

*Just Another Way of Information Retrieval*

# Penggunaan Hyperdimensional Computing dalam Implementasi Search Engine



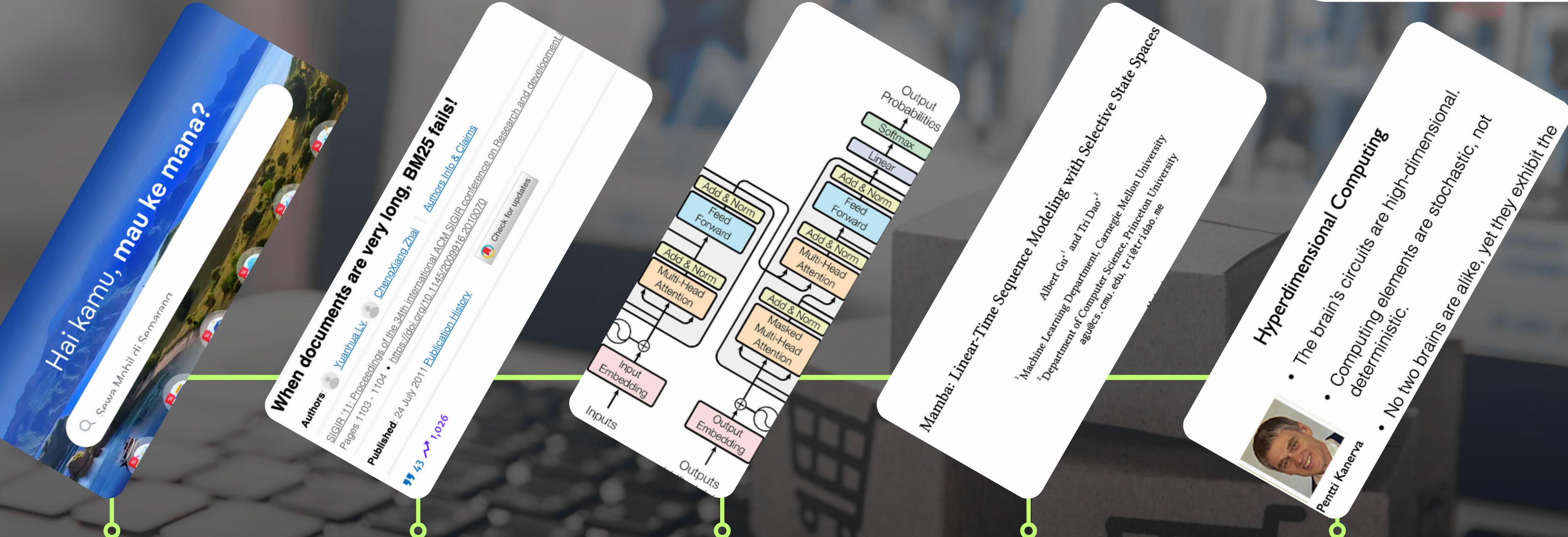
Yosef Nuraga Wicaksana



Vian Sebastian Bromokusumo



Louis Widi Anandaputra



E-commerce telah menjadi bagian penting masyarakat modern.

Metode konvensional tidak bisa menangkap konteks.

