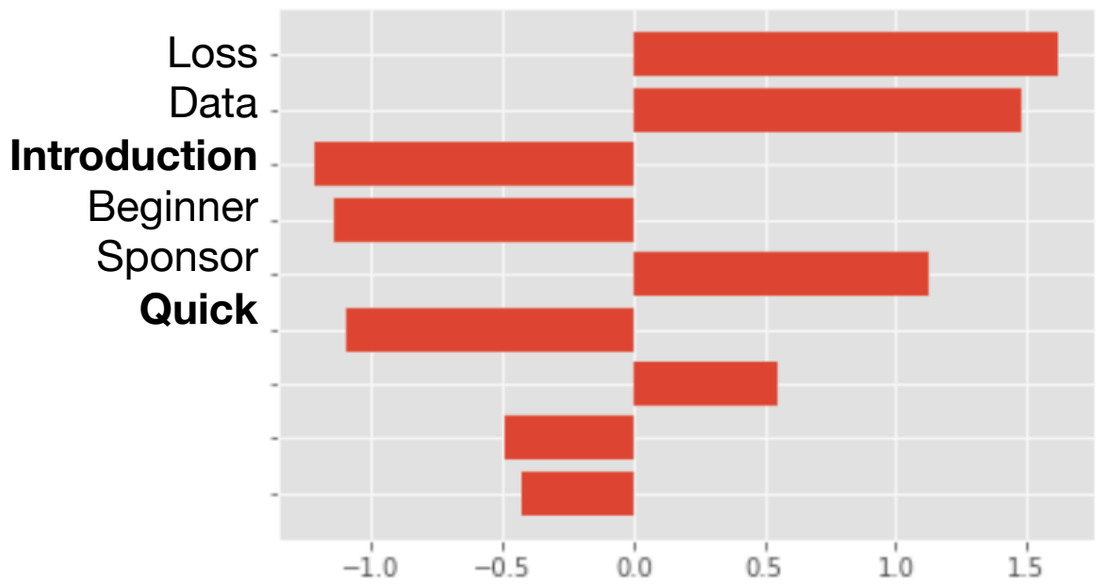
(c) Masked samples ($v = 0.1$, $k = 5$)

Evaluation

- How do you know if the explanation is correct?
- Is the explanation faithful to the underlying model?
  - We cannot really answer this for a black box model. The loss value gives you a hint but it depends on the neighbourhood.
- We can measure the “correctness” of the explanation though.
  - How close is the explanation to the *user intent*?

This talk is great! I need a **quick introduction** to learning to rank so that i can understand this better!





(Local) Model agnostic interpretability of rankers via intent modelling [FAccT'20]

- Can we devise a method to directly explain the intent of text rankers?

- Use the same intuition as other model agnostic local interpretability techniques

- Select a simple feature space —> words

- *Train a simple **RANKING** model with labels from the complex **RANKING** model on perturbed instances that is faithful to the complex model's decision boundary*

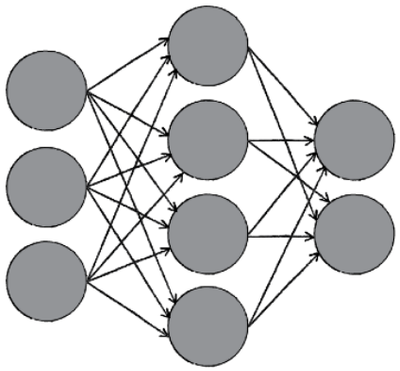
Simple Ranking Models as Explanations

In our view of intent based interpretability, an explanation for a query-ranking consists of:

Terms that encode the intent

Simple **scoring function** that uses intent terms to estimate document relevance: TF, IDF, position, proximity, semantic similarity…

Query **q**



Blackbox ranker

↓

d_1
d_3
d_5
d_2
d_4

Intent terms are used as query expansions

Query **q**

Intent Terms E_q

+

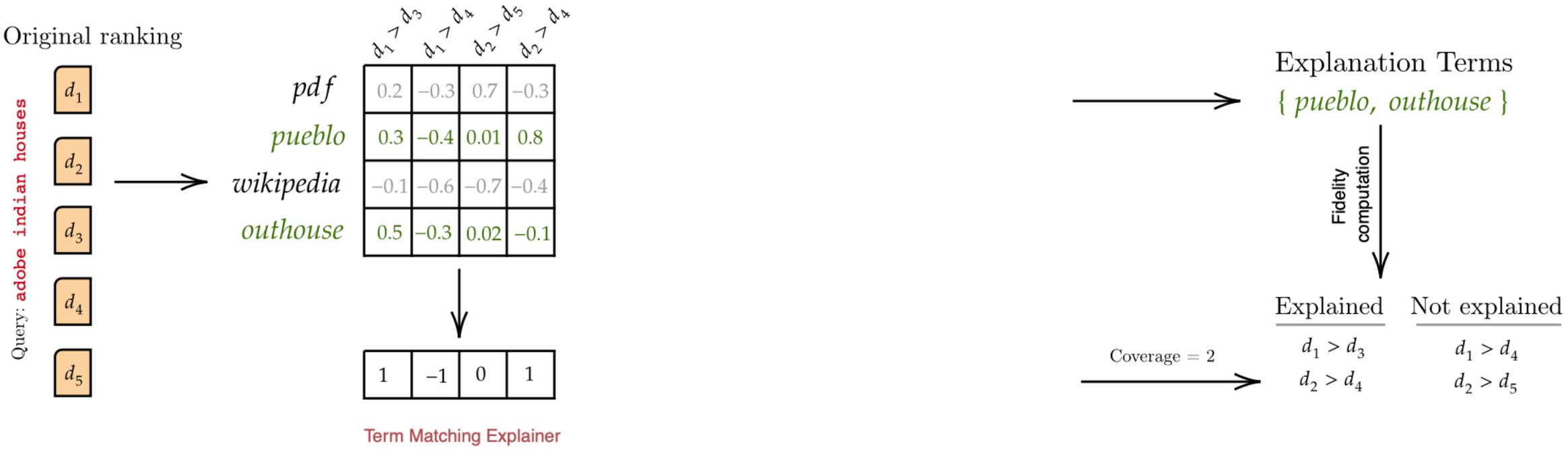
Simple scoring function R_E

↓

Rank order		
	d_1	d_1
	d_3	d_3
	d_5	d_5
	d_2	d_4
	d_4	d_2

Ranking: Aggregation of preference pairs

We find the explanation model that maximises the coverage of preference pairs in the ranking



Preference Matrix

$$w_{12} (R_E(w,d_1) - R_E(w,d_2))$$

Preference Pair sampling

health hazards

d1

d2

d3

d4

d5

Preference Pairs or Features

	d1 > d2	d1 > d3	d2 > d3	d1 > d4	d2 > d5
le	-0.2	-0.8	-0.12	-0.34	0.98
handle	-0.1	-0.34	-0.24	-0.0001	-0.7
doctor	-0.11	0.34	0.1	-0.223	-0.34
invert	-0.45	0.04	-0.67	-0.23	-0.003
medicin	-0.34	-0.31	0.5	0.8	0.01
Sum	-0.45	0.03	0.6	0.577	-0.33

Selecting Intent Terms

- Optimisation: Preference Coverage
- Select the minimum number of terms that maximises the coverage of preference pairs
- The PC framework using a single Ψ aims to choose a subset of rows $E \subseteq X$ (equivalent to selecting terms) from M so as to maximize the number of non-zero values in the aggregated vector.

Integer Linear Program formulation

$$maximize \sum_{i=1}^{m} \Big(\mathtt{sign}(\mathbf{x}^\top \mathbf{M}) \Big)_i \qquad \text{(PC)}$$

$$\text{s.t.} \;\; \mathbf{x} = [x_1, \cdots, x_n]; \;\; x_i \in \{0, 1\}$$

NP hard
Non-convex!
Greedy selection
(no guarantees)

	$d_1^\top d_3$	$d_1^\top d_4$	$d_2^\top d_5$	$d_2^\top d_4$
pdf	0.2	−0.3	0.7	−0.3
pueblo	0.3	−0.4	0.01	0.8
wikipedia	−0.1	−0.6	−0.7	−0.4
outhouse	0.5	−0.3	0.02	−0.1

$$minimize \quad \left(-\sum_{i=1}^{m} (\mathbf{tanh}(\mathbf{v}))_i + \|\mathbf{x}\| \right) \quad \text{(GPC)}$$

$$\text{s.t.} \quad \mathbf{v} = \sum_{j=1}^{p} \mathbf{tanh}(\mathbf{x}^\top \mathbf{M}_j),$$

$$0 \le x_i \le 1, \quad a \le \sum_{i=1}^{m} x_i \le b$$

Tanh normalizes and smoothes!

