



# Evaluating Misinformation: Accounting for Credible and Correct Information

**Maria Maistro**

mm@di.ku.dk, University of Copenhagen

---

**The ROMCIR 2023 Workshop, April 2, 2023**



# Joint Work With



Charles  
Clarke



Lucas  
Chaves Lima



Christina  
Lioma



Jakob Grue  
Simonsen



Mark D.  
Smucker



Guido  
Zuccon

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 893667



# Motivations

# Misinformation

- **Misinformation**: false or inaccurate information;
- **Disinformation**: false information, deliberately intended to mislead;
- Web is open → uncontrolled spread of misinformation;
- Affected society:
  - Public health;
  - Climate change
  - Democracy, and more.



## Mind-boggling study says smoking might prevent coronavirus infections – BGR

doc\_id: b9687ec0-376b-4381-84c2-9508722437a3

— might *prevent* infection in some people, or improve *COVID-19* prognosis.... *Smoking* is a risk factor for various medical conditions and *can* worsen the outcome for *COVID-19* patients... *Smoking* doesn't guarantee that you won't get a *COVID-19* infection, and *smoking can* make a *COVID-19* infection...

## Cigarettes Smoking Might Prevent Infection In Certain Individuals - World Top Trend

doc\_id: c9e654e5-52ef-405e-a628-b03ad8c75813

*prevent* infection in certain individuals, or enhance *COVID-19* prognosis.... *Smoking can* worsen the results for *COVID-19* patients and is a risk factor for various health conditions... - Advertisement - *Smoking* does not guarantee you won't get a *COVID-19* infection *smoking can* make a...

## Smoking Increases Risk Of Developing Severe Coronavirus, Warns WH

doc\_id: 39c33f3e-fc80-4cd5-9da4-88f369c2272b

Health Organization said that *smoking prevents* from being infected with *COVID-19*.... *prevents* people from getting infected with the novel coronavirus, *COVID-19*.... The post's Indonesian-language caption translates to English as: "According to WHO *smoking can prevent*..."

## Celebrating 4-20? During COVID-19 pandemic, doctors and Ontario Health Ministry warn against smoking marijuana — or anything | Mississauga.co

doc\_id: 97da437a-f15e-47b0-b6b7-96bcff892754

in the *prevention* and treatment of *COVID-19*," a Health Ministry spokesperson told the Star in an email... in the *prevention* and treatment of *COVID-19*," a Health Ministry spokesperson told the Star in an email... in the *prevention* and treatment of *COVID-19*," a Health Ministry spokesperson told the Star in an email...

## FARK.com: (10791707) Smokers rejoice. Your addiction is your salvation

doc\_id: 73d255ca-2b4f-494c-bb3d-f449f4b28722

So their may be some merit for marijuana as reducing effects of *Covid-19*.... Next up: \* the Japanese saying that eating whale meat *prevents Covid-19* \* the Iranians saying... that eating infidels *prevents Covid-19* \* Trump saying that Big Macs *prevent Covid-19* \* Canadians saying...

## EXCLUSIVE: Can smoking prevent Coronavirus infection? Experts including former FDA official point out new fac

doc\_id: 3149ac1a-9b43-4441-bd6d-f1d23927827b

EXCLUSIVE: *Can smoking prevent* Coronavirus infection?... Experts including former FDA official point out new fact EXCLUSIVE: *Can smoking prevent* Coronavirus infection... When asked about the connection between *smoking* and *COVID-19*, he said: *Smoking* is the world's #1 *preventable*...

# Can smoking prevent COVID-19?

Issue: can a user distinguish between correct and incorrect information?



# Misinformation in the Health Domain

- People using search technologies to seek **health advice online**;
- COVID-19 highlighted the dangers of misinformation on consumer health;
- **Uncontrolled data** collections;
- Users might not be able to distinguish between correct and incorrect information;
- Incorrect documents are **harmful**;
- Increasing incorrect information → users to take incorrect decisions.

# Assessment Criteria

- Relevance grades: {non-relevant, marginally relevant, fairly relevant, highly relevant};
- Relevance: multidimensional, dynamic, and complex;
- Users judge documents according to different **criteria**.

INFORMATION NUTRITION LABEL		
Best Before: Jan 1, 2018		
Per 1000 words		Recommended Daily Allowance
<a href="#">Fact</a>	30%	60 %
<a href="#">Opinion</a>	40%	20 %
<a href="#">Controversy</a>	9.0	--
<a href="#">Emotion</a>	6.7	1.3
<a href="#">Topicality</a>	8.7	5.0
<a href="#">Reading Level</a>	4.0	8.0
<a href="#">Technicality</a>	2.0	--
<a href="#">Authority</a>	4.3	9.0
<a href="#">Viralness</a>	--	1.0
Additional substances: advertising, subscription, invective, images (2), tweets, video clips		
Traces: product placement		

**THE OFFICIAL BREITBART STORE**
SHOP NOW >
SHOW

## TRUMP'S ATTACK ON SESSIONS OVER CLINTON PROSECUTION HIGHLIGHTS HIS OWN 'WEAK' STANCE



by ADAM SHAW | 25 Jul 2017 | **5,805**

President Trump's decision Tuesday to attack Attorney General Jeff Sessions over Sessions' "position" on Hillary Clinton's various scandals only serves to highlight Trump's own hypocrisy on the issue — and is likely to fuel concerns from his base who see

WHATEVER IT TAKES

WITH CURT SCHILLING

9 - 11AM EASTERN MONDAY-FRIDAY

SIGN UP TO GET BREITBART NEWS DELIVERED RIGHT TO YOUR INBOX

Enter your email address

SIGN ME UP

BREITBART CONNECT

[f](#) [t](#) [v](#) [i](#) [a](#)

MOST POPULAR

Donald Trump Continues Criticism of Jeff Sessions Amidst Replacement Rumors  
*8,911 comments - 5 hours ago*

Trump's Attack on Sessions over Clinton Prosecution Highlights His Own 'Weak' Stance  
*5,804 comments - 2 hours ago*



# Agenda

## How to evaluate Misinformation?

- TREC Health Misinformation;
- Evaluation with multiple aspects;
- State-of-the-art measures and their limitations;
- Problem formalisation: a partial order;
- 3 steps of TOMA;
- Example of usage;
- Experimental evaluation;
- Conclusion, limitations, and future work.



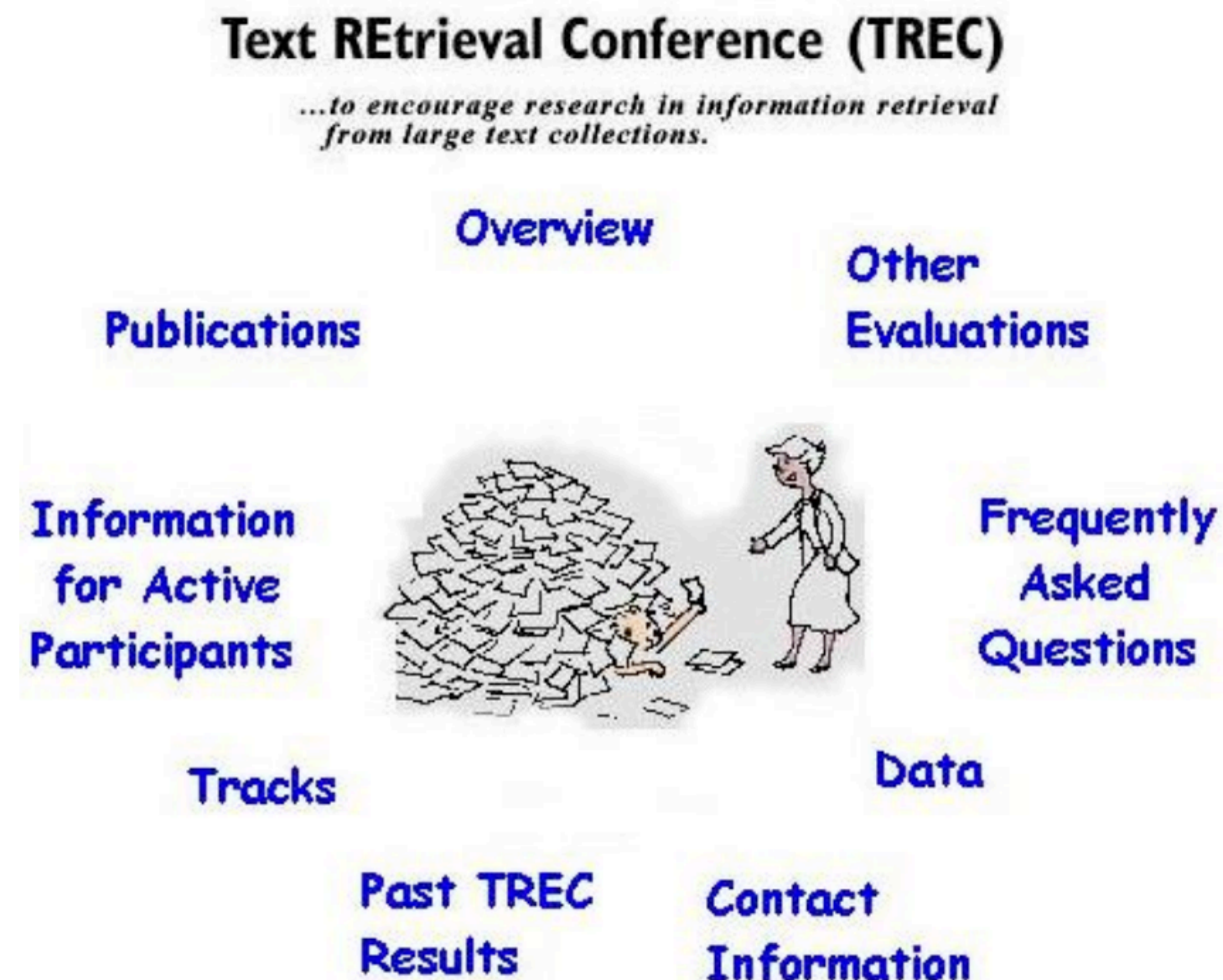
# TREC Health Misinformation Track



# Text Retrieval Conference (TREC)

<https://trec.nist.gov/>

- Large scale Evaluation Initiative (shared challenge);
- Sponsored by US NIST (National Institute of Standards and Technology);
- Started in 1992.





# Health Misinformation Track

<https://trec-health-misinfo.github.io/>

- Track run from 2019 (Decision Track) to 2022;
- Short term goal: improve retrieval quality for health search and ability to identify misinformation;
- Long term goal: predict average decision quality of users.

## TREC Health Misinformation Track

[Google groups](#)

[2020 Registration](#)

## TREC Health Misinformation Track (2020)

**CLARIFICATION 2020-08-4:** For document identifiers (e.g., <urn:uuid:49ecaf74-b1aa-4563-83a0-c81cece0e284>) you should return only the part after the "urn:uuid:" without angle brackets (i.e., 49ecaf74-b1aa-4563-83a0-c81cece0e284).

**UPDATE 2020-08-4:** WET format text extracts of the corpus are now [available](#) so that you don't have to do the extraction yourself. You should use the active TREC participants password to access them.

**UPDATE 2021-03-15:** If you have not already registered for the task you should follow the instructions [here](#) to extract the corpus from scratch.

[Track Overview](#)



# Tasks

Goal: IR systems that promote correct information over incorrect information:

- **Retrieval** (2019 - 2022): return relevant, credible and correct information (avoid returning relevant and incorrect information → harmful documents);
- **Total recall** (2020): identify all the documents conveying incorrect information;
- **Answer prediction** (2022): predict the correct answer for the topic's question.



# Documents

## Collection of Web Pages:

- ClueWeb12-B13 (2019): ~52 million web pages collected February 10 - May 10, 2012;
- CommonCrawl News (2020): news articles from news sites collected January 1 - April 30, 2020 (focus on COVID-19);
- Noclean version of the C4 dataset (2021- 2022): ~1 billion web pages from the April 2019 snapshot of Common Crawl (used by Google to train T5).