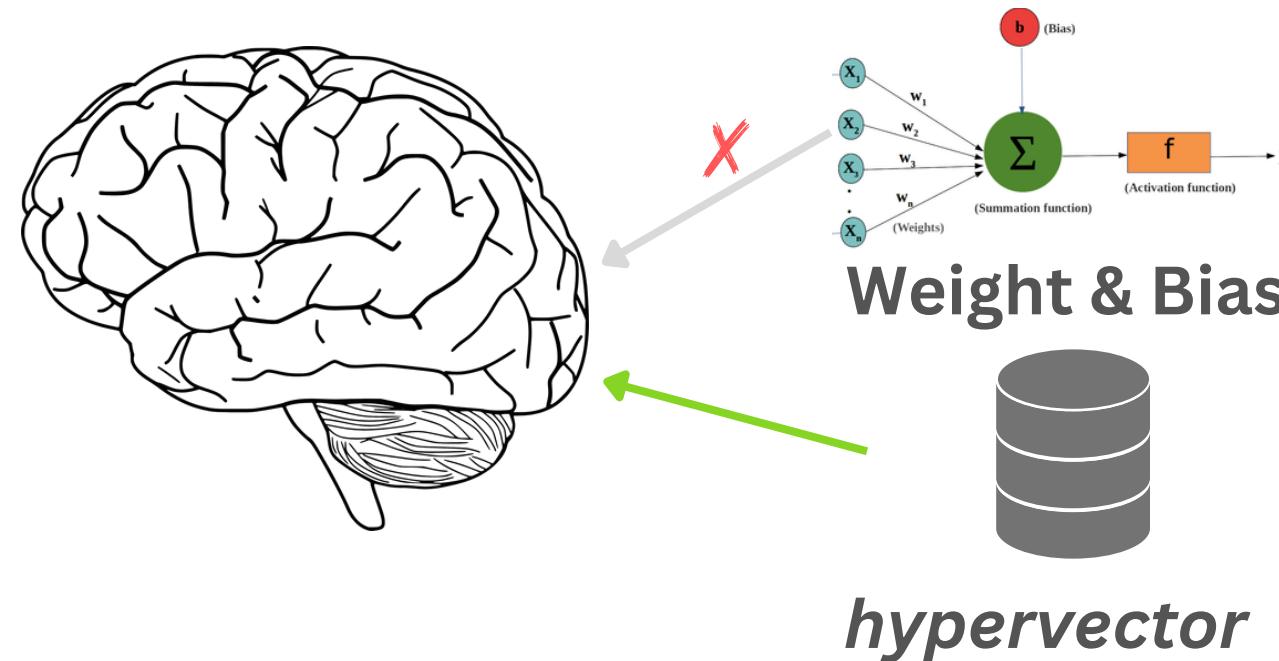
Transformer memerlukan komputasi yang berat.

Metode alternatif yang lebih ringan seperti MAMBA masih sangat kompleks.

*Hyperdimensional Computing* dipopulerkan oleh Kanerva pada 2009 sebagai alternatif *neural net* yang efisien

# Memperkenalkan! **Fast Vector Symbolic Search (FastVSS)!**

## Landasan Teori - Hyperdimensional Computing



Termotivasi dengan otak manusia yang mampu dalam **menyimpan** dan **menggabungkan** konsep **tanpa melupakan** konsep pembentuknya.

HDC memiliki perbedaan dengan *neural network*. Alih-alih menyimpan cara komputasi (weight & bias), HDC **menyimpan data** itu sendiri dalam representasi ruang vektor (menggunakan *hypervector/HV*).

Menggunakan operasi aljabar sederhana : **Adisi (+)**, **Multiplikasi (\*)**, dan **Permutasi ( $\Pi$ )**

### Contoh kasus- representasi konsep “Susu dari Solo”

Jika diasumsikan memiliki *HV-minuman* dan *HV-asal*, namun **tidak ada *HV-susu* dan *HV-solo***, maka dilakukan inisialisasi berikut

$$HV_{susu} = \Pi(\Pi HV_S + HV_U) + HV_S + HV_u$$

$$HV_{solo} = \Pi(\Pi(HV_S + HV_O) + HV_L) + HV_O$$

Selanjutnya, dilakukan penghubungan **variabel** dengan **objek observasi** menggunakan operasi **multiplikasi**:

$$HV_{minuman\ adalah\ susu} = HV_{minuman} * HV_{susu}$$

$$HV_{asal\ adalah\ solo} = HV_{asal} * HV_{solo}$$

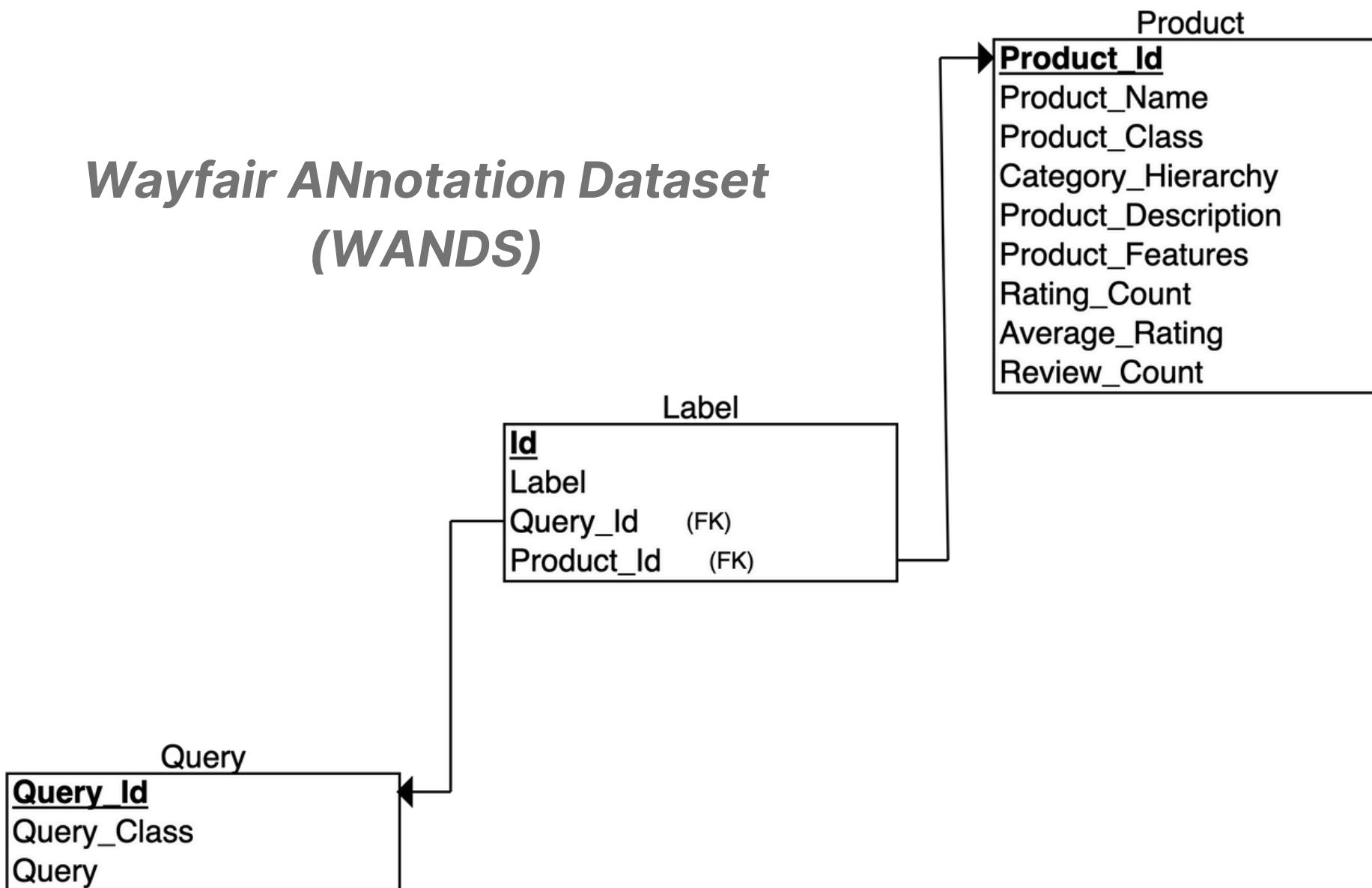
Terakhir, **dibentuk ruang vektor** representasi dari “susu dari solo” dengan operasi **adisi**, menggambarkan agregasi antara dua *HV*

$$HV_{susu\ dari\ solo} = HV_{minuman\ adalah\ susu} + HV_{asal\ adalah\ solo}$$

## Metodologi - Data dan Preprocessing

Preprocessing yang akan dilakukan relatif sederhana dan straightforward, yaitu Missing Values Imputation, Cleaning, dan Label Encoding.