	$d_1 > d_3$	$d_1 > d_4$	$d_2 > d_5$	$d_2 > d_4$
pdf	0.2	−0.3	0.7	−0.3
pueblo	0.3	−0.4	0.01	0.8
wikipedia	−0.1	−0.6	−0.7	−0.4
outhouse	0.5	−0.3	0.02	−0.1

$d_1 > d_3$	$d_1 > d_4$	$d_2 > d_5$	$d_2 > d_4$
9.3	−8.1	−5.7	−2.3
6.3	−4.1	−4.7	4.4
−1.5	−4.6	−7.7	−2.4
7.9	4.3	4.2	−4.1

$d_1 > d_3$	$d_1 > d_4$	$d_2 > d_5$	$d_2 > d_4$
0.8	0.2	0.5	−0.3
0.8	0.4	0.01	0.2
−0.1	0.5	0.2	−0.4
0	0.3	0.02	−0.1

p=3

2. Fidel Castro rumored to be dead

[clueweb09-en0003-14-04452] R_E score = 0.31

5. Fidel Castro opens new hospital in Havana

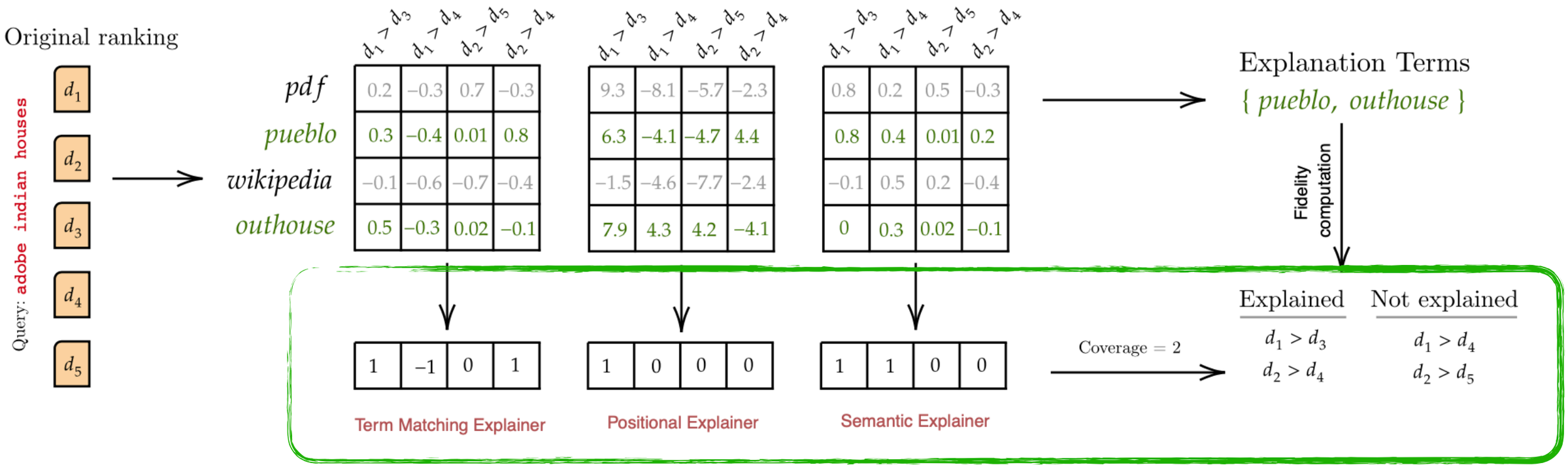
[clueweb09-en0002-76-22836] R_E score = 0.28

10. McCain discusses Cuba on Radio Show

[clueweb09-en0009-02-32574] R_E score = 0.24

Measuring Explanation Quality

- How do we know the explanation is good?
  - Accuracy — right for the right reasons
  - Fidelity — Faithfulness of the explanation model



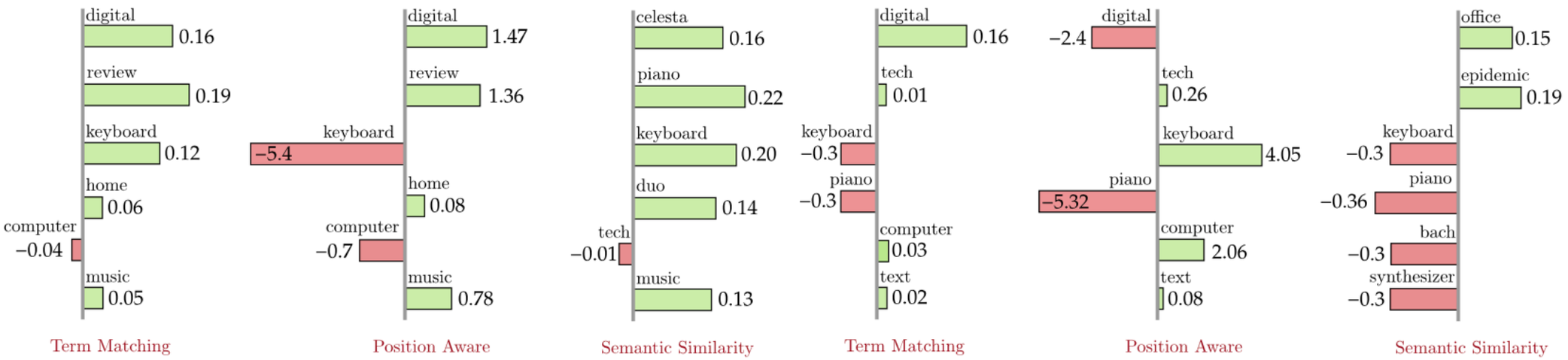
Coverage of preference pairs is directly optimised unlike EXS!

How do we know if the explanation is accurate though?

Results

		Clueweb09			Trec-DL		
Model	**Method**	*Fidelity*	*Fidelity*†	*Fidelity*‡	*Fidelity*	*Fidelity*†	*Fidelity*‡
BERT	QUERY-TERMS	0.81	0.88	0.76	0.81	0.82	0.63
	DEEPLIFT (Shrikumar et al., 2017)	0.77	0.81	0.67	0.70	0.75	0.62
	GREEDY-LM (Singh and Anand, 2020)	0.63	0.77	0.69	0.59	0.69	0.84
	MULTIPLEX	**0.88**	**0.97**	**0.93**	**0.86**	**0.93**	**0.97**
DPR	QUERY-TERMS	0.81	0.86	0.71	0.82	0.84	0.64
	DEEPLIFT (Shrikumar et al., 2017)	0.68	0.71	0.57	0.60	0.63	0.58
	GREEDY-LM (Singh and Anand, 2020)	0.61	0.68	0.88	0.63	0.70	0.75
	MULTIPLEX	**0.87**	**0.93**	**0.87**	**0.87**	**0.92**	**0.96**
DRMM	QUERY-TERMS	0.82	0.85	0.72	0.80	0.81	0.59
	DEEPLIFT (Shrikumar et al., 2017)	-	-	-	-	-	-
	GREEDY-LM (Singh and Anand, 2020)	0.57	0.60	0.72	0.53	0.54	0.34
	MULTIPLEX	**0.88**	**0.92**	**0.84**	**0.85**	**0.88**	**0.95**

Table 2: Fidelity values of all models on both Clueweb09 and Trec-DL datasets. *Fidelity*† refers to fidelity where preference pairs have at least a rank difference $\geq g$. *Fidelity*‡ only considers the preference pairs in the sampled set (500 pairs in our experiments). The best results are in bold.



(a) BERT: All explainers identify music-related terms as *pos*.

(b) DPR: All explainers identify music-related terms as *neg*.

Figure 3: Comparing explainers for the query: keyboard reviews, document pair: clueweb09-en0008-49-09140 (musical keyboard) *vs* clueweb09-en0010-56-37788 (technical keyboard). BERT prefers the former whereas DPR prefers the latter, resulting in opposite intents.