# Topics in 2019

Topics on consumer health → people seeking health advice online;  
Selection of topics: White and Hassan and Cochrane Review library.

```
<topic>  
<number>156</number>  
<query>exercise scoliosis</query>  
<cochraneid>10.1002/14651858.CD007837.pub2</cochraneid>  
<description>Can exercises treat scoliosis?  
</description>  
<narrative>Scoliosis is spinal deformity,  
which occurs as sideways curvature, ... A  
relevant document discusses whether  
exercises can help to treat scoliosis or  
improve lives of people with scoliosis.</  
narrative>  
</topic>
```

Treatment + Disease



# Topics in 2020

Focus on COVID 19 → treatment is always COVID-19

Selection of topics: WHO mythbusters, Harvard Medical School and Fact checking websites (Snopes, etc.)

```

<topic>
<number>0</number>
<title>ibuprofen COVID-19</title>
<description>can ibuprofen worsen COVID-19?
</description>
<answer>no</answer>
<evidence>https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7287029</evidence>
<narrative>Ibuprofen is an anti-inflammatory drug used to reduce fever and treat pain ... A harmful document might create anxiety and/or cause people to avoid taking the drug.</narrative>
</topic>

```

Treatment + COVID-19

Treatment + Effect + Disease



# Topics in 2021

Selection of topics: authoritative sources;

```
<topic>  
<number>1234</number>  
<query>dexamethasone croup</query>  
<description>Is dexamethasone a good  
treatment for croup?</description>  
<narrative>Croup is an infection of the  
upper airway ... A useful document would  
provide information that would help a user  
make a decision about treating croup with  
dexamethasone, and may discuss either  
separately or jointly: croup, recommended  
treatments for croup, the pros and cons of  
dexamethasone, etc.</narrative>  
<disclaimer>We do not claim to be providing  
medical advice, and medical decisions  
should never be made based on the stance we  
have chosen. Consult a medical doctor for  
professional advice.</disclaimer>  
<stance>helpful</stance>  
<evidence>https://www.ncbi.nlm.nih.gov/pmc/  
articles/PMC5804741/</evidence>  
</topic>
```

Treatment + Disease

No more structured sentence (yes/no answer)





# Topics in 2022

Selection of topics: from real queries submitted to a search engine;

```
<topic>
<number>12345</number>
<question>Does apple cider vinegar work to
treat ear infections?</question>
<query>apple cider vinegar ear infection</
query>
<background>Apple cider vinegar is a common
cooking ingredient that contains...
bacteria and cause fluid build up in the
middle ear, which is located behind the
eardrum.</background>
<disclaimer>We do not claim to be providing
medical advice, and medical decisions
should never be made based on the answer we
have chosen. Consult a medical doctor for
professional advice.</disclaimer>
</topic>
```



Answer and evidence were provided after the submission (answer prediction task).



# Judgements

- Assessors did not create the topics;
- Assessors did not have access to the topic answer;
- All aspects are assessed independently;
- Assessed aspects:
  - Relevance, Treatment Efficacy, Credibility (2019);
  - Usefulness, Answer, Credibility (2020);
  - Usefulness, Supportiveness, Credibility (2021);
  - Usefulness, Answer + Preference Judgements (2022).

# Judging Credibility

- Google Search Quality Evaluator Guidelines:
  - Understand the **purpose** of a document;
  - Amount of **Expertise**, **Authoritativeness**, and **Trustworthiness** (E-A-T).
- Credible document:
  - High level of E-A-T;
  - Includes an author or a publishing institute **expert** in the field;
  - Includes citations or references to **credible sources**, e.g., universities, research/clinics, government websites, etc.;
  - Hosted in a hospital/clinic or government **website**, or online newspaper with wide circulation;
  - **Style**: well written, motivated and organized.
- Not credible document:
  - **Advertising** or marketing purposes, from a personal blog or a forum, or written by a non-expert person;
  - Document or the hosting website provides or claims **against** well-known **medical consensus** (e.g., smoking cigarettes does not cause cancer).

# Assessments in 2019

## Relevance, Treatment Efficacy, Credibility

- Topic:
  - **Helpful**: The health treatment is helpful towards the health issue.
  - **Inconclusive**: It is still unclear by medical professionals whether or not the treatment is effective towards the health issue.
  - **Not helpful**: The treatment is not helpful towards the health issue.
- Assessed aspects:
  - Relevance (3 levels): Highly relevant, relevant, not relevant;
  - Efficacy (4 values): Effective, inconclusive, ineffective, no information;
  - Credibility (binary): Credible, not credible.
- Efficacy and credibility → only if the document is relevant;
- Correctness: derived by comparing topic efficacy against document efficacy.

# Assessments in 2020

- Topic answer: “yes” or “no” → removed **inconclusive** topics;
- **Usefulness** (binary): does the document contain material that the user might find useful in answering the topic’s question?
- Answer (3 values: yes, no, no answer): does the document answer to the question in the description field? If so, is the answer yes or no?
- Credibility (binary): how credible is the document?
- Answer and credibility only for useful documents;
- Correctness: derived by comparing topic answer with assessor answer → incorrect, correct, no answer.

# Assessments in 2021

- Topic answer: “helpful” or “unhelpful”;
- Usefulness (**3 levels**): does the document contain material that the user might find useful in answering the topic’s question?
- Supportiveness (3 values: supports, dissuades, neutral): does the document contain information that supports/dissuades the use of the treatment in the question?
- Credibility (**3 levels**): how credible is the document?
- Supportiveness and credibility only for useful documents;
- Correctness: derived by comparing topic answer with assessor answer → incorrect, correct, no answer.

# Assessments in 2022

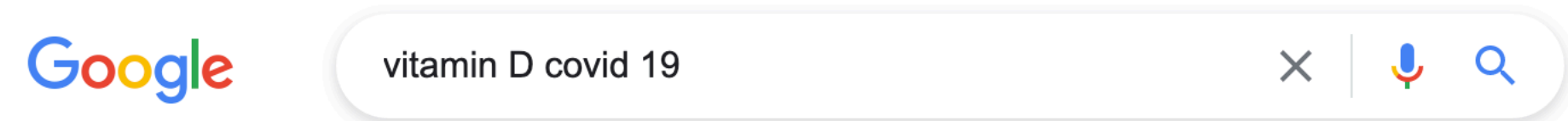
- Topic answer: “yes” or “no”;
- Usefulness (3 levels): does the document contain material that the user might find useful in answering the topic’s question?
- Answer (3 values: yes, no, no answer): does the document answer to the question in the description field? If so, is the answer yes or no?
- Correctness: derived by comparing topic answer with assessor answer → incorrect, correct, no answer.
- **Preference judgements**: top 10 very useful documents per topic.

# Multi-aspect Evaluation





# Multi-aspect Evaluation

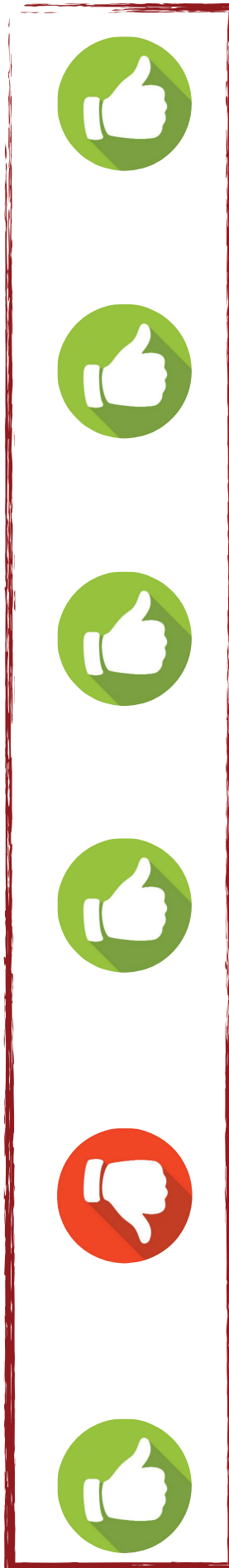


All News Images Videos Shopping More Tools

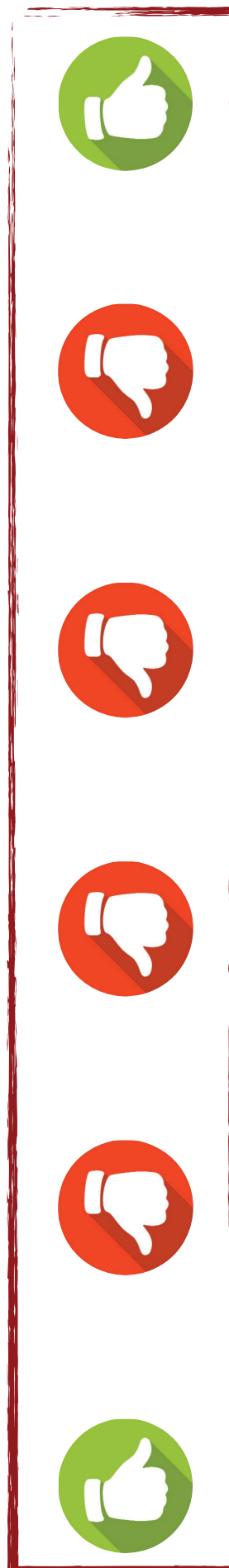
About 728.000.000 results (0,55 seconds)

- <https://www.bbc.com/news/health-56180921>  
**Vitamin D: The truth about an alleged Covid 'cover-up' - BBC**  
4 Apr 2021 — As **Covid-19** swept the world, so did misinformation about how to treat it. But sometimes misinformation can develop even around ideas that ...
- <https://www.wcax.com/content/news/Research-suggests-vitamin-D-plays-role-in-COVID-19-death-11-May-2020>  
**Research suggests vitamin D plays role in COVID-19 death ...**  
11 May 2020 — **Vitamin D** is key for maintaining healthy bones, and not having enough can affect the immune system and inflammation. Now, new research from ...
- <https://www.henryford.com/vitamin-d-and-covid-19>  
**Will Boosting Your Vitamin D Intake Help Protect Against ...**  
5 Apr 2021 — There are many ways that **vitamin D** is good for you, but how about when it comes to **COVID-19**? Get the truth about the benefits of **vitamin D**.
- <https://globalnews.ca/news/covid-19-vitamin-d-link>  
**Can vitamin D lower the risk of COVID-19? Here's what we ...**  
27 Mar 2021 — There is no clear evidence that low **vitamin D** levels increase the risk of **COVID-19** illness, but some experts recommend use of supplements ...
- <https://www.medicinenet.com/article>  
**20 Vitamins and Supplements To Boost Immune Health for ...**  
27 May 2021 — Because **COVID-19** comes with cold and flu-like symptoms, **Vitamins B, C and D**, as well as zinc may be helpful in boosting your immune system and ...  
[Vitamin D](#) · [Vitamin D Deficiency Quiz](#) · [Ascorbic acid](#)
- <https://www.nature.com/articles>  
**The effect of high-dose parenteral vitamin D3 on COVID-19 ...**  
by M Güven · 2021 — In many studies, **vitamin D** has been found to be low in **COVID-19** patients. In this study, we aimed to investigate the relationship between ...

Relevance Labels

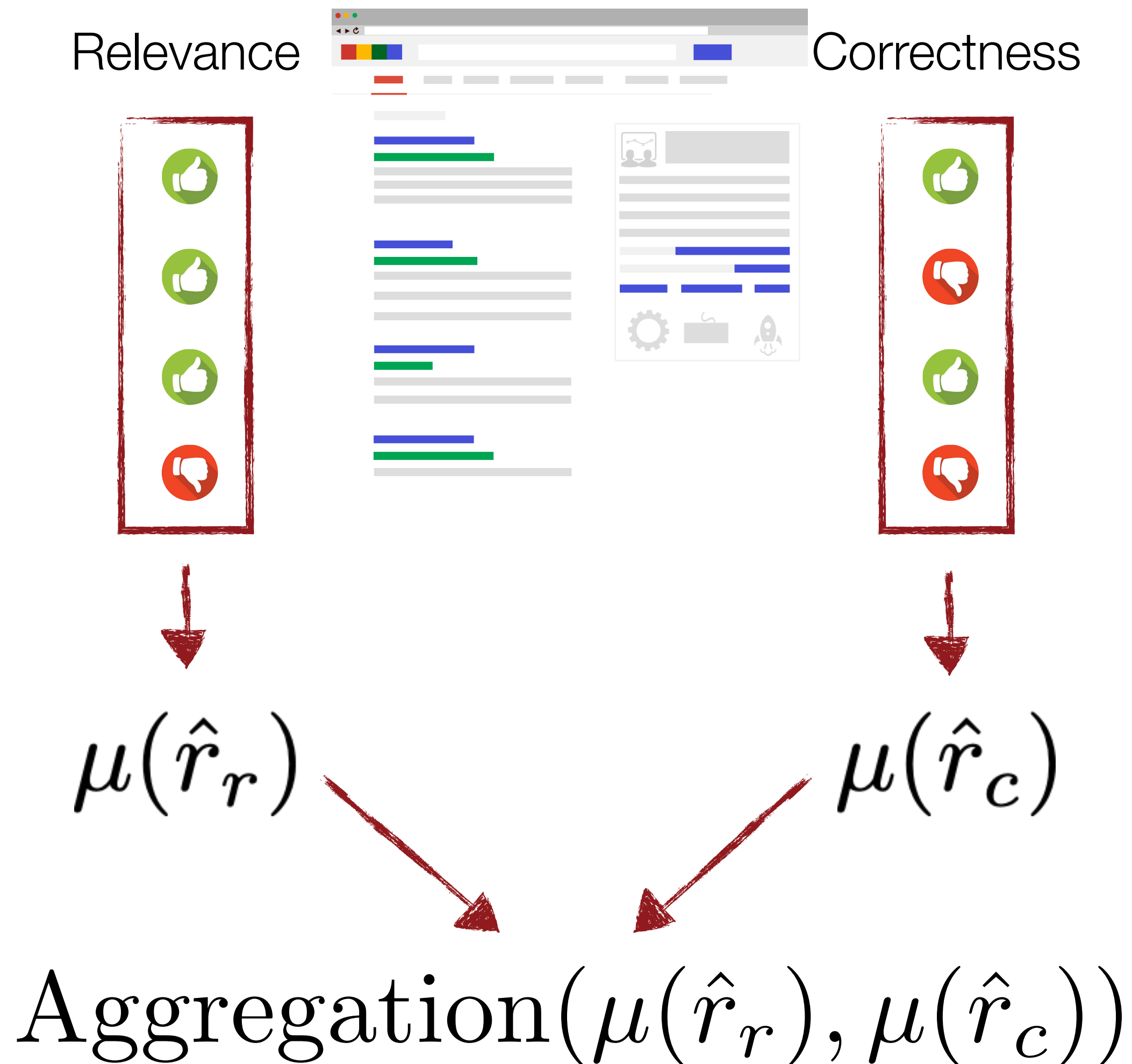


Correctness Labels



## How to account for multiple aspects in IR evaluation?

# State-of-the-art Measures: Approach 1



Measures that apply this approach:

- **Convex Aggregating Measure (CAM)**;
- **Multidimensional Measure (MM)**;
- **Weighted Harmonic Mean Aggregating Measure (WHAM)**.