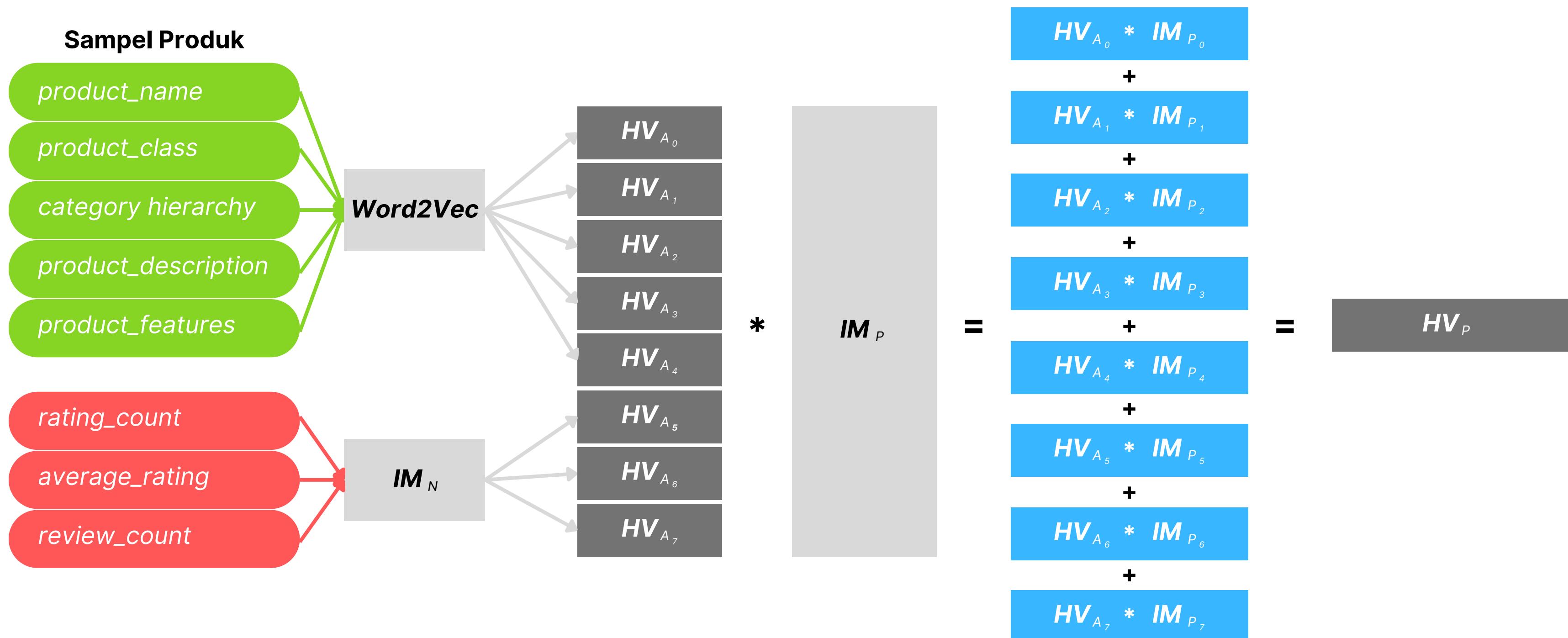
### Wayfair ANnotation Dataset (WANDS)



- **Missing Values Imputation**
  - Imputasi “ ” (empty string) pada data teks
  - Imputasi 0 pada data numerik
- **Cleaning**
  - Menghilangkan kata/huruf non-alfanumerik
  - Menghilangkan kata penghubung (konjungsi)
  - Lemmatization
  - Tokenization
- **Label Encoding**
  - Encoding Label dengan gain

## Metodologi - Vektorisasi Produk

Vektorisasi dengan Word2Vec yang telah dilatih oleh kami dilakukan dengan melakukan rata-rata dari  $HV$  representasi tiap kata



## Metodologi - Vektorisasi Query Class

Pembentukan konsep *query class* dilakukan dengan agregasi adisi untuk tiap kelompok kelas

$$\begin{array}{c}
 \text{Class 1} \\
 \begin{array}{c} \mathbf{HV}_{P_0} \\ \hline \mathbf{HV}_{P_1} \\ \hline \mathbf{HV}_{P_2} \end{array} = \begin{array}{c} \mathbf{HV}_{P_0} \\ + \\ \mathbf{HV}_{P_1} \\ + \\ \mathbf{HV}_{P_2} \end{array} = \mathbf{HV}_{C_{\text{class 1}}} \\
 \\ \\
 \text{Class 2} \\
 \begin{array}{c} \mathbf{HV}_{P_3} \\ \hline \mathbf{HV}_{P_4} \end{array} = \begin{array}{c} \mathbf{HV}_{P_3} \\ + \\ \mathbf{HV}_{P_4} \end{array} = \mathbf{HV}_{C_{\text{class 2}}}
 \end{array}$$

**Kuantisasi** dilakukan sebagai tahap akhir untuk memastikan tiap elemen berada di rentang nilai yang sama dengan memanfaatkan fungsi Tanh.