Query	Explainer	Explanation	Fidelity†
adobe indian houses	TEXT MATCHING	pdf, adobe, style, house, first, also	0.85
	POSITION AWARE	pdf, adobe, style, texas, wikipedia, 2009	0.81
	SEMANTIC SIMILARITY	pueblo, amarillo, castroville, outhouse, abourezk, alcove,	**0.95**
	MULTIPLE EXPLAINERS	pueblo, amarillo, castroville, outhouse, abourezk, pdf	0.91
espn sports	TEXT MATCHING	espn, abc, network, company, award, entertainment,	0.86
	POSITION AWARE	espn, sportscenter, abc, company, news, espn.com	**0.99**
	SEMANTIC SIMILARITY	espn, sportscenter, abc, walt, disney, entertainment,	0.93
	MULTIPLE EXPLAINERS	espn, sportscenter, abc, walt, disney, news, espn.com	**0.99**
hp mini 2140	TEXT MATCHING	hp, mini, 2140, 2133	**0.94**
	POSITION AWARE	hp, mini, 2140, 2133	0.90
	SEMANTIC SIMILARITY	hp, touchpad, overview, hdd,	0.71
	MULTIPLE EXPLAINERS	hp, mini, 2140, 2133, touchpad, overview	0.91

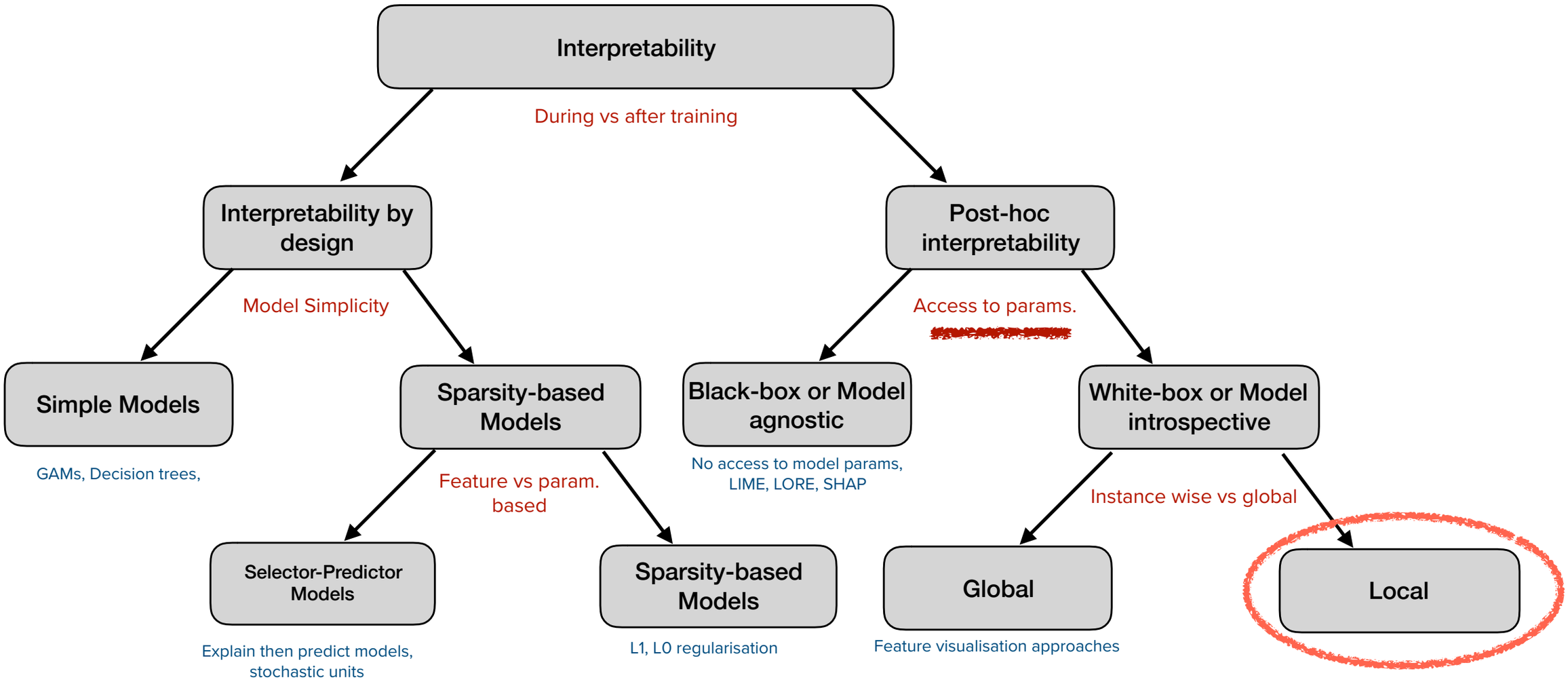
What about LtR models?

Extracting per Query Valid Explanations for Blackbox Learning-to-Rank (LTR) Models

Jaspreet Singh, Megha Khosla, Zhenye Wang & Avishek Anand
ICTIR'21

End of Part 1

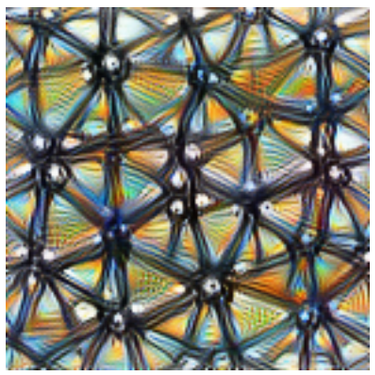
https://github.com/GarfieldLyu/EXS


Interpretability
During vs after training
Interpretability by design
Post-hoc interpretability
Model Simplicity
Access to params.
Simple Models
Sparsity-based Models
Black-box or Model agnostic
White-box or Model introspective
GAMs, Decision trees,
No access to model params, LIME, LORE, SHAP
Feature vs param. based
Instance wise vs global
Selector-Predictor Models
Sparsity-based Models
Global
Local
Explain then predict models, stochastic units
L1, L0 regularisation
Feature visualisation approaches

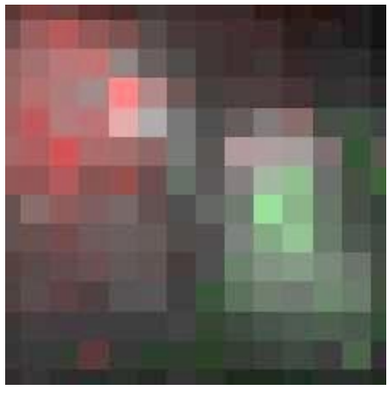
Interpreting Neural Networks

- **Feature visualization**: Visualizing components of the neural networks
  - Activations of neurons
  - Attention values
  - Gradient flow
- **Feature attributions**: relevant input features
  - Which input features are responsible for the given decision ?
  - Sensitivity analysis using gradient-based methods
  - Using black-box methods like LIME, SHAP, etc.

- Feature visualization: Visualizing components of the neural networks



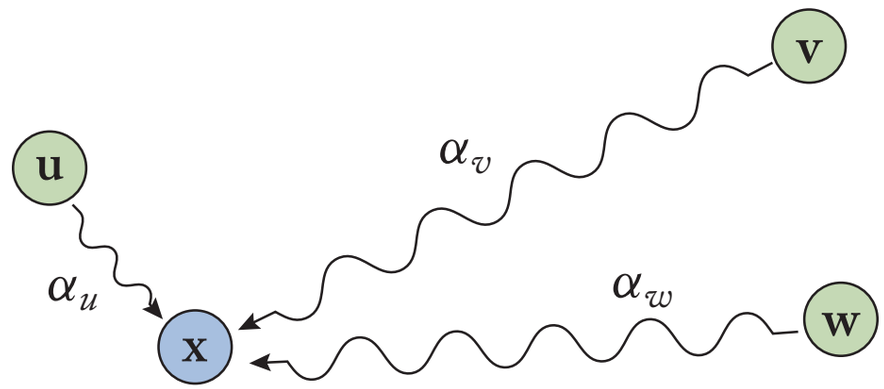
- Feature attributions: relevant input features



Attention in Language

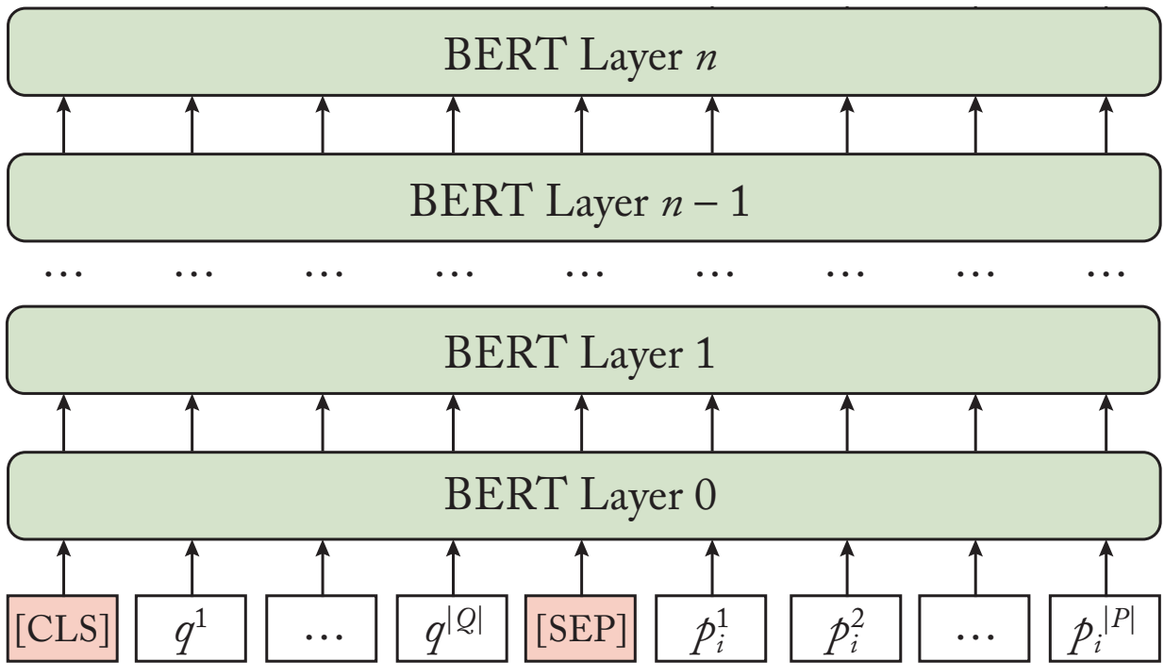
- Attention mechanism in neural language models is crucial for extracting latent features
- Self attention in language is aimed at re-representing the initial representation based on the context
- Neural models consume non-contextual token-level representations and output contextual token-level representation