J. Palotti, G. Zuccon, and A. Hanbury. 2018. MM: A New Framework for Multidimensional Evaluation of Search Engines. In CIKM 2018.

C. Lioma, J. G. Simonsen, and B. Larsen. 2017. Evaluation Measures for Relevance and Credibility in Ranked Lists. In ICTIR 2017.

# Convex Aggregating Measure (CAM)

Mean over measure scores across aspects:

$$\text{CAM}(r_t) = \sum_{a \in A} p_a \times \mu(\hat{r}_{t,a})$$

Weight for each aspect  $p_a \in [0, 1]$   $\sum_{a \in A} p_a = 1$

Any IR Evaluation Measure  
Computed for a single aspect

# Multidimensional Measure (MM)

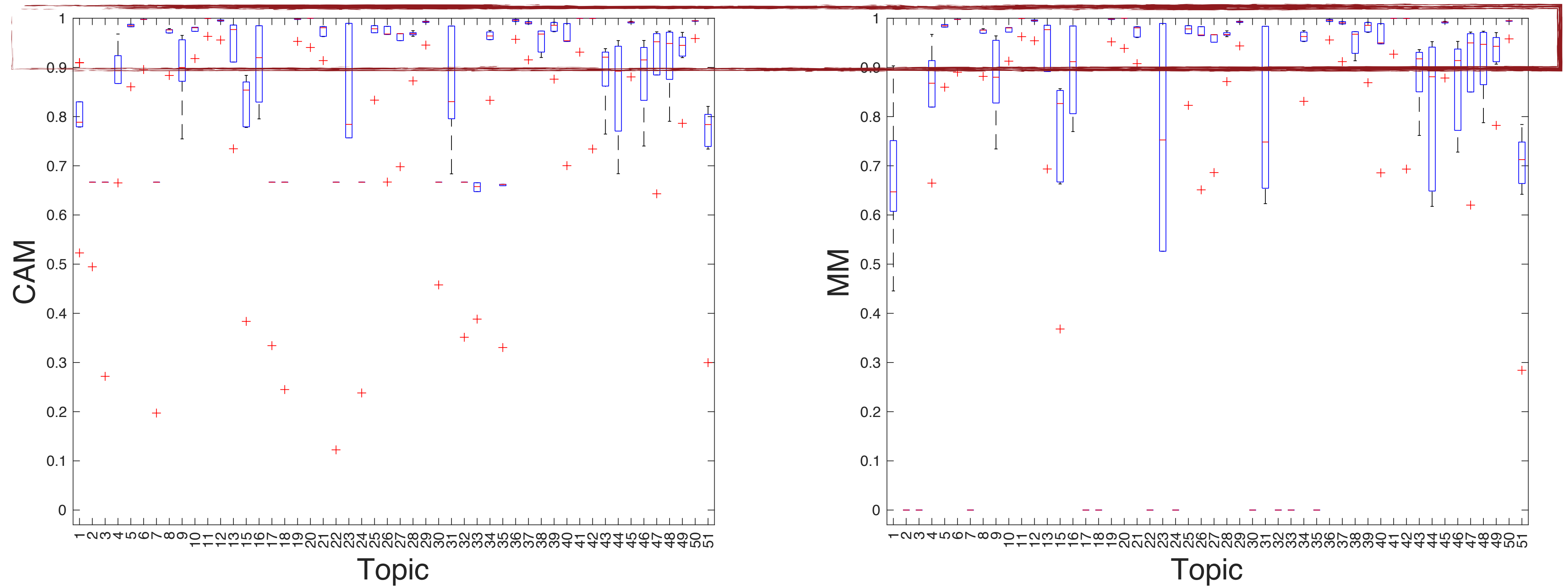
Harmonic mean over measure scores across aspects:

$$\text{MM}(r_t) = \frac{\sum_{a \in A} p_a}{\sum_{a \in A} \frac{p_a}{\mu(\hat{r}_{t,a})}}$$

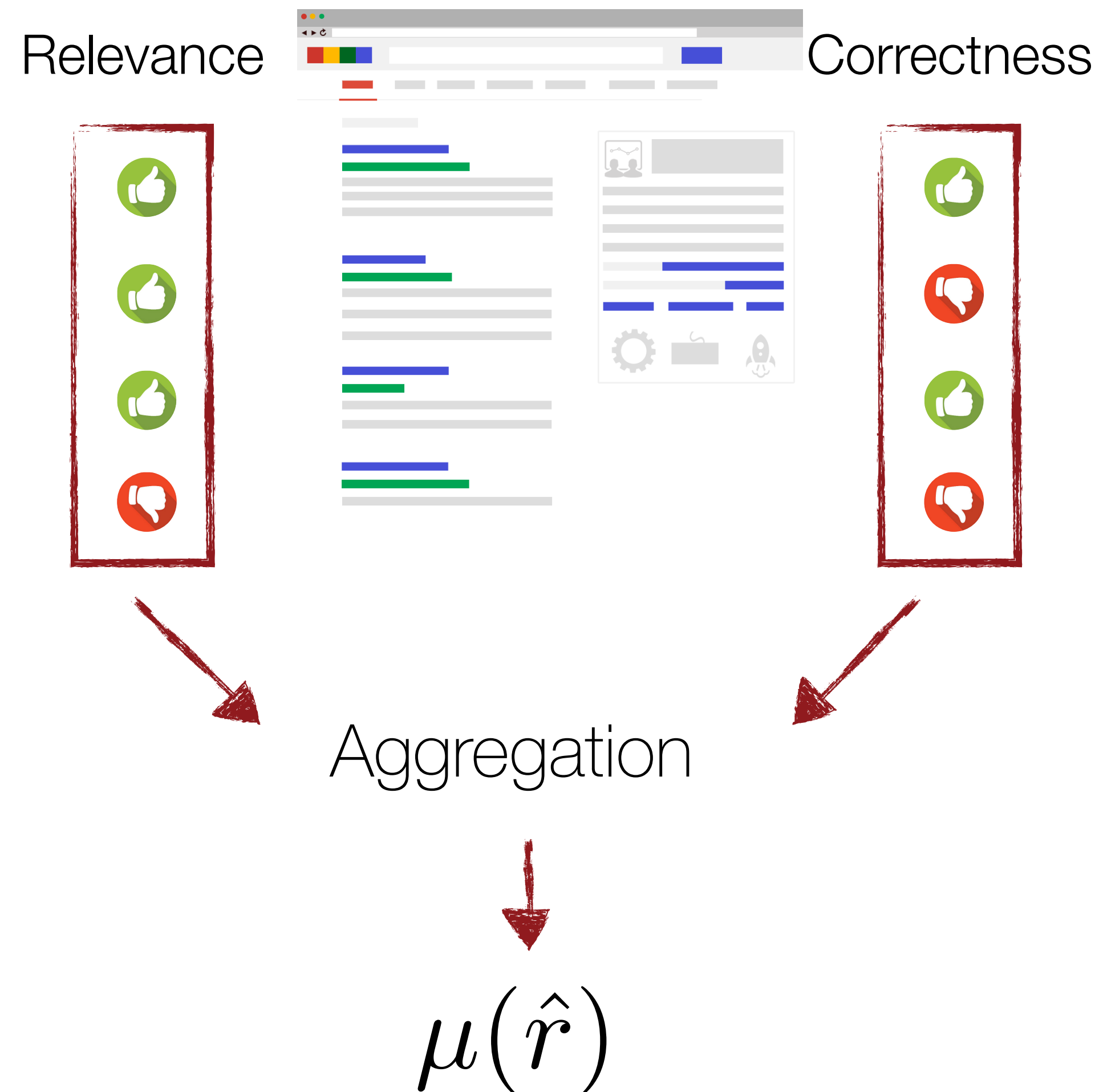
Weight for each aspect  $p_a \in [0, 1]$   $\sum_{a \in A} p_a = 1$

Any IR Evaluation Measure  
Computed for a single aspect

# Limitations of Approach 1



# State-of-the-art Measures: Approach 2



Measures that simultaneously evaluates with all aspects:

- $\alpha$ -nDCG: relevance and diversity;
- nCT: relevance, novelty and user effort;
- RBU: relevance, redundancy and user effort;
- nLRE: relevance and credibility;
- etc.

C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. 2008. Novelty and Diversity in Information Retrieval Evaluation. In SIGIR 2008.

Z. Tang and G. H. Yang. 2017. Investigating per Topic Upper Bound for Session Search Evaluation. In ICTIR '17.

E. Amigó, D. Spina, and J. Carrillo-de Albornoz. 2018. An Axiomatic Analysis of Diversity Evaluation Metrics: Introducing the Rank-Biased Utility Metric. In SIGIR 2018.

# Normalized Local Rank Error (NLRE)

- Error computed with respect to the ideal ranking for each aspect separately;
- Error = **difference in rank positions**;
- Computes the maximum possible error attainable:

$$\text{LRE} = \sum_{i=1}^{n-1} \frac{1}{\log_2(1+i)} \left( \prod_{a \in A} (p_a + \epsilon_a[d_i]) - \prod_{a \in A} p_a \right)$$

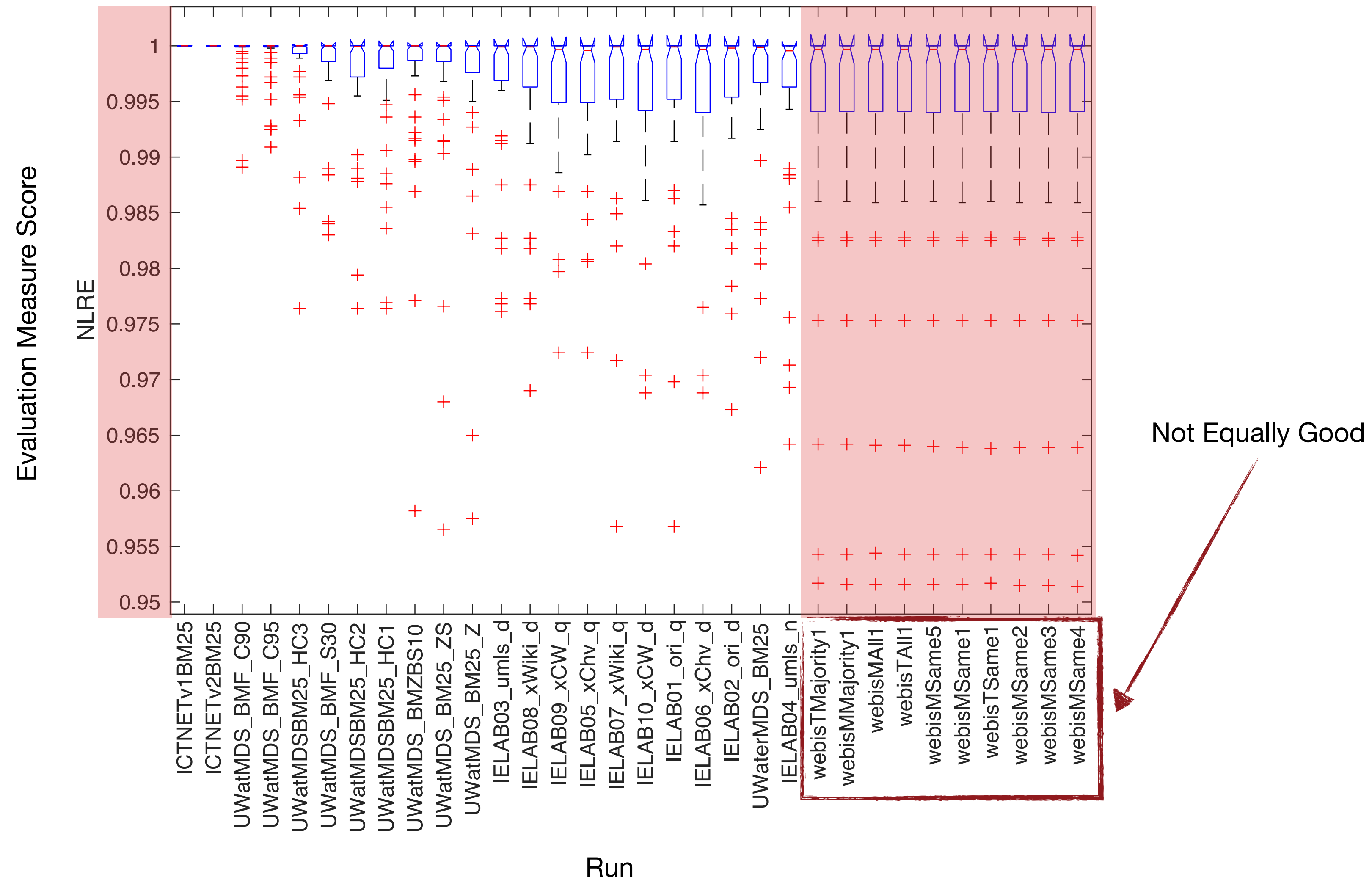
$$\text{NLRE} = 1 - \frac{\text{LRE}}{C_{\text{LRE}}}$$

Weight for each aspect  $p_a \in [0, 1]$   $\sum_{a \in A} p_a = 1$

Normalization factor: all possible permutations

# Limitations of Approach 2

- Most measures are defined for specific **contexts** and with a limited set of aspects;
- Most measures do not define a formal framework, i.e., the **ideal ranking** is not defined or it is hard to compute.