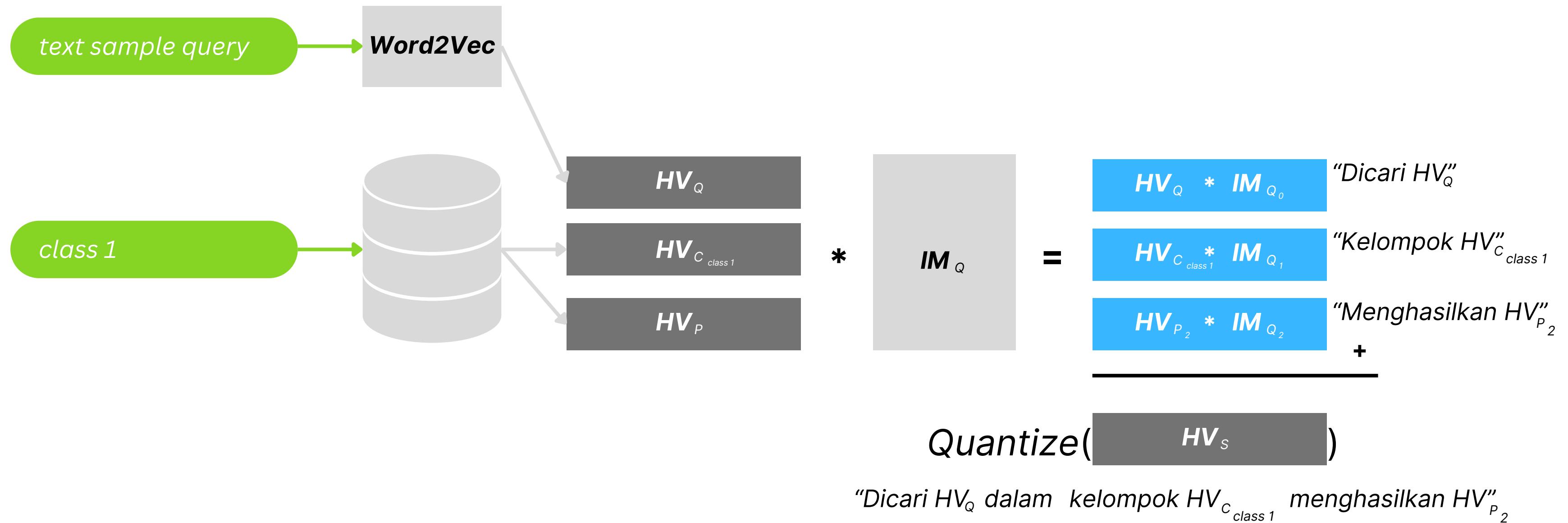
*Quantize(  $\mathbf{HV}_P$  )*

*Quantize(  $\mathbf{HV}_{C_{\text{class 1}}}$  )*

*Quantize(  $\mathbf{HV}_{C_{\text{class 2}}}$  )*

## Metodologi - Vektorisasi Query

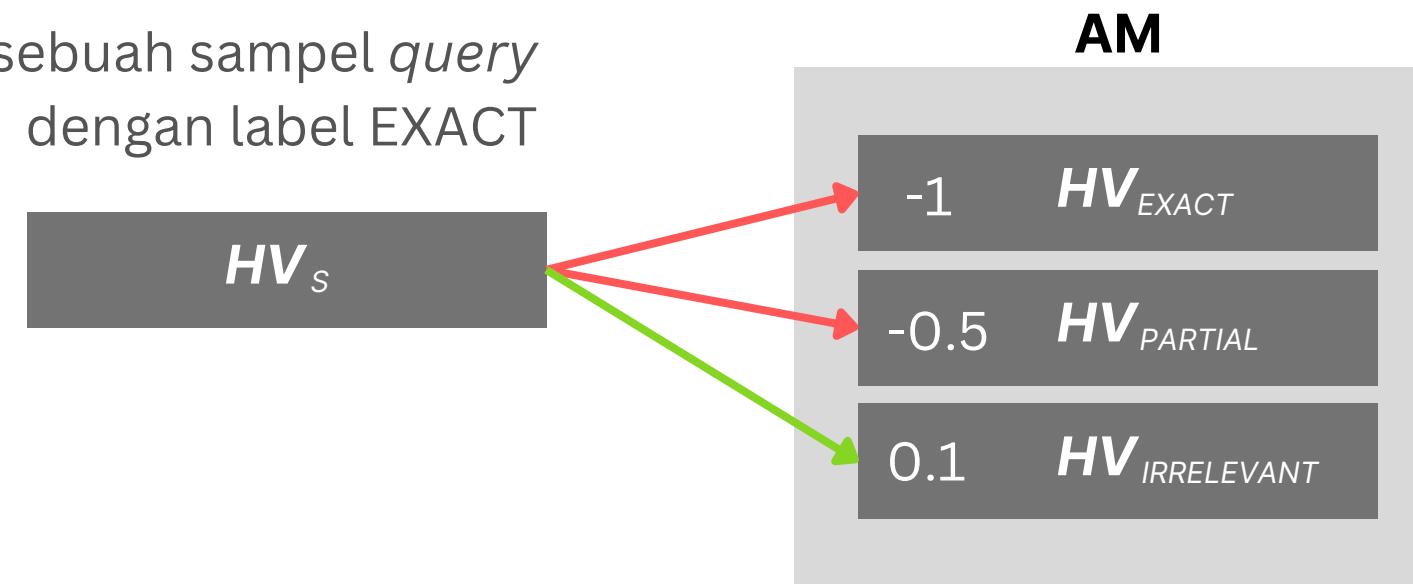
Pembentukan  $HV$  query dilakukan dengan menggabungkan tiga konsep dari *query*, *query class*, dan *product*



## Metodologi - Proses Training

Proses *training* dilakukan dengan memanfaatkan fungsi ***cosine similarity*** untuk mendapatkan label yang tepat.

Diketahui sebuah sampel *query* dengan label EXACT



**Skenario koreksi** - koreksi label

$$HV_{EXACT} = HV_{EXACT} + HV_s \times \text{learning rate}$$

**Skenario koreksi** - koreksi prediksi

$$HV_{IRRELEVANT} = HV_{IRRELEVANT} - HV_s \times \text{learning rate}$$

**Syarat 1**- antisipasi *overfitting*

Jika sebuah **prediksi salah** namun nilai similaritasnya telah tinggi seperti 0.95 maka **koreksi label tidak akan dilakukan**.