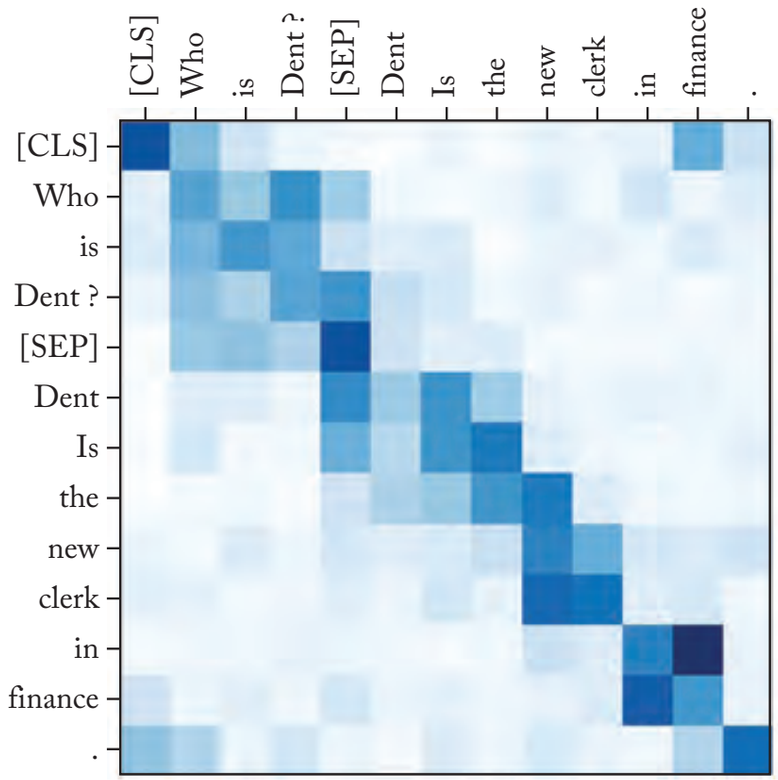
New contextual representation

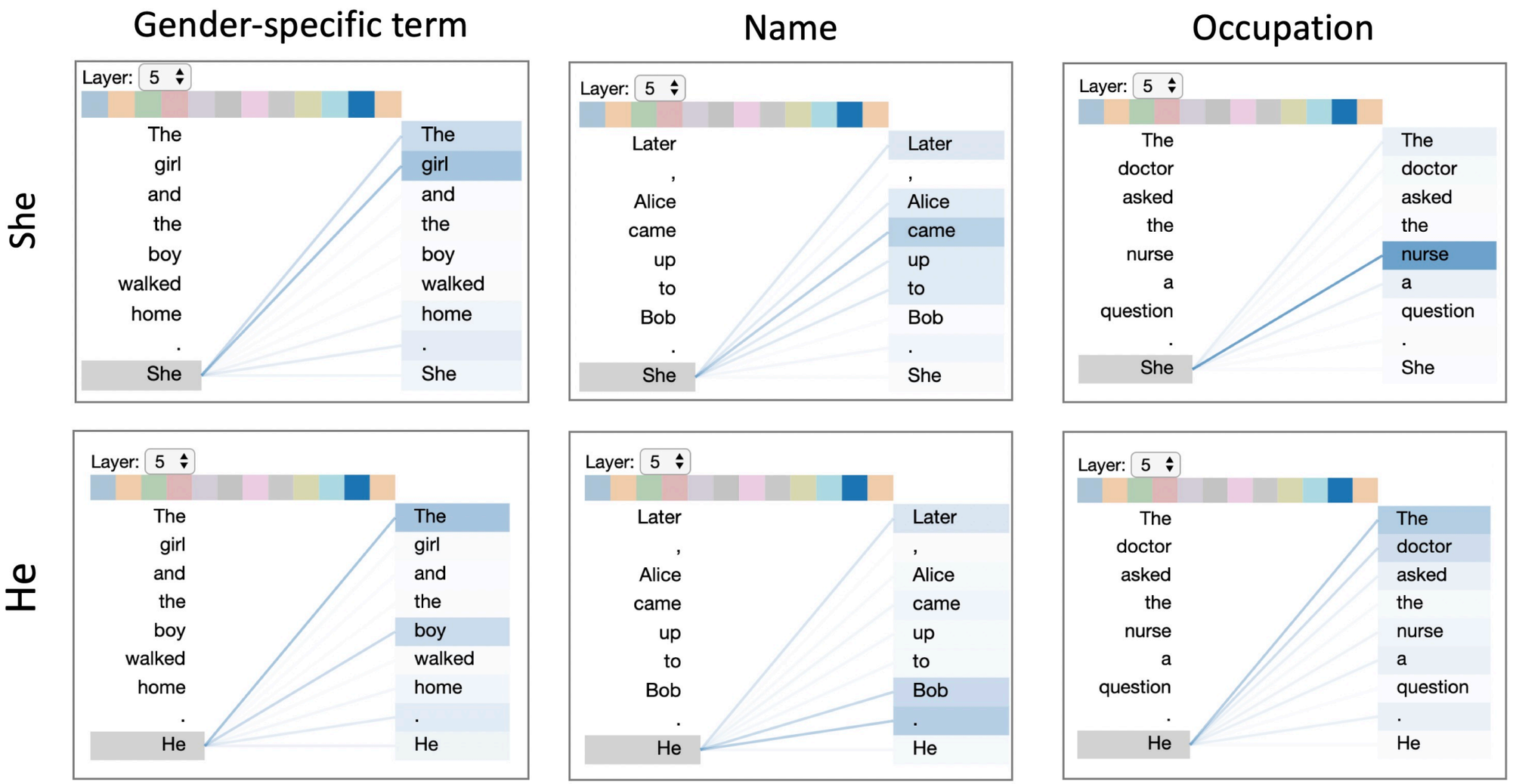


$$\mathbf{x}' = \alpha_u \mathbf{u} + \alpha_v \mathbf{v} + \alpha_w \mathbf{w}$$

$$\alpha_u = \frac{e^{\text{sim}(\mathbf{u},\mathbf{x})}}{e^{\text{sim}(\mathbf{u},\mathbf{x})} + e^{\text{sim}(\mathbf{v},\mathbf{x})} + e^{\text{sim}(\mathbf{w},\mathbf{x})}}; \quad \text{sim}(\mathbf{u},\mathbf{x}) = \mathbf{x} \cdot \mathbf{W}\mathbf{u}$$


BERT Layer n
BERT Layer $n-1$
BERT Layer 1
BERT Layer 0
[CLS]
q^1
$q^{|Q|}$
[SEP]
p_i^1
p_i^2
$p_i^{|P|}$


[CLS]
Who
is
Dent ?
[SEP]
Dent
Is
the
new
clerk
in
finance
.