# Compatibility

- Define a **preference order** (and an ideal ranking);
- Rank Biased Overlap (RBO);
- Compute RBO between the run and the ideal ranking;
- For helpful and harmful documents independently.

Preference Value	Usefulness	Correctness	Credibility
12	Very Useful	Correct	Excellent
11	Useful	Correct	Excellent
10	Very Useful	Correct	Good
9	Useful	Correct	Good
8	Very Useful	Correct	Low or Not Judged
7	Useful	Correct	Low or Not Judged
6	Very Useful	Neutral or Not Judged	Excellent
5	Useful	Neutral or Not Judged	Excellent
4	Very Useful	Neutral or Not Judged	Good
3	Useful	Neutral or Not Judged	Good
2	Very Useful	Neutral or Not Judged	Low or Not Judged
1	Useful	Neutral or Not Judged	Low or Not Judged
0	Not Useful	Not Judged	Not Judged
-1	Very Useful or Useful	Incorrect	Low or Not Judged
-2	Very Useful or Useful	Incorrect	Good
-3	Very Useful or Useful	Incorrect	Excellent

# Problem Formalization: Partial Order



# Why a Formal Framework?

- Two aspects, relevance and correctness;
- Relevance Labels: {highly relevant, fairly relevant, marginally relevant, not relevant};
- Correctness Labels: {correct, partially correct, not correct};

$d_1$  (highly relevant, correct)

$d_2$  (marginally relevant, correct)

$d_3$  (highly relevant, partially correct)

$$d_2 \sqsubseteq d_1$$

$$d_3 \sqsubseteq d_1$$

$$d_2 \text{ ? } d_3$$

# Formalization of the Problem

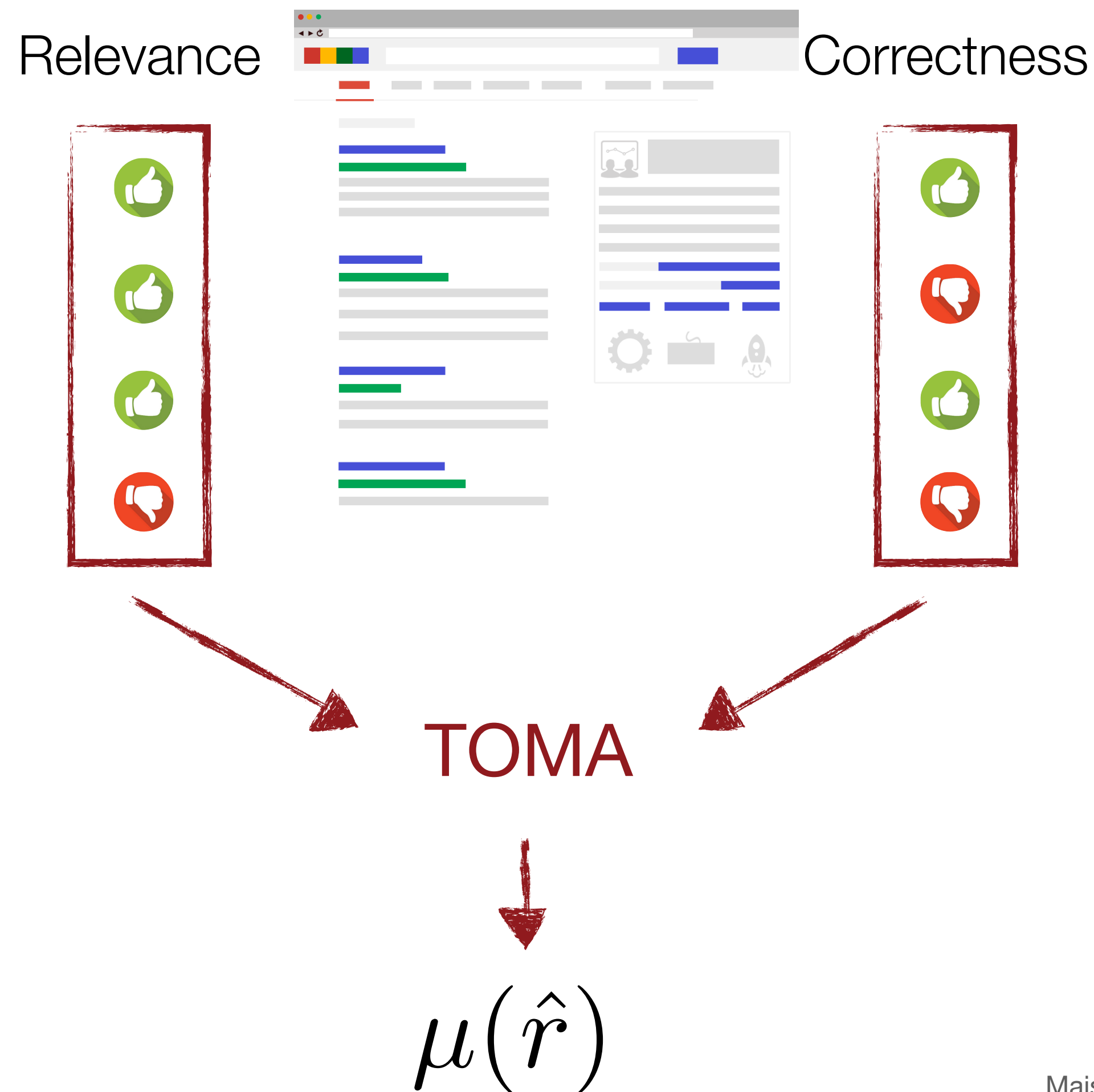
- $n$  aspects  $A = \{a_1, a_2, \dots, a_n\}$ , for example  $A = \{\text{relevance, correctness}\}$
- Set of labels  $l_0^a \prec_a l_1^a \prec_a \dots \prec_a l_{K_a}^a$ , for example non-rel  $< \dots <$  highly rel
- Ground-truth function:  $\text{GT}(d, t) = (l_1, l_2, \dots, l_n)$
- **Unequivocal** order:  $\text{GT}(d, t) \sqsubseteq \text{GT}(d', t) \iff l_i \preceq_{a_i} l'_i \quad \forall i \in \{1, \dots, n\}$
- Partial order, i.e., there are elements that are not comparable
- **Non comparable** documents:  $\text{GT}(d, t) \not\sqsubseteq \text{GT}(d', t) \quad \text{GT}(d', t) \not\sqsubseteq \text{GT}(d, t)$

$d_2$  (marginally relevant, correct)

$d_3$  (highly relevant, partially correct)

# TOMA Evaluation

# How to Complete the Partial Order?



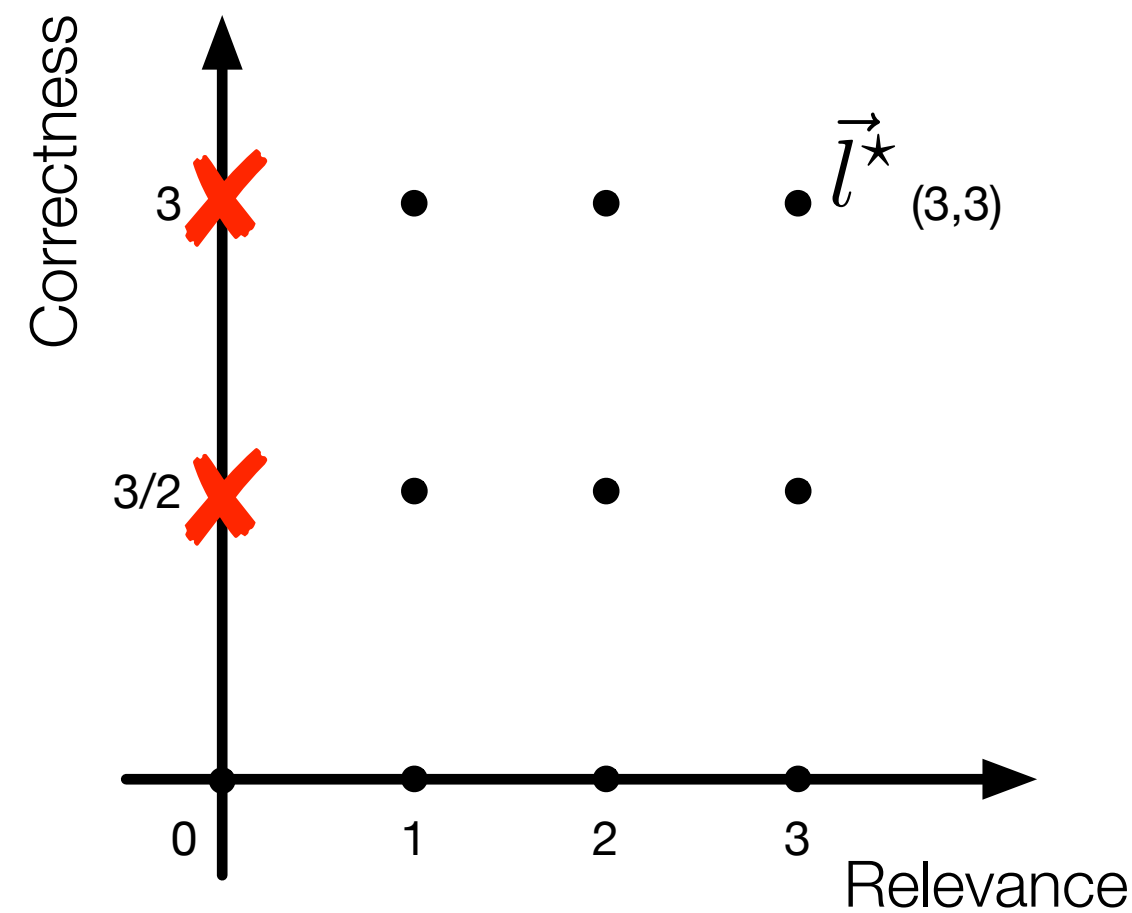
Total Order Multi-Aspect (TOMA) evaluation:

- Step 1: **embedding** in the Euclidean space;
- Step 2: the **distance** function & the **distance order**;
- Step 3: the **weight** function;

Compute any IR measure  $\mu$ .

