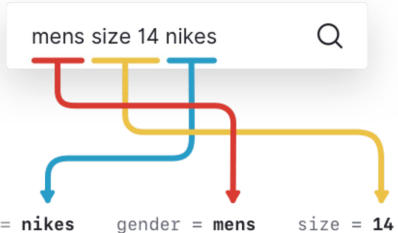# SAJARI

Highly efficient dense retrieval using neural hashes

hello@sajari.com

Core components of great search
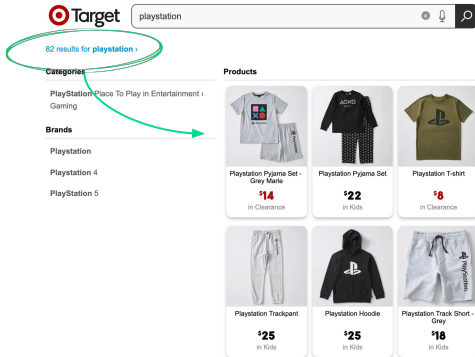
Query understanding

Modify the query to help improve the relevance process
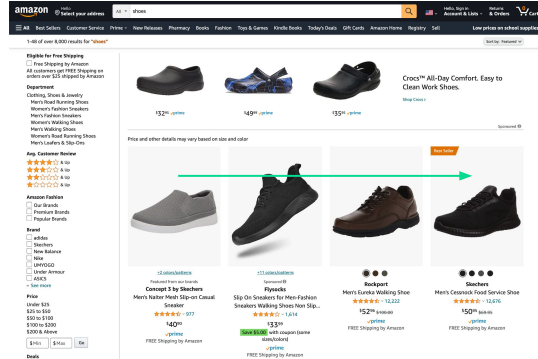

Relevance

Filter results to best match the query intent (user satisfaction)

Ranking

Order results to maximise a goal (e.g. units sold, profit, etc)

What is AI relevance?
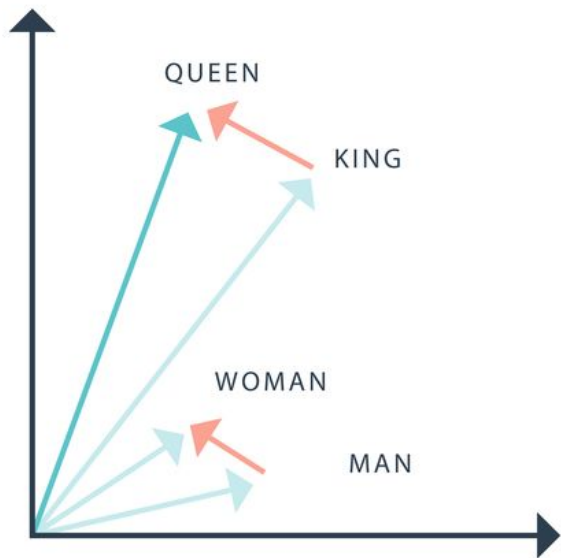
Represent words (and images) mathematically

Keyword search is binary, terms either match or they don't. Synonyms and rules help to make this work ok.

AI relevance instead uses similarity. Close concepts are represented very similarly mathematically.

Learn and improve with data

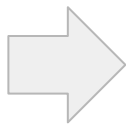Data such as purchases can be used to amplify relationships and thus continuously improve relevance.

KING - MAN + WOMAN = QUEEN

The next evolution of search indexes

Keywords

2000 - present
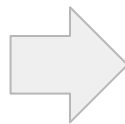
Vectors

2013 - present

Neural hashes

2021

???

First generation AI. Better quality results, but very expensive and/or very slow at scale

Next generation AI.

Extremely fast and efficient at scale, with the same result quality as vectors!

Why hashes over vectors?

	Vectors	Neural hashes
Storage size	**Large**. Must store the vector and some kind of graph or tree based structure	**Tiny**. Hashes for 99% nDCG use **~10%** of the vector size and there is no structure needed for storage. Can also go smaller!
Query speed	**Moderate to fast**. Various, but commonly need at least some float arithmetic.	**Ridiculous**. Simple binary bit comparison such as XOR. GPUs are not needed at all.
Data velocity	**Moderate**. Varies, but typically immutable, thus deletion and/or editing not supported	**Great**. Fits in a modified posting list. Can be inserted, modified or deleted easily. No graph or tree based structures to maintain.

Inevitable that hashes will replace vectors for dense search

Demo time!

__1 cpu__, 25k products → hash index = 15MB

What this does do ✅

Old Best Buy catalog (~25k products)

AI based queries (universal sentence encoder vectors → neural hashes)

Nearest neighbor lookup (ANN)

Applies historical and/or real-time learnings to improve results

What this does <u>not</u> do ❌

No keywords or tokenization

No language processing

No boosting

No rules

No hacks

No ranking aside from the AI driven nearest neighbor

SAJARI

Thank you

Keep in touch: @hamishogilvy

Relevance is a hard problem

Today relevance is largely still determined by keywords, synonyms, stemming, stop words, rules tokenization, ...

Without **enormous human assistance** this approach works poorly...

Even for very large companies