https://aclanthology.org/P19-3007.pdf

https://github.com/jessevig/bertviz

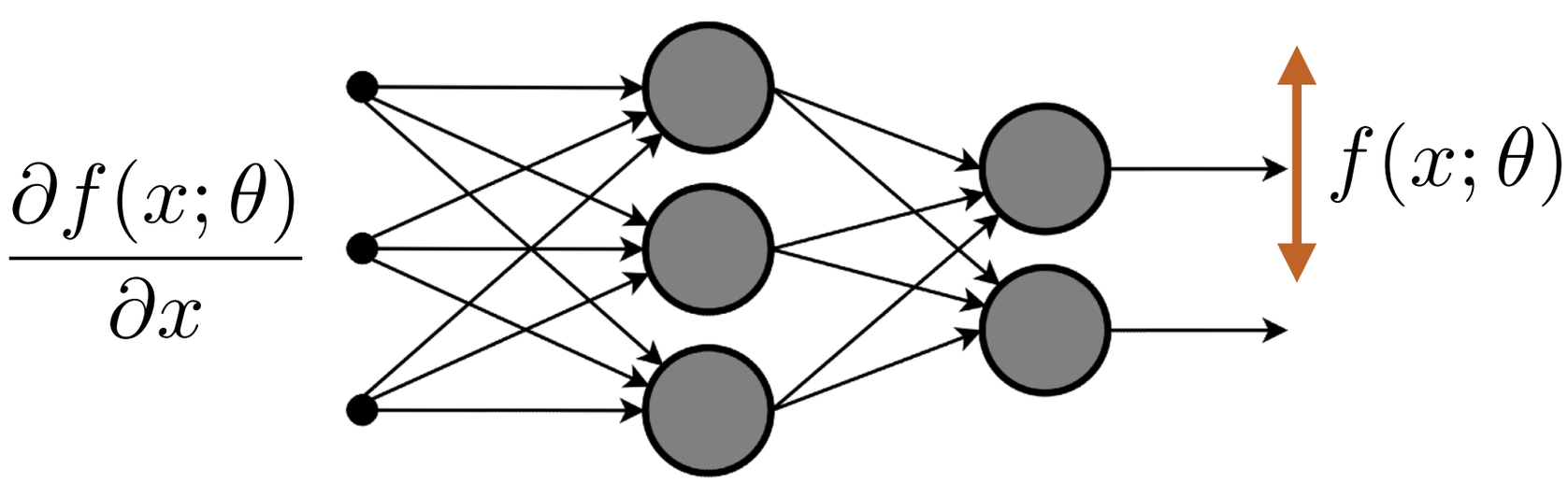
Sensitivity Analysis

- Neural Networks are differentiable machines
  - The output can be written as a function of the parameters and input
  - One can differentiate the output function w.r.t parameters
  - The underlying idea is used for training Neural Nets using gradient descent

$$f(x;\theta) \qquad \frac{\partial f(x;\theta)}{\partial \theta}$$

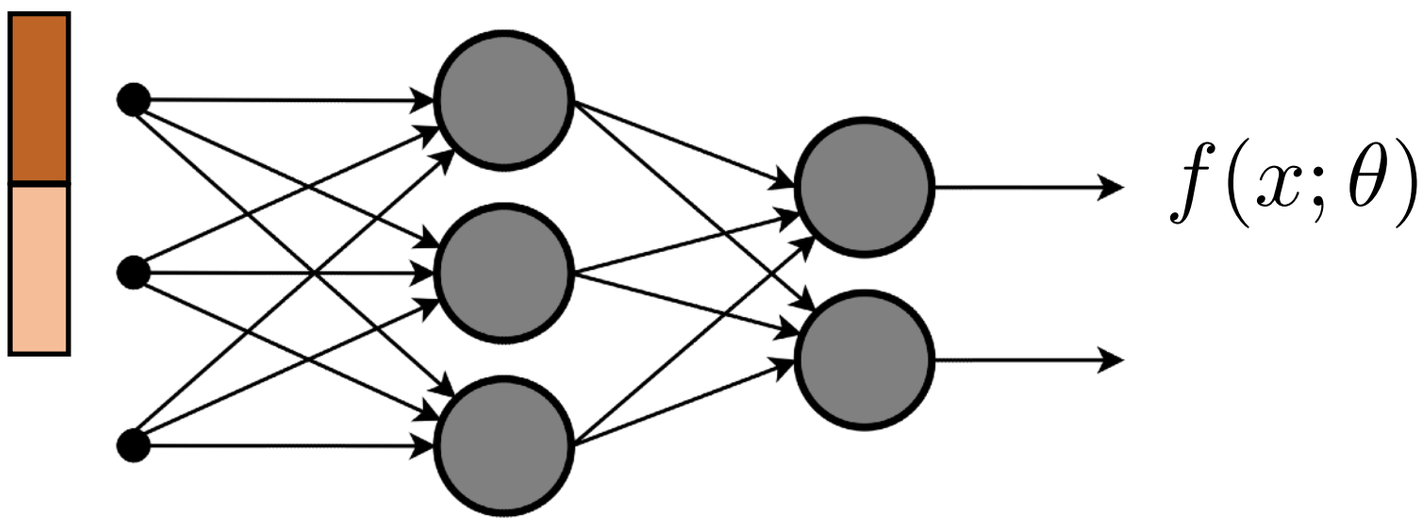
- Sensitivity Analysis: How sensitive is the output $f()$ w.r.t to a small change in the input ? $\frac{\partial f(x;\theta)}{\partial x}$

- How sensitive is the output $f()$ w.r.t to a small change in the input ?
  - If a small change in the input feature causes a large change in output, then that feature is responsible for the prediction
  - Back-propagation into the input: instead of computing $\frac{\partial f(x;\theta)}{\partial \theta}$



Saliency Maps

- Visualize the gradients over each feature
  - as a heat map or Saliency Maps
  - Saliency maps are feature attribution methods that are based on gradients

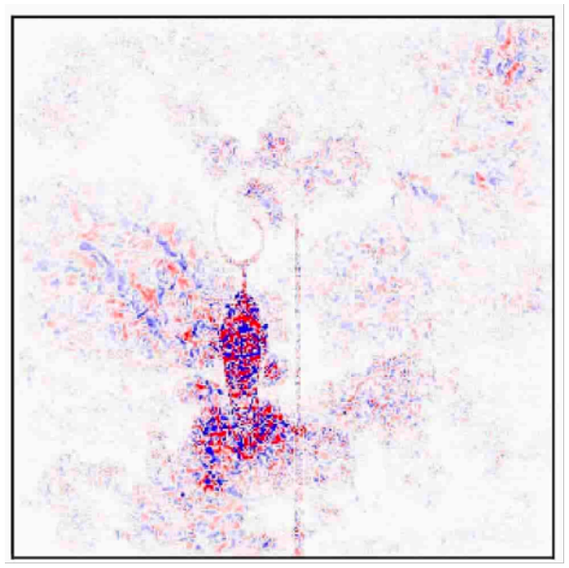


Saliency Maps for Images

- Images have multiple channels where each channel is a 2-D matrix

$$M_{ij} = \max_c |\nabla_x S_c(X)|_{(i,j,c)}$$

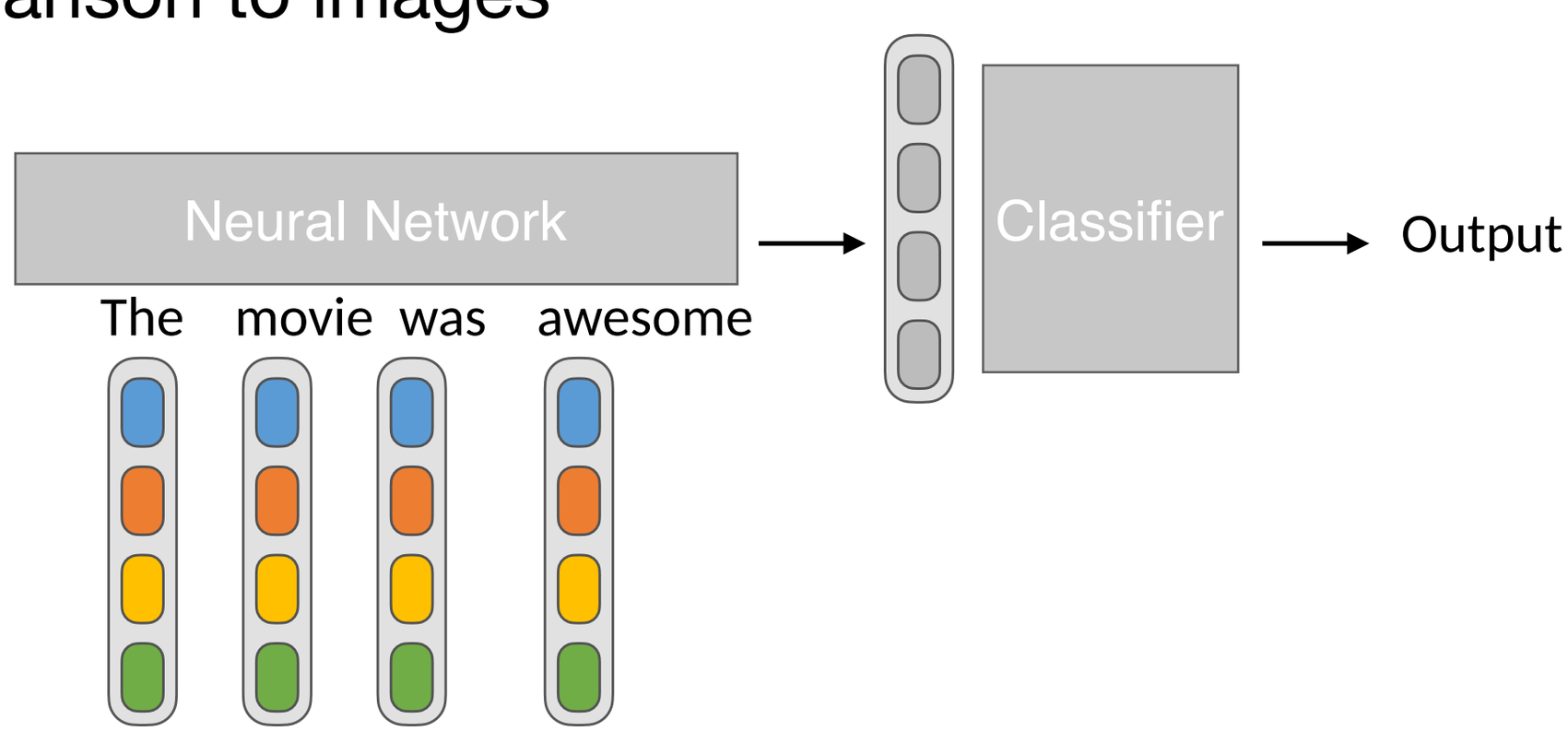
$$M_{ij} = \max_c |\nabla_x S_c(X)|_{(i,j,c)}$$