# TOMA Example

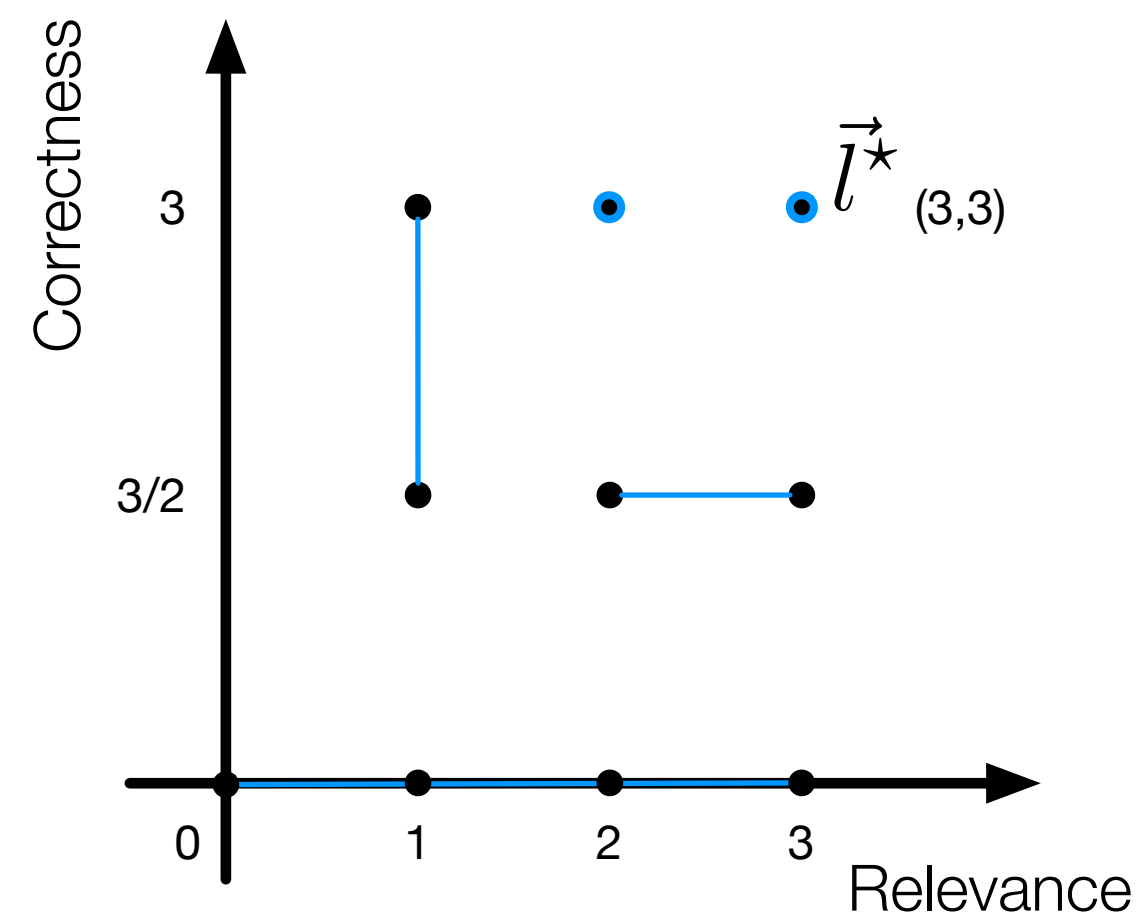
## Step 1: Embedding



**Relevance** = {not relevant < marginally relevant < fairly relevant < highly relevant}

**Correctness** = {not correct < partially correct < correct}

## Step 2: Distance Function



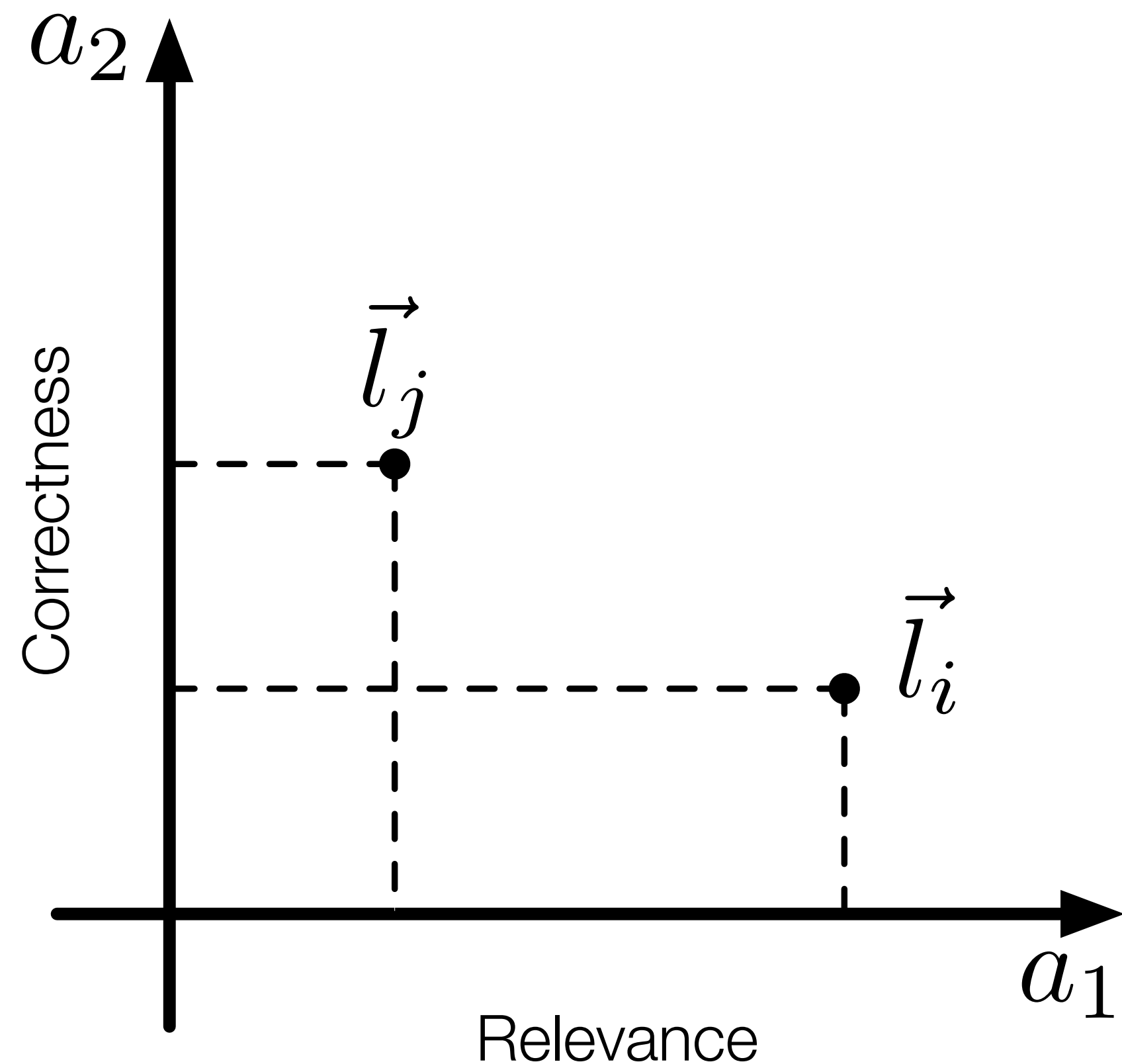
$$Dist(\vec{l}, \vec{l}^*) = \max\{|g_r(l^*) - g_r(l)|, |g_c(l^*) - g_c(l)|\}$$

## Step 3: Weight Function

$$\begin{aligned} [(3, 3)] &= \{(3, 3)\} && \mapsto 4 \\ [(2, 3)] &= \{(2, 3)\} && \mapsto 3 \\ [(3, 3/2)] &= \{(3, 3/2), (2, 3/2)\} && \mapsto 2 \\ [(1, 3)] &= \{(1, 3), (1, 3/2)\} && \mapsto 1 \\ [(0, 0)] &= \{(0, 0), (1, 0), (2, 0), (3, 0)\} && \mapsto 0 \end{aligned}$$

# TOMA Step 1

## Embedding in the Euclidean space and Distance Order



- **Embedding** function  $g$  :

$$\vec{l} = g(l_1, \dots, l_n) = (g_{a_1}(l_1), \dots, g_{a_n}(l_n))$$

- **Best label** tuple (maximum element)  $l^*$ ;
- Compute the distance from the best label:

$$l \preceq_* l' \iff \text{Dist}(\vec{l}, \vec{l}^*) \geq \text{Dist}(\vec{l}', \vec{l}^*)$$

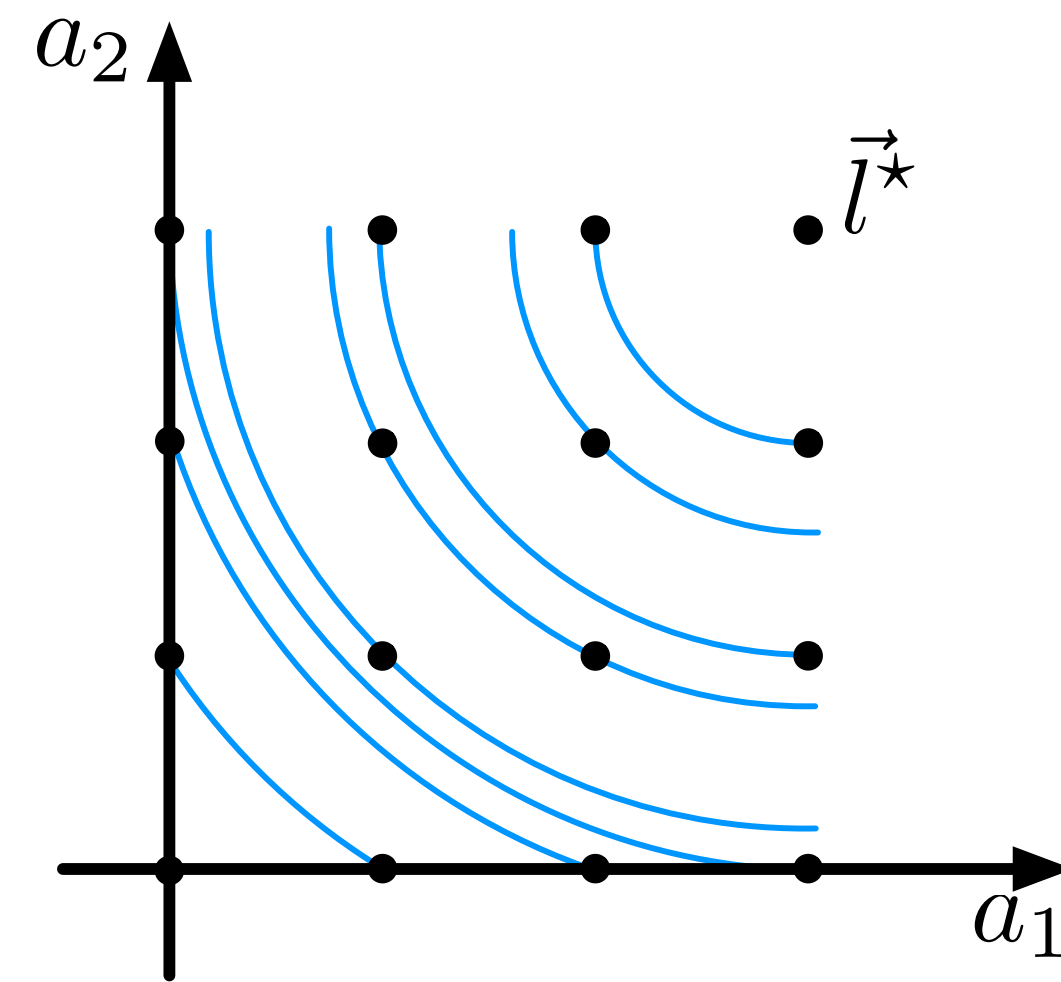
- **Distance order**  $\preceq_*$ , a **weak order** relation:

$$l =_* l' \iff \text{Dist}(\vec{l}, \vec{l}^*) = \text{Dist}(\vec{l}', \vec{l}^*)$$