**Syarat 2**- antisipasi *underfitting* (usulan kami)

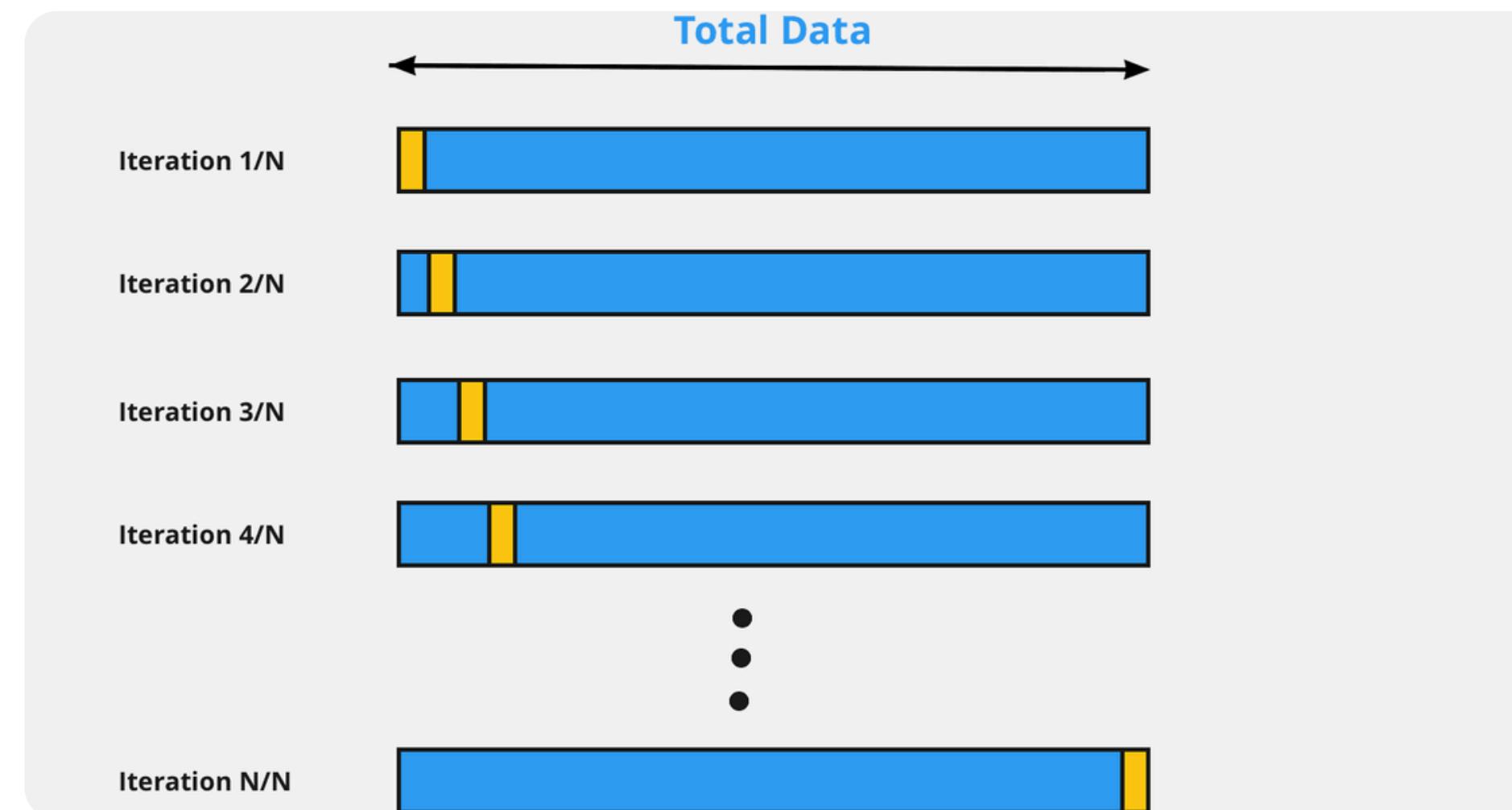
Jika sebuah **prediksi benar** namun nilai similaritasnya berada di rentang -1 hingga 0 maka **koreksi label akan dilakukan**.

## Metodologi - Validasi

- **Metode Ranking, “double sort”**

Proses Ranking akan dilakukan dengan cara mengurutkan berdasarkan label, lalu mengurutkannya kembali berdasarkan nilai *cosine similarity*-nya.

- **Leave-One-Group-Out Cross Validation**



## Metodologi - Metrik

- **Mean Reciprocal Rank (MRR)**

Penggunaan MRR adalah untuk mengetahui kualitas pengurutan (Ranking).

$$MRR = \frac{1}{Q} \sum_{i=1}^Q \frac{1}{rank\ i}$$

- **Normalized Discounted Cumulative Gain (NDCG)**

Penggunaan NDCG difokuskan untuk mengetahui kemampuan model memberi skor relevansi.

$$CG = \sum_{i=1}^n Gain_i \longrightarrow DCG = \sum_{i=1}^{ranks} \frac{Gains}{log(i+1)} \longrightarrow IDCG = \sum_{i=1}^{ranks} \frac{sorted(Gains)}{log(i+1)} \longrightarrow NDCG = \frac{DCG}{IDCG}$$

- **Rata-rata latensi**

Nilai rata-rata latensi digunakan untuk memperkirakan efisiensi model dalam memprediksi sebuah kandidat produk.