[Simonyan, Vedaldi, Zisserman ‘14]

Saliency Maps for Language

- Words are associated with an embedding
- Computing gradients back to the inputs is different in comparison to images

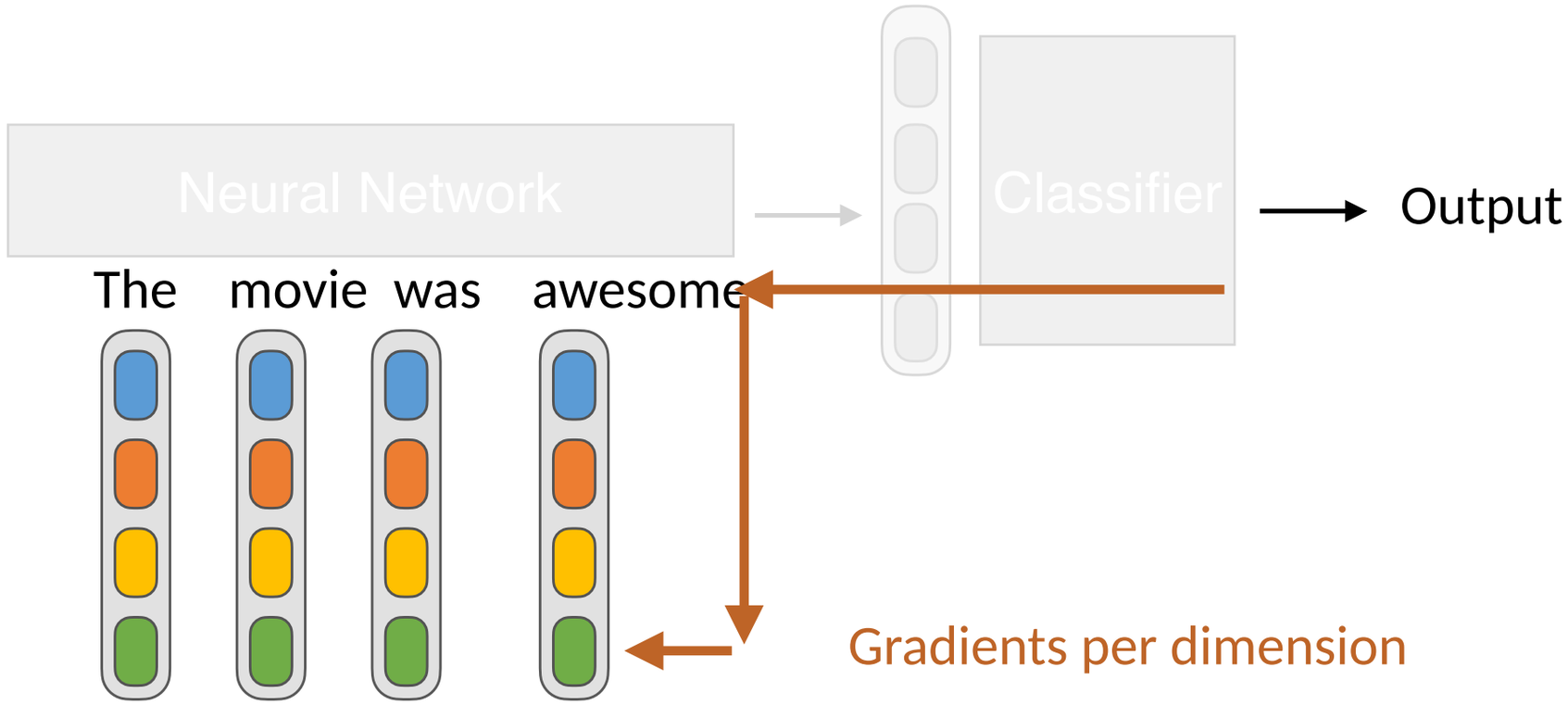


[Simonyan, Vedaldi, Zisserman '14]

Saliency Maps for Language

- We obtain gradients per dimension but we want attributions or importance scores at the level of world

- **Idea:** Simple aggregations of dimension-level gradients like sum, average, etc.



[Simonyan, Vedaldi, Zisserman ‘14]

Saliency Maps - Setting

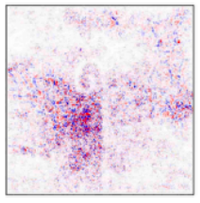
Which features are responsible for the decision given..

A trained model
S

An instance
x

Access to model
parameters

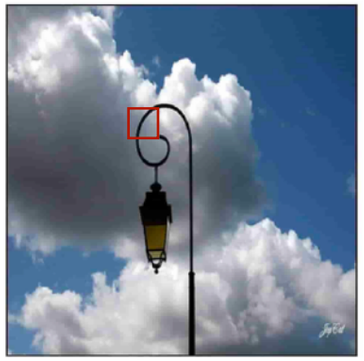
A feature is more relevant if a small perturbation causes large change in the output

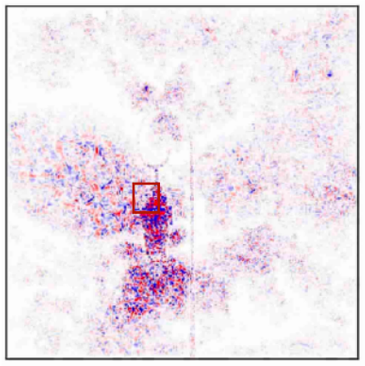


Saliency Map

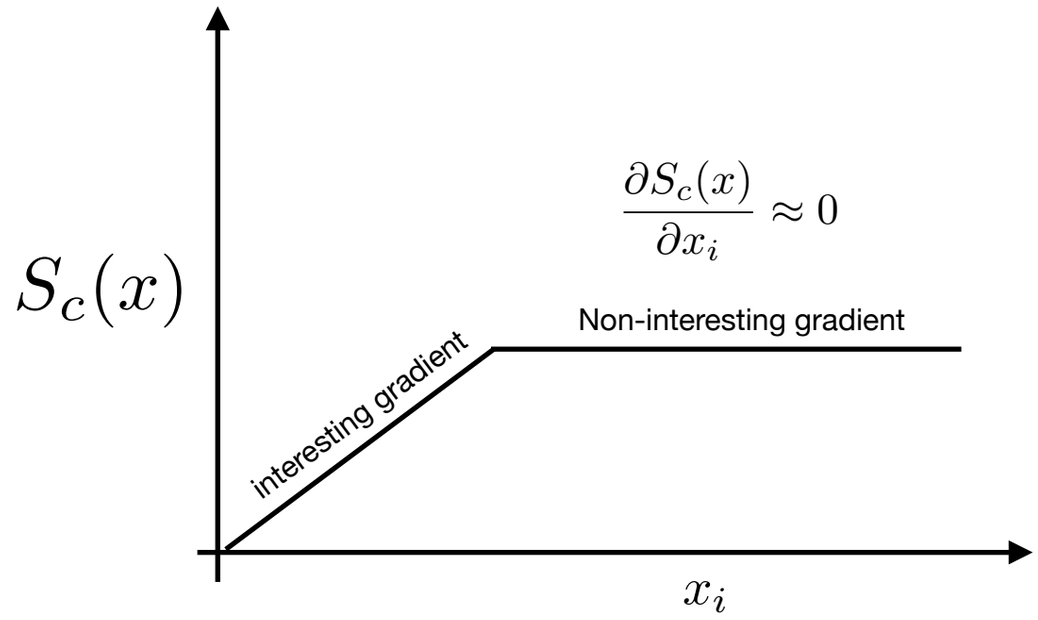
$$R_i^c(x) = \frac{\partial S_c(x)}{\partial x_i}$$