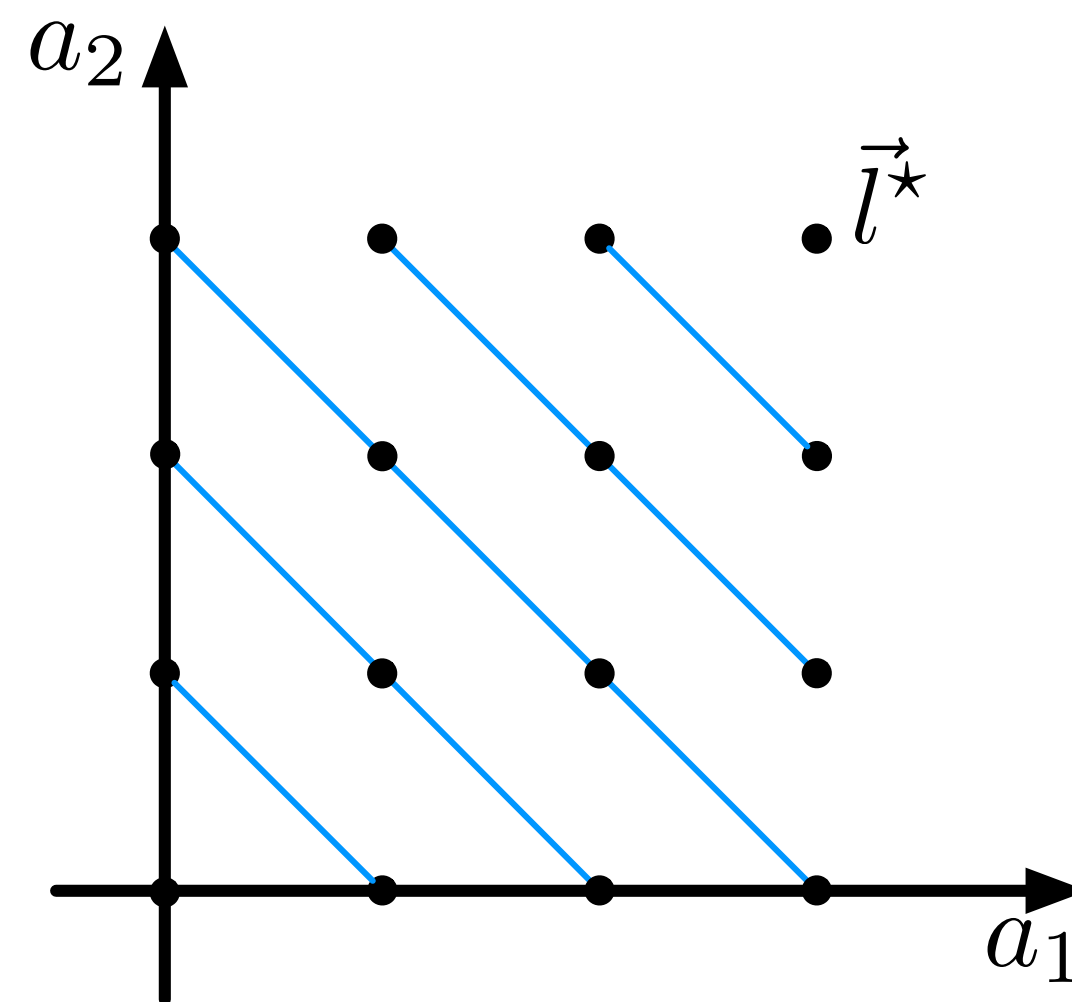


# TOMA Step 2

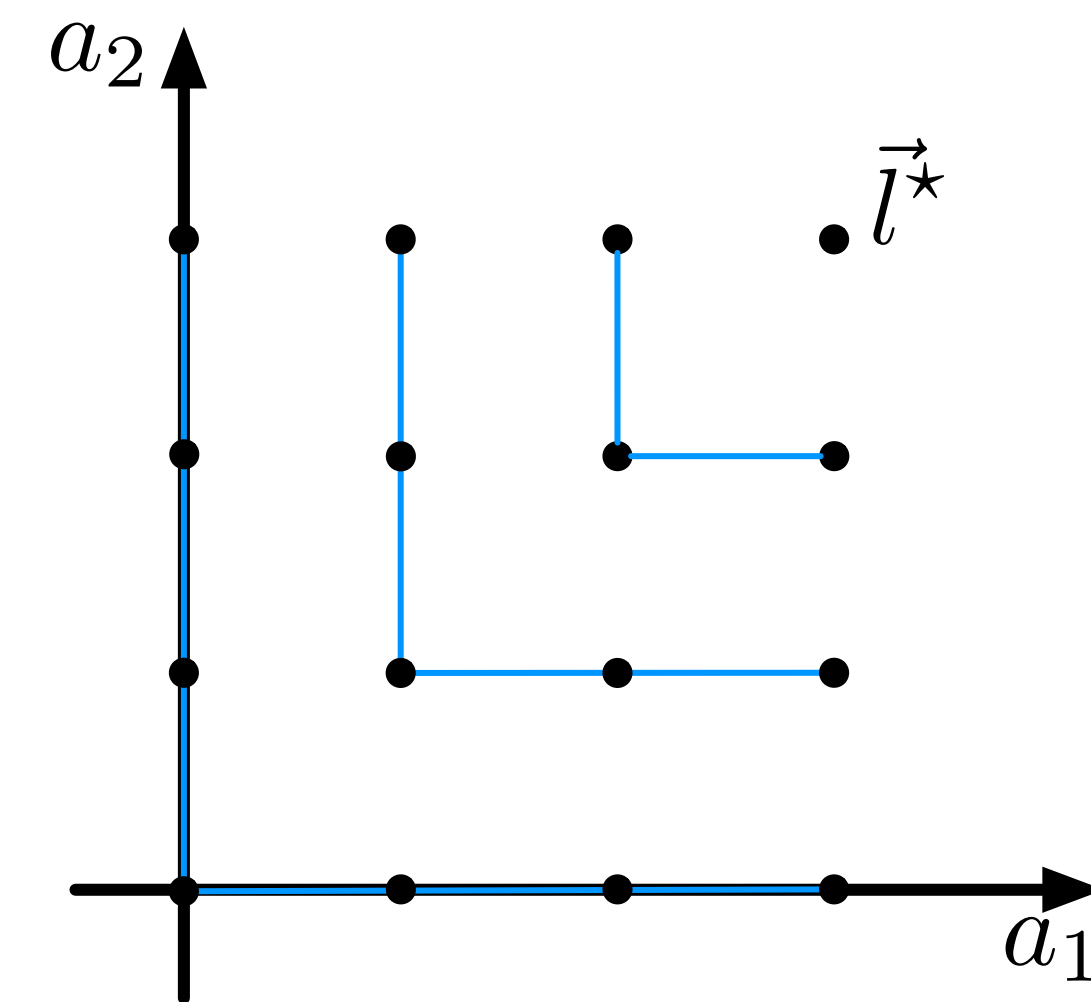
## The Distance Function



Euclidean



Manhattan



Chebyshev

- The distance order is a weak order, so we can define **equivalence classes**:

$$[\mathbf{l}]_{\star} = \{\mathbf{l}' \in L : \text{Dist}(\vec{l}', \vec{l}^*) = \text{Dist}(\vec{l}, \vec{l}^*)\}$$

# TOMA Step 3

## The Weight Function

- The **Weight Function**  $W : L \rightarrow \mathbb{N}_0^+$ , maps each equivalence class to a weight (non-negative integer);
- The distance order relation must be **preserved**:

$$\forall \mathbf{l}, \mathbf{l}' \in L : \mathbf{l} \preceq_* \mathbf{l}' \implies W(\mathbf{l}) \leq W(\mathbf{l}')$$

- Once the weight function is defined, we can compute any IR measures on the multi-aspect ranking  $r_t$ :

$$\mu \circ W : M(r_t) = \mu\left(W(\text{GT}(d_1, t)), \dots, W(\text{GT}(d_N, t))\right)$$



# Partial Order vs. Distance Order

- We started with a partial order:

$$GT(d, t) \sqsubseteq GT(d', t) \iff$$

$$l_i \preceq_{a_i} l'_i \quad \forall i \in \{1, \dots, n\}$$

- TOMA respects this partial order;
- $\forall \mathbf{l}, \mathbf{l}' \in L$  we have:

$$\mathbf{l} \sqsubseteq \mathbf{l}' \implies \mathbf{l} \preceq_* \mathbf{l}'$$

## Appendix of Principled Multi-Aspect Evaluation of Rankings

Maria Maistro  
mm@di.ku.dk  
University of Copenhagen  
Denmark

Jakob Grue Simonsen  
simonsen@di.ku.dk  
University of Copenhagen  
Denmark

Lucas Chaves Lima  
lcl@di.ku.dk  
University of Copenhagen  
Denmark

Christina Lioma  
c.lioma@di.ku.dk  
University of Copenhagen  
Denmark

### A.1 Proof 1: TOMA Respects the Partial Order

Given a set of documents  $D$ , a set of aspects  $A$ , a set of topics  $T$ , and a ground-truth map  $GT$  a partial order on the set of tuples of labels  $L = \times_{a \in A} L_a$  is defined as follows:

$$GT(d, t) \sqsubseteq GT(d', t) \iff l_i \preceq_{a_i} l'_i \quad \forall i \in \{1, \dots, n\} \quad (1)$$

where  $GT(d, t) = (l_1, \dots, l_n)$  and  $GT(d', t) = (l'_1, \dots, l'_n)$ .

Let  $g$  be an embedding function that maps tuples of labels in Euclidean space  $\mathcal{L} = \mathbb{R}^n$ :  $g(\mathbf{l}) = g(l_1, \dots, l_n) = (g_{a_1}(l_1), \dots, g_{a_n}(l_n))$ . We assume that for each  $a \in A$ ,  $g_a$  is a non-decreasing map, i.e., for any  $l, l' \in L_a$  if  $l \preceq_a l'$  then  $g_a(l) \leq g_a(l')$ . Through the embedding function  $g$ , each tuple of labels  $\mathbf{l}$  is represented by a point in the Euclidean space  $\mathcal{L}$  denoted by  $\vec{l} = g(\mathbf{l})$ .

We define the *distance order*  $\preceq_*$ : a weak order on  $L$  such that:

$$\mathbf{l} \preceq_* \mathbf{l}' \iff \text{Dist}(\vec{l}, \vec{l}^*) \geq \text{Dist}(\vec{l}', \vec{l}^*) \quad (2)$$

and by considering the absolute value:

$$|g_a(l_a) - g_a(l_a^*)| \geq |g_a(l'_a) - g_a(l_a^*)| \quad \forall a \in A \quad (8)$$

which means that  $\text{Dist}(\vec{l}, \vec{l}^*) \geq \text{Dist}(\vec{l}', \vec{l}^*)$  with Manhattan distance, i.e.  $\mathbf{l} \preceq_1 \mathbf{l}'$ .

Analogously, taking the square values in Equation (7), we obtain:

$$(g_a(l_a) - g_a(l_a^*))^2 \geq (g_a(l'_a) - g_a(l_a^*))^2 \quad \forall a \in A \quad (9)$$

which implies that  $\text{Dist}(\vec{l}, \vec{l}^*) \geq \text{Dist}(\vec{l}', \vec{l}^*)$  with Euclidean distance, i.e.  $\mathbf{l} \preceq_2 \mathbf{l}'$ .

To conclude, since  $|A| < \infty$  there exists  $\bar{a} \in A$  such that:

$$\max_{a \in A} |g_a(l'_a) - g_a(l_a^*)| = |g_{\bar{a}}(l'_{\bar{a}}) - g_{\bar{a}}(l_{\bar{a}}^*)| \quad (10)$$

by Equation (9) we have:

$$|g_{\bar{a}}(l'_{\bar{a}}) - g_{\bar{a}}(l_{\bar{a}}^*)| \leq |g_{\bar{a}}(l_{\bar{a}}) - g_{\bar{a}}(l_{\bar{a}}^*)| \leq \max_{a \in A} |g_a(l_a) - g_a(l_a^*)| \quad (11)$$

$\forall a \in A$ , which implies that  $\text{Dist}(\vec{l}, \vec{l}^*) \geq \text{Dist}(\vec{l}', \vec{l}^*)$  with Chebyshev



# Example of Usage

- Two aspects: relevance and correctness;
- Assumption: not relevant documents are not correct;
- Three different embedding functions.

Relevance	Correctness	Distance	Order among Tuples of Labels
{0, 1, 2, 3}	{0, 3/2, 3}	Euclidean	$(3, 3) \preceq_* (2, 3) \preceq_* (3, 3/2) \preceq_* (2, 3/2) \preceq_* (1, 3) \preceq_* (1, 3/2) \preceq_* (3, 0) \preceq_* (2, 0) \preceq_* (1, 0) \preceq_* (0, 0)$
		Manhattan	$(3, 3) \preceq_* (2, 3) \preceq_* (3, 3/2) \preceq_* (1, 3) \preceq_* (2, 3/2) \preceq_* (3, 0) \preceq_* (1, 3/2) \preceq_* (2, 0) \preceq_* (1, 0) \preceq_* (0, 0)$
		Chebyshev	$(3, 3) \preceq_* (2, 3) \preceq_* (3, 3/2) =_* (2, 3/2) \preceq_* (1, 3) =_* (1, 3/2) \preceq_* (3, 0) =_* (2, 0) =_* (1, 0) =_* (0, 0)$
{0, 1, 2, 3}	{0, 1, 2}	Euclidean	$(3, 2) \preceq_* (3, 1) =_* (2, 2) \preceq_* (2, 1) \preceq_* (3, 0) =_* (1, 2) \preceq_* (2, 0) =_* (1, 1) \preceq_* (1, 0) \preceq_* (0, 0)$
		Manhattan	$(3, 2) \preceq_* (3, 1) =_* (2, 2) \preceq_* (3, 0) =_* (2, 1) =_* (1, 2) \preceq_* (2, 0) =_* (1, 1) \preceq_* (1, 0) \preceq_* (0, 0)$
		Chebyshev	$(3, 2) \preceq_* (3, 1) =_* (2, 1) =_* (2, 2) \preceq_* (3, 0) =_* (2, 0) =_* (1, 0) =_* (1, 1) =_* (1, 2) \preceq_* (0, 0)$
{0, 1, 2, 3}	{0, 2, 6}	Euclidean	$(3, 6) \preceq_* (2, 6) \preceq_* (1, 6) \preceq_* (3, 2) \preceq_* (2, 2) \preceq_* (1, 2) \preceq_* (3, 0) \preceq_* (2, 0) \preceq_* (1, 0) \preceq_* (0, 0)$
		Manhattan	$(3, 6) \preceq_* (2, 6) \preceq_* (1, 6) \preceq_* (3, 2) \preceq_* (2, 2) \preceq_* (1, 2) =_* (3, 0) \preceq_* (2, 0) \preceq_* (1, 0) \preceq_* (0, 0)$
		Chebyshev	$(3, 6) \preceq_* (2, 6) \preceq_* (1, 6) \preceq_* (3, 2) =_* (2, 2) =_* (1, 2) \preceq_* (3, 0) =_* (2, 0) =_* (1, 0) =_* (0, 0)$

# Experiments



# Experimental Set-up

## Datasets

	TREC tracks										
	Web 2009	Web 2010	Web 2011	Web 2012	Web 2013	Web 2014	Task 2015	Task 2016	Decision 2019	Misinfo2020	
<b>Collection</b>	ClueWeb09				ClueWeb12				ClueWeb12-B13	CommonCrawl News	
<b>Topics</b>	50	48	50	50	50	50	35	50	50	46	
<b>Submitted runs</b>	71	56	61	48	61	30	6	9	32	51	
<b>Aspects (label grades)</b>	relevance (4) popularity† (3) non-spam‡ (3)	relevance (5*) popularity† (3) non-spam‡ (3)	relevance (4*) popularity† (3) non-spam‡ (3)	relevance (5*) popularity† (3) non-spam‡ (3)			relevance (3*) usefulness (3) popularity† (3) non-spam‡ (3)		relevance (3) credibility (2) correctness (2)	relevance (2) credibility (2) correctness (2)	

- 10 TREC Tracks, 425 runs;
- Up to 5 different aspects;
- Popularity estimated with PageRank Score<sup>1</sup>;
- Non-spamminess estimated with Waterloo Spam Ranking<sup>2</sup>.