## Pembahasan - Evaluasi Training FastVSS

Model	MRR	NDCG@50	Rata-Rata Latensi
FastVSS	0.9	0.820	0.11

Table 1: Evaluasi FastVSS

**NDCG@50 : 0.82**

mengindikasikan kemampuan model yang baik dalam me-retrieve produk-produk dengan relevansi tinggi

**MRR : 0.9**

mengindikasikan kemampuan model dalam mengurutkan produk-produk dekat dengan kondisi ideal

**Rata-Rata Latensi : 0.11 ms** (prediksi satu kandidat produk)

mengindikasikan efisiensi model yang cukup tinggi

## Pembahasan - FastVSS vs Conventional LLMs

Model	NDCG@50
Roberta Base	0.772
Roberta Large	0.773
SimCSE Large	0.768
<b>FastVSS</b>	<b>0.820</b>

Table 2: FastVSS vs conventional LLMs

- **FastVSS** memiliki kemampuan *retrieval* yang sangat baik dan **mampu bersaing** dengan model-model LLM konvensional.
- **FastVSS** memiliki perbedaan nilai lebih dari 0.05 untuk setiap model, mengindikasikan adanya **potensi** yang sangat menjanjikan bagi **perkembangan HDC** dalam *information retrieval* pada search engine.

## Pembahasan - A bit of Demo

Input : Query, Query Class

Output : 50 produk dengan relevansi tertinggi terhadap Query

```
query = 'full length mirror'
jumlah= 50
qclass = 'Wall & Accent Mirrors'
model.retrieve(query, qclass,jumlah)
✓ 0.1s
```

Predicting WANDS: 100%

424/424 [00:00<00:00, 2062.65it/s]

Time taken for this query: 62.5 ms  
shape: (50, 4)

item	group	type	score
---	---	---	---
str	str	i64	f64
berman slim over the door full...	Wall & Accent Mirrors	2	0.026242
twig rustic beveled accent mir...	Wall & Accent Mirrors	2	0.01515
lafontaine rustic distressed a...	Wall & Accent Mirrors	2	0.014778
belle meade rectangular molded...	Wall & Accent Mirrors	2	0.014723
merseyside distressed accent m...	Wall & Accent Mirrors	2	0.012977
...	...	...	...
lunado full length mirror	Wall & Accent Mirrors	2	0.004494
talmadge coastal beveled distr...	Wall & Accent Mirrors	2	0.004337
alessandra soft corner metal a...	Wall & Accent Mirrors	2	0.004184
mississauga weathered mirror	Wall & Accent Mirrors	2	0.004169
emert beveled distressed accen...	Wall & Accent Mirrors	2	0.004102

```
query = 'ergonomic chair'
jumlah= 50
qclass = 'Office Chairs'
model.retrieve(query, qclass,jumlah)
✓ 0.3s
```

Predicting WANDS: 100%

519/519 [00:00<00:00, 1701.78it/s]

Time taken for this query: 78.125 ms  
shape: (50, 4)

item	group	type	score
---	---	---	---
str	str	i64	f64
office chair	Office Chairs	2	0.037395
tristani executive chair	Office Chairs	2	0.032584
pierron ergonomic task chair	Office Chairs	2	0.031353
swivel executive chair	Office Chairs	2	0.030982
opheim conference chair	Office Chairs	2	0.030855
...	...	...	...
nettles task chair	Office Chairs	2	0.024981
office chair	Office Chairs	2	0.024905
lollie executive chair	Office Chairs	2	0.024843
task chair	Office Chairs	2	0.024823
partain executive chair	Office Chairs	2	0.024747

## Kesimpulan - dan Saran

Secara umum, model memiliki **kemampuan baik dan cepat** untuk melakukan proses *information retrieval* dengan indikasi bahwa model dapat memiliki performa pemodelan yang kuat dalam merepresentasikan informasi. Memenuhi tujuan awal penelitian ini untuk **memperkenalkan metode *information retrieval* yang mengimplementasikan HDC pada search engine.**

Penelitian ini memberikan arah positif bagi pengembangan model *IR* baru guna memiliki performa yang lebih cepat, namun tetap memiliki kemampuan pemodelan yang baik.

Beberapa saran bagi penelitian selanjutnya:

- Penggunaan **metode vektorisasi awal** yang memiliki performa lebih tinggi dibandingkan dengan Word2Vec
- Pengujian FastVSS pada **data kondisi bervariasi** untuk melihat lebih dalam kemampuan pemodelan dan efisiensi model
- Pengujian FastVSS pada **distributed computing**
- Pengujian FastVSS pada **implementasi *information retrieval* lainnya** seperti RAG dan question-answering.

***Terima Kasih  
Feedbacks are Open!***