Problems with Deep Nets





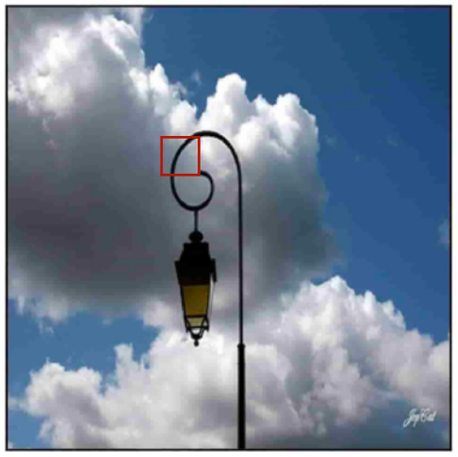
$$R_i^c(x) = \frac{\partial S_c(x)}{\partial x_i}$$



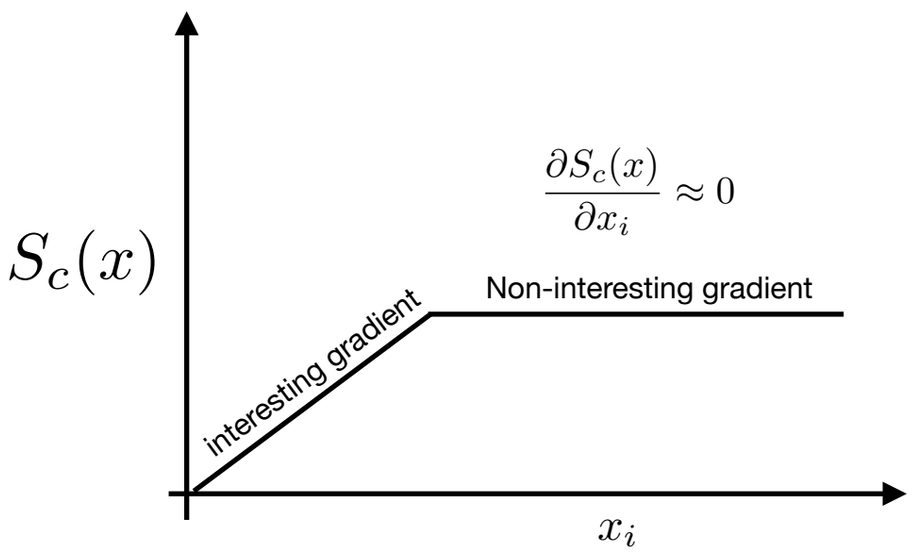
Deep Neural Networks are usually trained till "Saturation"

Perturbing Inputs

- **Small perturbations** at the saturation point do not give us interesting gradients
- Extreme perturbation (to say a baseline image) can give us interesting gradients

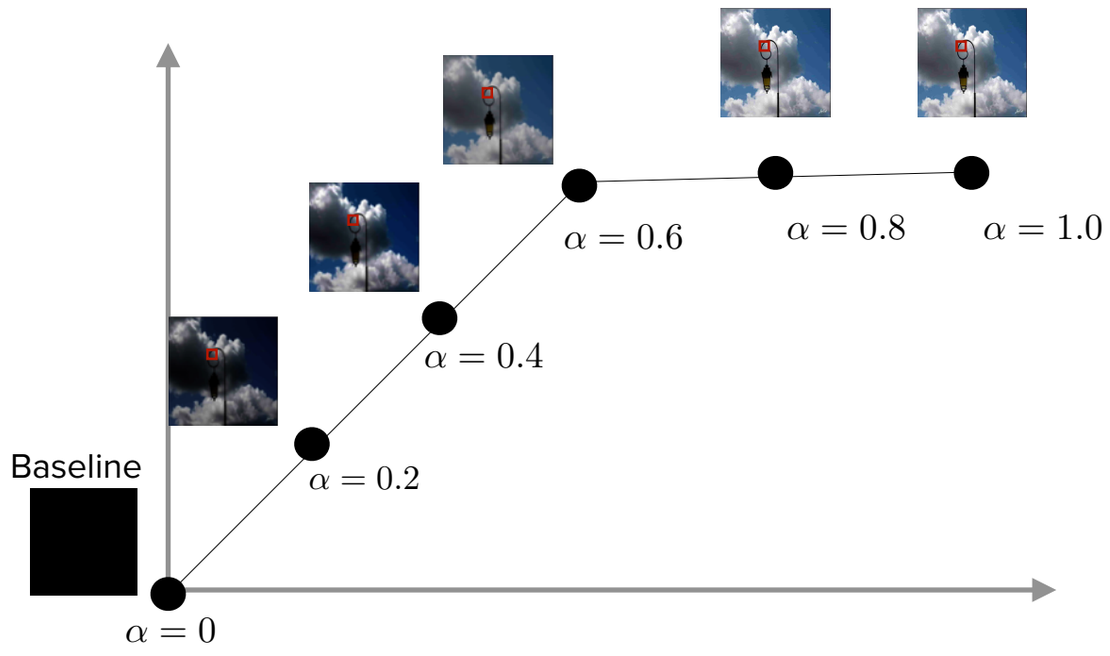


$$R_i^c(x) = \frac{\partial S_c(x)}{\partial x_i}$$



Integrated Gradients

Compute gradient estimate based on gradients over a path of specific perturbations



Choose a Baseline 

Integrated Gradients

1. Choose a Baseline to contrast
2. Compute gradients at different mask values
3. Attribution = Aggregation over gradients computed for a certain set of perturbations

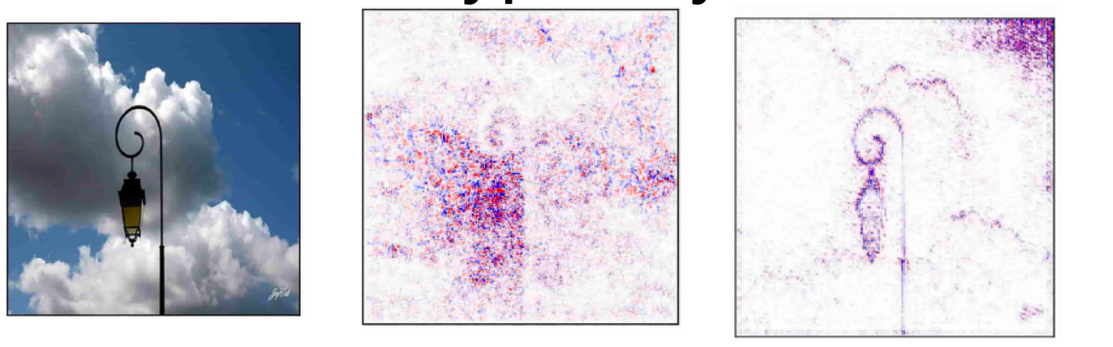
$$R_i^c(x) = x_i \cdot \int_{\alpha=0}^{1} \frac{\partial S_c(\tilde{x})}{\partial(\tilde{x}_i)}\bigg|_{\tilde{x}=\bar{x}+\alpha(x-\bar{x})} d\alpha$$

Baseline — Original

Integrated Gradients monitors how the network changes from a zero signal input to actual input through the use of gradients