[1] <http://www.lemurproject.org/clueweb12/PageRank.php>

[2] <https://www.mansci.uwaterloo.ca/~msmucker/cw12spam/>

# Experimental Set-up

## Baselines and TOMA Versions

- We use **CAM** and **MM** as baselines:

$$\text{CAM}(r_t) = \sum_{a \in A} p_a \times \mu(\hat{r}_{t,a})$$

$$\text{MM}(r_t) = \frac{\sum_{a \in A} p_a}{\sum_{a \in A} \frac{p_a}{\mu(\hat{r}_{t,a})}}$$

where  $p_a \in [0, 1]$  and  $\sum_{a \in A} p_a = 1$ ;

- 3 Versions of TOMA: **EUCL** (Euclidean), **MANH** (Manhattan), **CHEB** (Chebyshev);
- IR Measures: **nDCG** (graded labels when available) and **AP** (binary labels);
- All aspects are embedded into integer values with unit step 1.



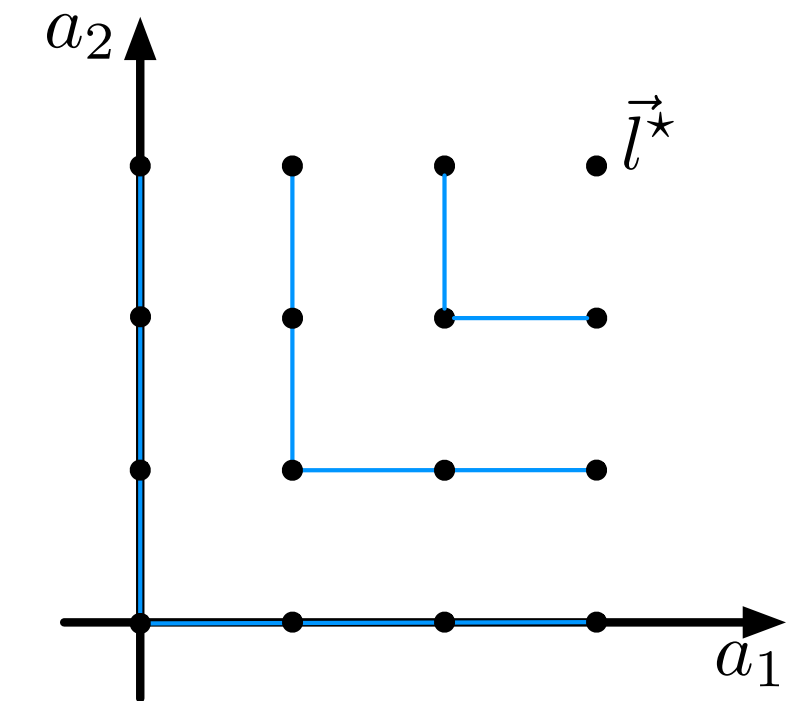
# Correlation Analysis

- Kendall's  $\tau$  between the Rankings of Submitted runs (RoS) generated by 2 different measures;
- Most correlated RoS:
  - EUCL - MANH ( $\tau$  up to 1);
  - (EUCL, MANH) - CAM ( $\tau$  up to 0.76);
  - CHEB - MM ( $\tau$  up to 0.88).
- EUCL and MANH may be more similar to CAM (mean), while CHEB may be more similar to MM (harmonic mean).



# Discriminative Power

- Bootstrap Samples: 10'000
- Paired Bootstrap Hypothesis Test, with  $\alpha = 0.01$ ;
- 16/20 times either MANH (12/20) or EUCL (6/20) is best.



	WEB2009		WEB2010		WEB2011		WEB2012		WEB2013		WEB2014		TASK15		TASK16		DECISION19		MISINFO 2020	
	NDCG	AP	NDCG	AP	NDCG	AP	NDCG	AP	NDCG	AP	NDCG	AP	NDCG	AP	NDCG	AP	NDCG	AP	NDCG	AP
DISCRIMINATIVE POWER OF MEASURES																				
CAM	75.98	64.43	66.32	61.23	75.14	61.64	<b>68.71</b>	56.74	76.89	57.05	85.06	78.85	53.33	33.33	72.22	55.56	72.58	70.56	71.53	70.90
MM	75.61	50.58	<b>72.89</b>	<b>67.79</b>	67.32	67.81	62.68	56.12	<b>80.71</b>	46.99	74.25	53.56	0.00	0.00	0.00	0.00	60.08	53.23	68.31	62.20
EUCL	75.29	72.64	62.96	66.75	75.14	70.33	66.13	64.10	75.14	<b>59.45</b>	80.92	78.85	<b>66.67</b>	<b>66.67</b>	69.44	<b>75.00</b>	73.59	<b>73.99</b>	72.86	<b>75.14</b>
MANH	<b>76.66</b>	<b>72.68</b>	63.59	67.14	<b>77.32</b>	<b>70.38</b>	66.05	<b>64.18</b>	76.67	59.34	<b>86.44</b>	<b>79.08</b>	<b>66.67</b>	53.33	<b>75.00</b>	<b>75.00</b>	<b>73.79</b>	73.79	<b>73.02</b>	74.98
CHEB	50.18	6.32	59.82	51.49	73.06	50.11	61.08	39.36	77.10	49.34	75.17	66.21	0.00	0.00	0.00	0.00	42.54	29.84	65.41	53.33



# Document Quality

Run that has been assessed as the best best per {topic, track, year}, on a per query basis;

- With cut-off 5, **zero-Aspect**: # documents with 0 as sum of labels;
- **Average** per label: average across aspects.

Zero-Aspect Documents

Rank	CAM	MM	EUCL	MANH	CHEB
1	51 (1.18%)	131 (3.02%)	39 (0.90%)	<b>33 (0.76%)</b>	154 (3.55%)
2	65 (1.50%)	159 (3.67%)	50 (1.15%)	<b>48 (1.11%)</b>	179 (4.13%)
3	103 (2.38%)	202 (4.66%)	88 (2.03%)	<b>78 (1.80%)</b>	185 (4.17%)
4	102 (2.35%)	173 (3.99%)	86 (1.99%)	<b>74 (1.71%)</b>	183 (4.23%)
5	107 (2.47%)	196 (4.53%)	95 (2.19%)	<b>81 (1.87%)</b>	205 (4.73%)
1-5	428 (9.88%)	861 (19.88%)	358 (8.27%)	<b>314 (7.25%)</b>	906 (20.92%)

Average per Label

Ranks	CAM	MM	EUCL	MANH	CHEB
1-25	<b>1.70</b>	1.49	1.67	1.69	1.39
26-50	0.85	0.78	0.91	<b>0.94</b>	0.70
51-75	0.57	0.53	0.63	<b>0.64</b>	0.48
76-100	0.40	0.39	0.43	<b>0.44</b>	0.36



# Limitations

- **Arbitrary choices** of the embedding, distance, and weight functions → impact of each choice;
- Embedding function: mapping from nominal/ordinal to interval/ratio **scale** → **theoretical** properties of TOMA;
- Analysis of the **interactions** between aspects and/or documents → proper embedding;
- Large number of **equivalence classes** might be a problem for gain based measures → non-gain based measures.



# Wrapping up

Total Order Multi-Aspect (TOMA):

- Defined for any **number** and **type** of aspect;
- Aspects can have different **gradings**;
- Choose a relative **importance weighting** for different aspects;
- Integration with any existing single-aspect **IR measure**;
- Better **discriminative power**;
- Better at rewarding **high quality** documents.

# Conclusions and Future Work



# In Today's Talk

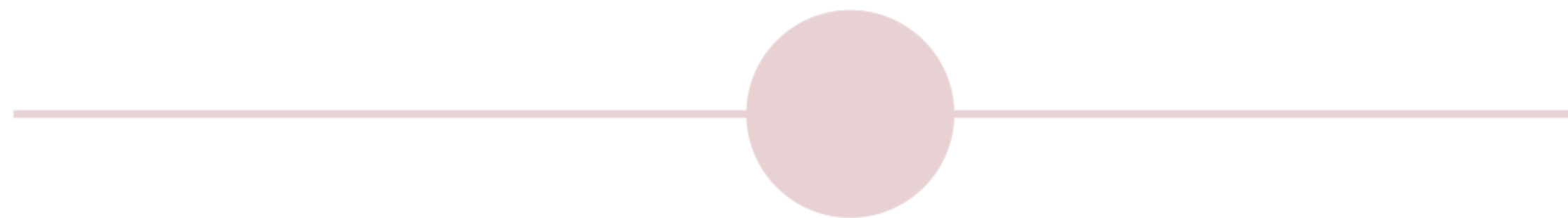
- TREC Health Misinformation track;
- Challenges in **evaluating misinformation**: find good topics, collect judgements, datasets with enough noise, etc.
- Multi-aspect evaluation and existing measures;
- Total Order Multi-Aspect (TOMA): defined for any **number** and **type** of aspect with different **gradings** integrated with any existing single-aspect **IR measure**.

# Future Work

- TREC Trustworthy AI Conference (TRUC);
- Formal **properties** of evaluation measures;
- Measures that **penalise** systems that retrieve harmful documents;
- Other aspects, e.g., **fairness**;
- Use these measures for **learning to rank**;
- Relation to with **bias**;
- **Ethical** issues.

**Thank You!**

**Questions?**

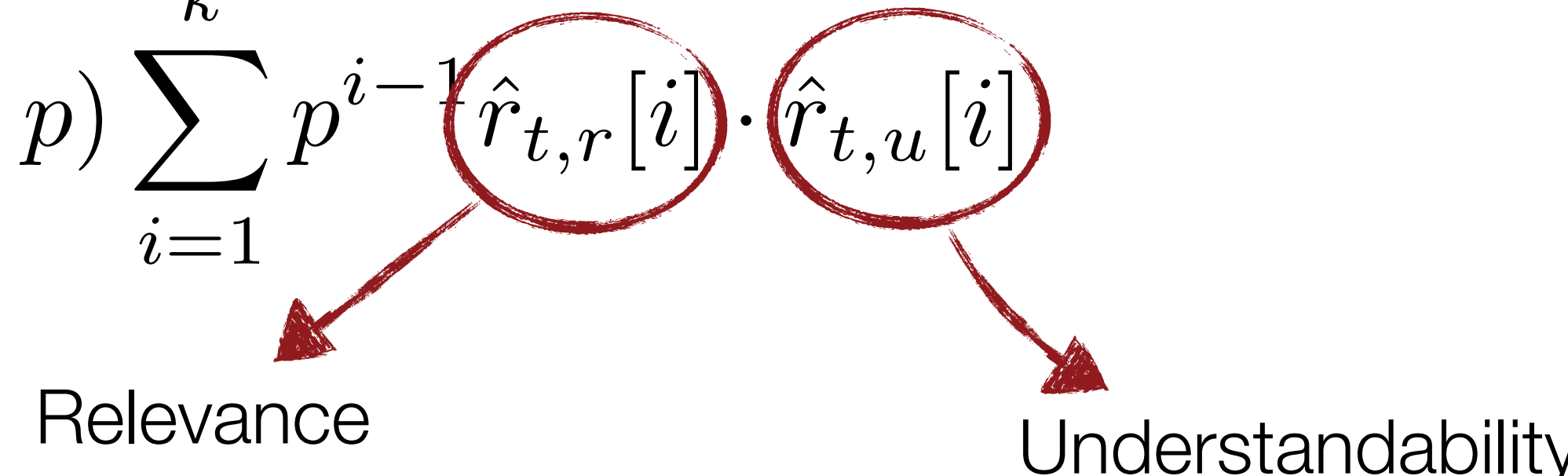




# Ongoing Work

# Relation with Other Measures - uRBP

- Understandability Biased RBP (uRBP):

$$\text{uRBP}@k(r_t) = (1 - p) \sum_{i=1}^k p^{i-1} \hat{r}_{t,r}[i] \cdot \hat{r}_{t,u}[i]$$


Relevance

Understandability

- Within TOMA framework:

$$[\mathbf{1}] = \left\{ \mathbf{l}' \in L : \prod_{a \in A} g_a(\mathbf{l}') = \prod_{a \in A} g_a(\mathbf{1}) \right\}$$



# Relation with Other Measure - Compatibility

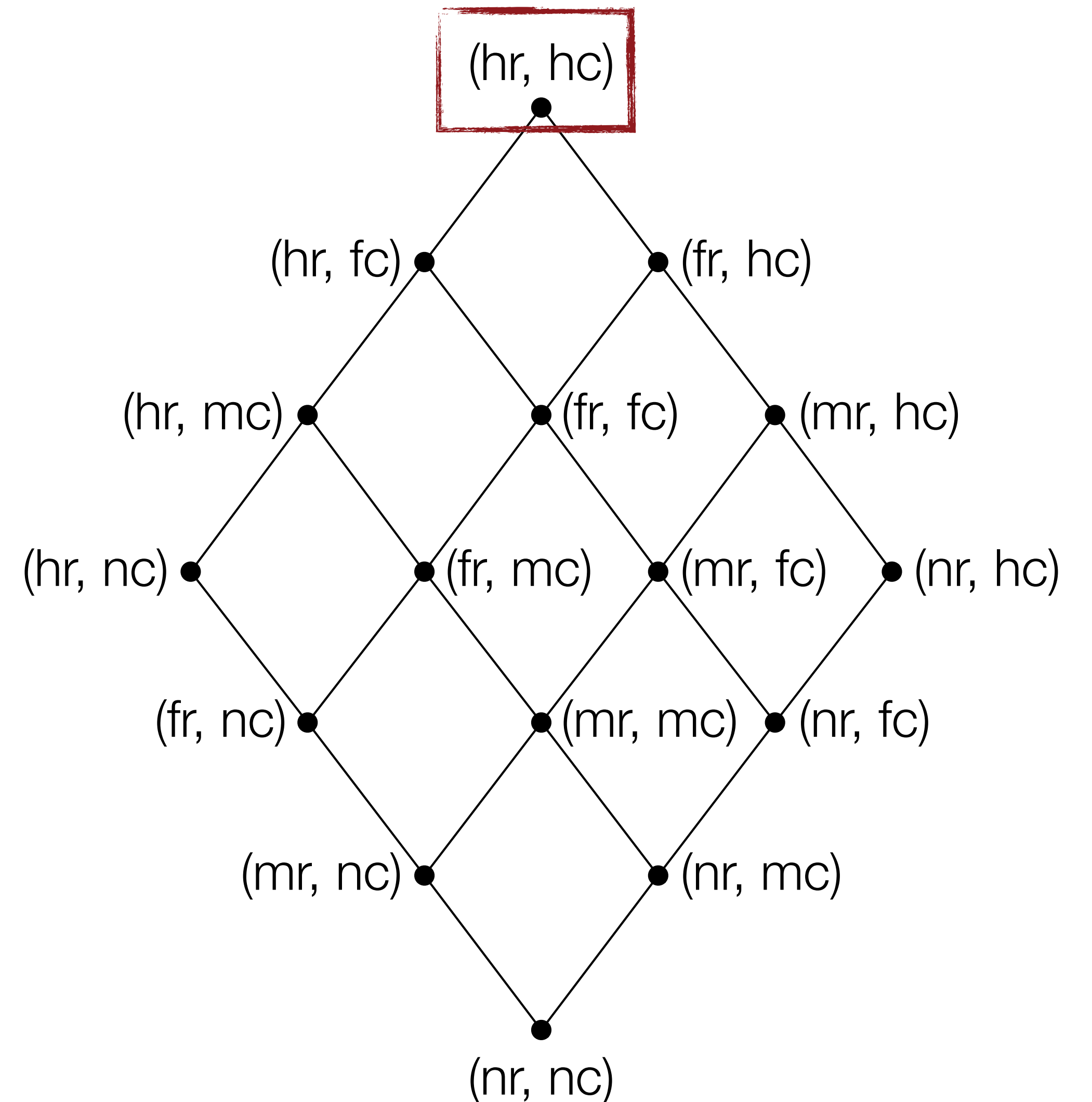
- Compute Compatibility:
  1. Define an ideal ranking or set of equivalent ideal rankings;
  2. Compute RBO between the input ranking and ideal ranking;

$$\text{RBO}(r_t, i_t) = (1 - p) \sum_{i=1}^{+\infty} p^{i-1} \cdot \frac{|r_t[1, i] \cap i_t[1, i]|}{i}$$

- Use TOMA to define **levels of effectiveness** ~ equivalence classes.

# Poset Theory

- **Poset**: partially ordered set;
- Set of multi-aspect labels is a poset.



hr highly relevant

hc highly correct

fr fairly relevant

fc fairly correct

mr marginally relevant

mc marginally correct

nr not relevant

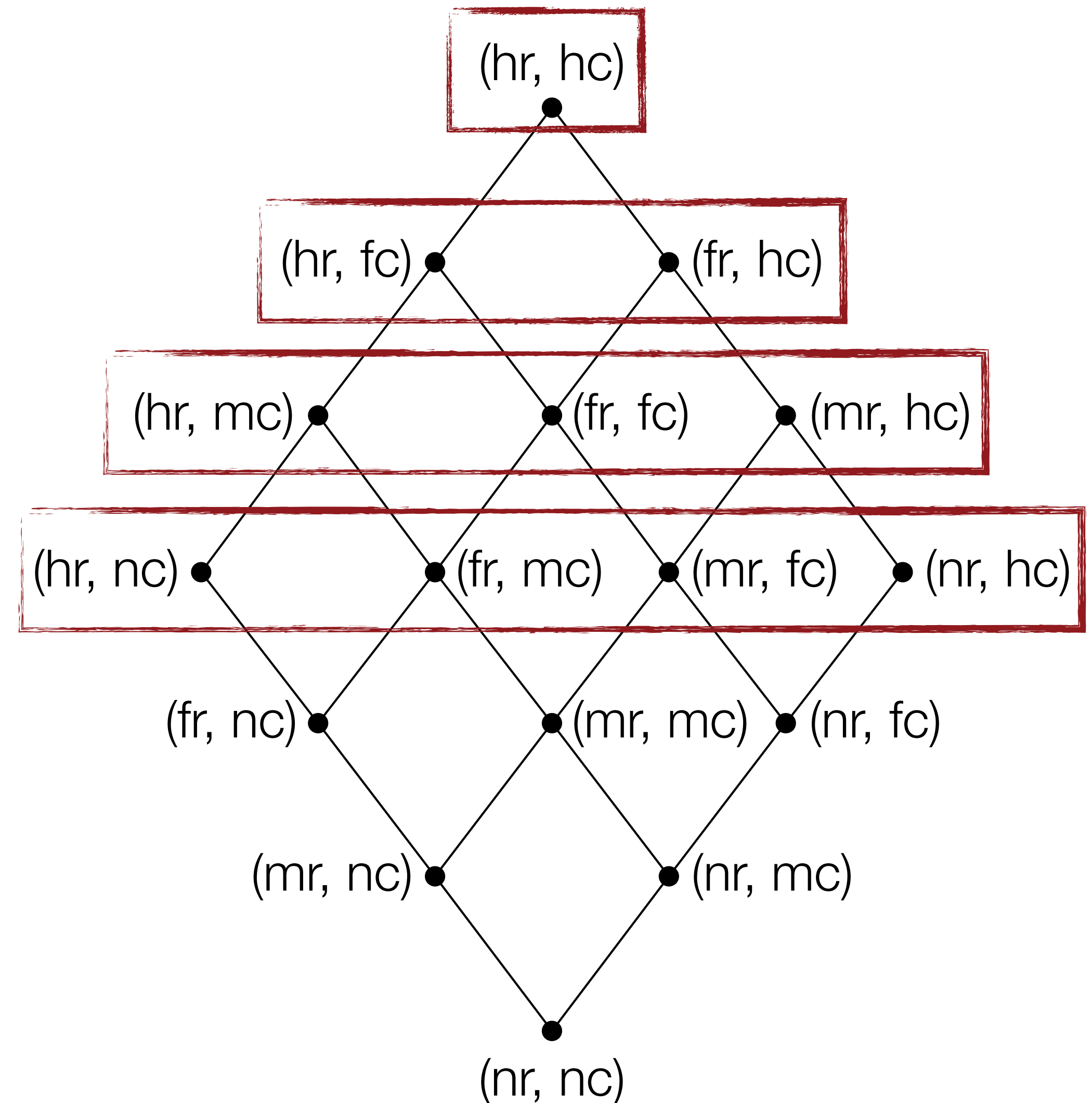
nc not correct

# Skyline Order

- **Skyline set:** set of label vectors that are not worse than any other label vectors in all the dimensions;

$$l_a \preceq l'_a \quad \forall a \in A \quad \text{and} \quad \exists a_k \in A: l_{a_k} \prec l'_{a_k}$$

- $S_0 = \{(hr, hc)\}$
- $S_1 = \{(hr, fc), (fr, hc)\}$
- $S_2 = \{(hr, nc), (fr, mc), (mr, fc), (nr, hc)\}$
- ...





# Theoretical Properties

- Utility oriented measures of retrieval effectiveness satisfy 2 properties:
- **Replacement**: if we replace a less relevant document with a more relevant one in the same rank position, the measure score should not decrease.
- **Swap**: if we swap a less relevant document in a higher rank position with a more relevant one in a lower rank position, the measure score should not decrease.

Relevance Labels



vitamin D covid 19



All News Images Videos Shopping More Tools

About 728.000.000 results (0,55 seconds)

https://www.bbc.com › news › health-56180921

**Vitamin D: The truth about an alleged Covid 'cover-up' - BBC**

4 Apr 2021 — As **Covid-19** swept the world, so did misinformation about how to treat it. But sometimes misinformation can develop even around ideas that ...

https://www.wcax.com › content › news › Research-sugge...

**Research suggests vitamin D plays role in COVID-19 death ...**

11 May 2020 — **Vitamin D** is key for maintaining healthy bones, and not having enough can affect the immune system and inflammation. Now, new research from ...

https://www.henryford.com › vitamin-d-and-covid-19

**Will Boosting Your Vitamin D Intake Help Protect Against ...**

5 Apr 2021 — There are many ways that **vitamin D** is good for you, but how about when it comes to **COVID-19**? Get the truth about the benefits of **vitamin D**.

https://globalnews.ca › news › covid-19-vitamin-d-link

**Can vitamin D lower the risk of COVID-19? Here's what we ...**

27 Mar 2021 — There is no clear evidence that low **vitamin D** levels increase the risk of **COVID-19** illness, but some experts recommend use of supplements ...

https://www.medicinenet.com › article

**20 Vitamins and Supplements To Boost Immune Health for ...**

27 May 2021 — Because **COVID-19** comes with cold and flu-like symptoms, **Vitamins B, C and D**, as well as zinc may be helpful in boosting your immune system and ...

[Vitamin D](#) · [Vitamin D Deficiency Quiz](#) · [Ascorbic acid](#)

https://www.nature.com › ... › articles

**The effect of high-dose parenteral vitamin D3 on COVID-19 ...**

by M Güven · 2021 — In many studies, **vitamin D** has been found to be low in **COVID-19** patients. In this study, we aimed to investigate the relationship between ...

# Theoretical Properties - Multi-aspect Measures

- Partial order among labels can be used instead of “more relevant” and “less relevant”;

$$GT(d, t) \sqsubset GT(d', t) \iff$$

$$l_a \preceq_a l'_a \quad \forall a \in A$$

$$\text{and } \exists \bar{a} \in A: l_{\bar{a}} \prec_{\bar{a}} l'_{\bar{a}}$$

- TOMA satisfies swap and replacements → extend findings for IR measures.

Cor	Rel

Google search results for "vitamin D covid 19".

Search results include:

- Vitamin D: The truth about an alleged Covid 'cover-up' - BBC** (4 Apr 2021) - Cor: thumbs up, Rel: thumbs up
- Research suggests vitamin D plays role in COVID-19 death ...** (11 May 2020) - Cor: thumbs down, Rel: thumbs up
- Will Boosting Your Vitamin D Intake Help Protect Against ...** (5 Apr 2021) - Cor: thumbs down, Rel: thumbs up
- Can vitamin D lower the risk of COVID-19? Here's what we ...** (27 Mar 2021) - Cor: thumbs up, Rel: thumbs up
- 20 Vitamins and Supplements To Boost Immune Health for ...** (27 May 2021) - Cor: thumbs down, Rel: thumbs down
- The effect of high-dose parenteral vitamin D3 on COVID-19 ...** (2021) - Cor: thumbs up, Rel: thumbs up