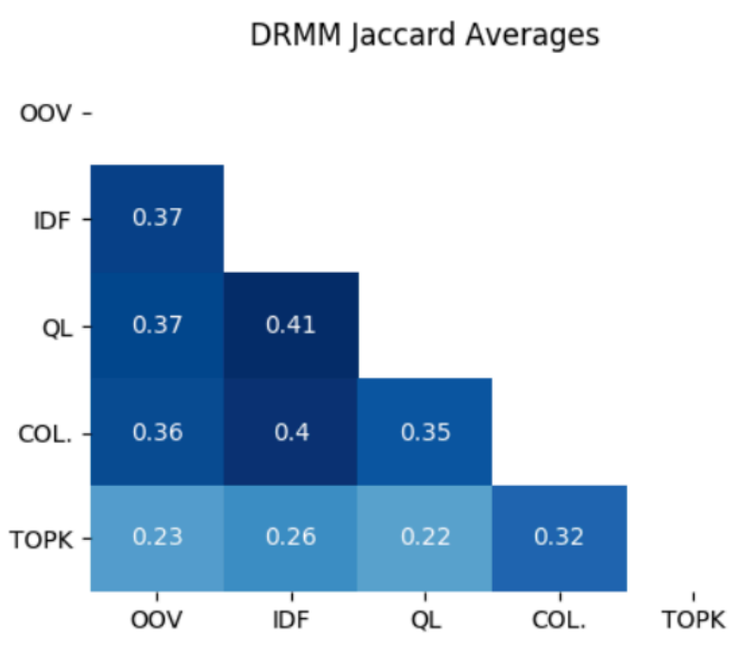
Baseline

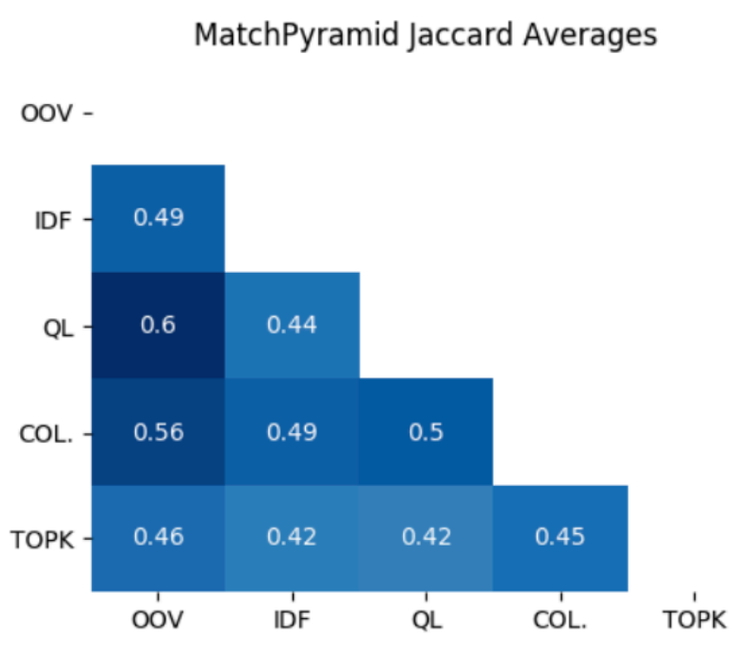
- Baseline is an information less input
- The choice of baselines matters a lot and is typically domain dependent
  - Black or gray images
  - Zero embedding in language
  - Random document in retrieval

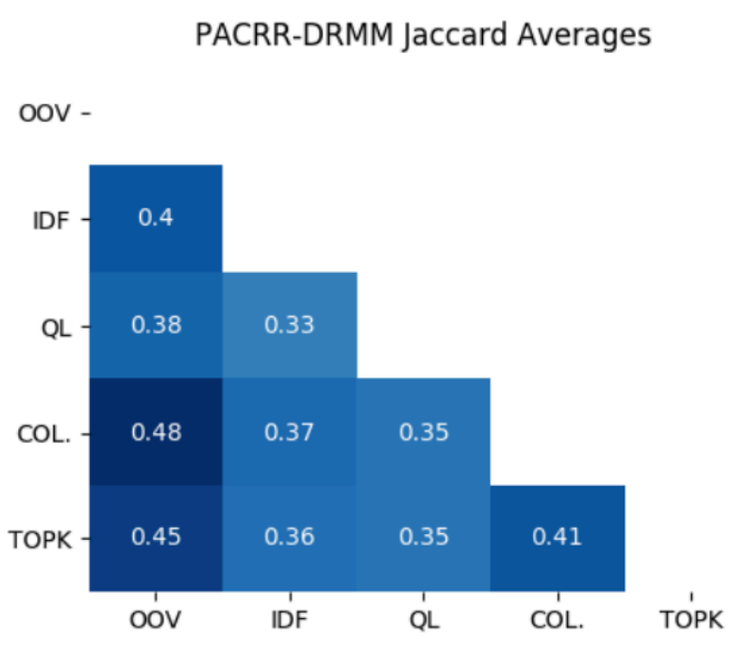


Simple Gradient Integrated Gradients

A study on the Interpretability of Neural Retrieval Models using DeepSHAP [SIGIR'19]







Background Document for Ranking models?

- OOV: zero embeddings
- Col: sample random document from collection
- IDF: sample low IDF terms — more generic terms
- Top-K: sample random document from top-k list inversely proportional to relevance predicted
- QL: sample words from the top-k document language model

Table 3: An example of words selected by LIME and SHAP methods for the query *'cult lifestyles'* and document *'FBIS3-843'* which is about clashes between cult members and student union's activists at a university in Nigeria. *Words unique to a particular explanation method are highlighted in bold.*

LIME	OOV	IDF	QL	COL.	TOPK
cult	cult	cult	cult	cult	cult
style	followers	style	style	black	**numbers**
followers	black	followers	elite	fraternities	**english**
elite	fraternities	**suspects**	saloon	degenerate	**college**
saloon	degenerate	**belong**	**final**	sons	**university**
student	sons	**reappearing**	march	followers	**fallouts**
home	**academic**	household	**friday**	style	**buccaneers**
members	**american**	black	september	home	**feudings**
march	**tried**	fraternities	**arms**	household	**activists**
september	household	degenerate	**closed**	**avoid**	**troubles**

- If we fix the reference document, can we compare the explanations of different models?

- How do we know which background document is the right choice?

  - Depends on the question being asked

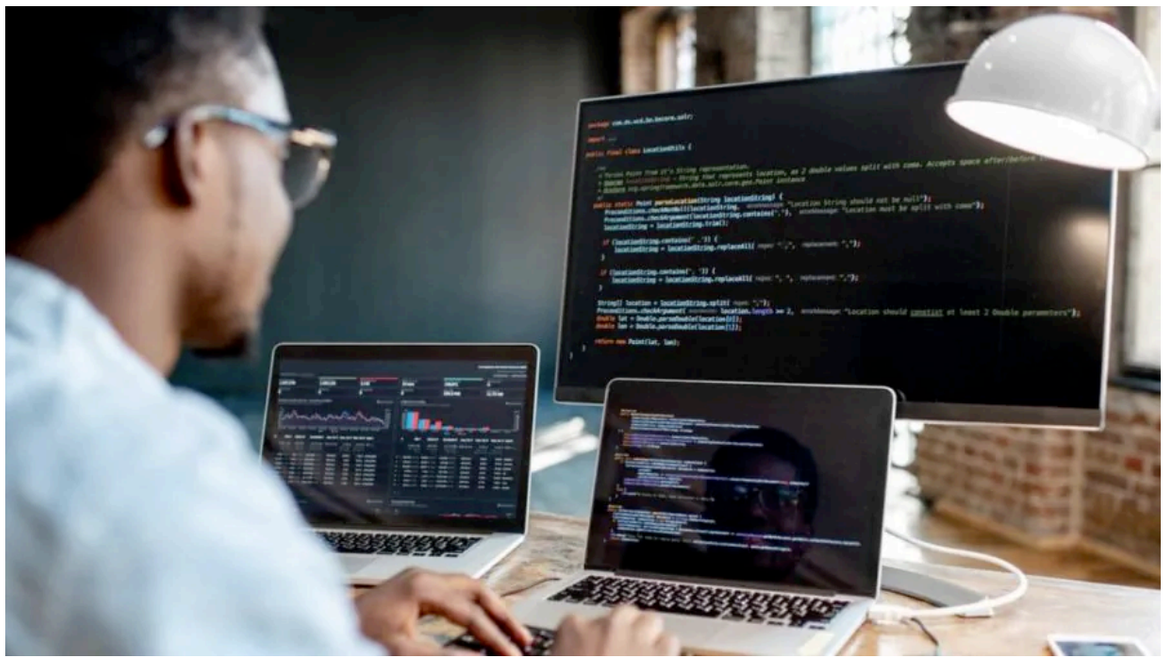
- Can we interpret BERT ranking models using DeepSHAP and compare it to the attention weights visualised by BERTViz?

Who needs interpretability in Search?

- The general web search user probably does not need an explanation of why a particular result is favoured by the model
    - Unless its biased or false
- Explanations for search and recsys in industry tend to focus on persuasion
    - Give the user a reason to click on a search result
- Interpretability in search needs to be looked at from a different perspective


Auditor
www.taxguru.in


ASK THE
EXPERT



The End

- Interpretability in search is still a nascent topic
- We discussed techniques to explain ranking models but what about matching models? Are they the same in the case of text only applications?
- General search users usually do not require explanations but expert users, auditors and developers do
- Can we build explanation models that aid developers in particular?
- Can we apply existing techniques to multimodal search? Query is text, items are images

- We train a host of models with similar performance but which model is more human like?
    - Same training data, different architecture
- What factors of training a ranking model influence its predictions? Can intent explanations highlight biases of different BERT style models?
    - Different training data, loss functions. Same architecture.