

# Science Behind Search

**Venkatesh Vinayakarao**

<http://vvtesh.co.in>

---

Principal Engineer, HERE Technologies

---

Information is the resolution of uncertainty. - **Claude Shannon.**

# Disclaimer

The views expressed in the presentation are solely those of the presenter and not necessarily those at HERE Technologies. HERE Technologies does not guarantee the accuracy or reliability of the information provided herein.

# About Me

**BE (Computer Science and Engineering)**

Java/J2EE  
Developer

**MS (Information Technology)**

SDE, Search  
Technologies  
Group, Bing,  
Microsoft

Principal  
Engineer, Cloud  
Platforms Group,  
Yahoo

**PhD (Computer Science)**

Principal  
Engineer, Search,  
Here  
Technologies

Intern, Porting ML  
Models to Azure,  
Microsoft  
Research

**8.5 Billion  
searches per  
day!**

**Google**

Google Search

I'm Feeling Lucky

Google offered in: हिन्दी বাংলা తెలుగు मराठी ಕನ್ನಡ ལྷོ་ཁྲིའི་སྐད་ ལྷོ་ཁྲིའི་སྐད་

**Bing**

**yahoo!**

**Life without search engines is difficult  
to imagine!**

# Role of Information

- We are always seeking information
  - How to be safe during COVID times?
  - Which universities provide the specialization I need?
  - How to get into a top college?
  - Which course to register for?
  - What are people talking about that new movie?
  - How to prepare for interviews?
  - ...
- If only you had the information, you could rule this world!

**What happens when all the information  
is deprived from you?**



# Solitary Confinement is Cruel

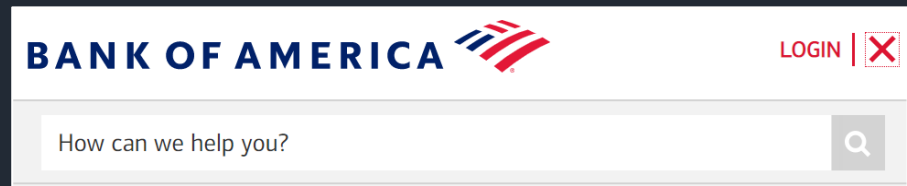
Picture Source: <https://jjustice.org/resources/solitary-confinement>

**There is search beyond the web search engines.**

# Search in Banking and Finance

Lots of  
products to  
sell

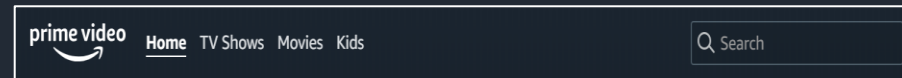
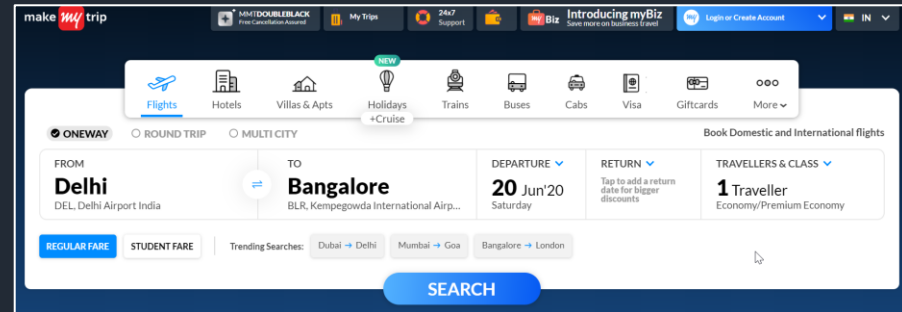
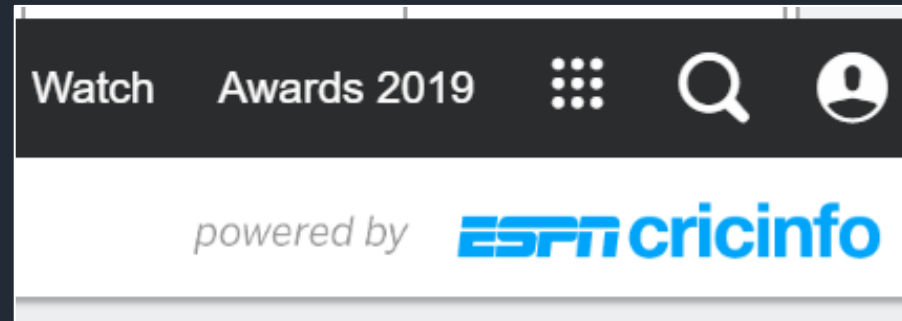
Reach a part of  
documentation  
faster





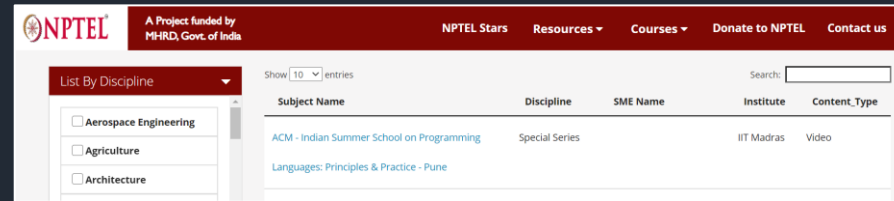
# Search in Sports, Travel & Entertainment

Search events,  
programs, and  
schedules



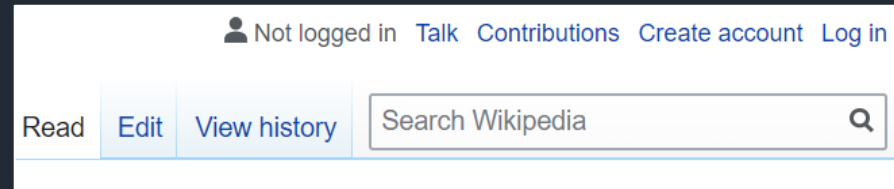
# Search in Education, Ecommerce and Healthcare

Search courses,  
articles,  
symptoms,  
books, etc.

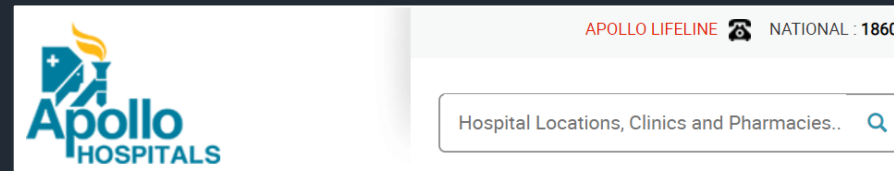


The image shows the NPTEL website header and search results. The header includes the NPTEL logo, a note that it is a project funded by MHRD, Gov. of India, and navigation links for NPTEL Stars, Resources, Courses, Donate to NPTEL, and Contact us. Below the header is a search bar and a table of search results.

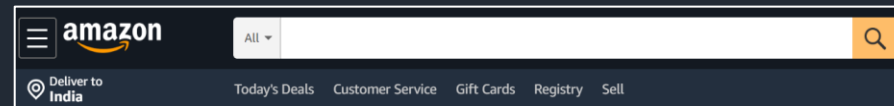
Subject Name	Discipline	SME Name	Institute	Content_Type
ACM - Indian Summer School on Programming	Special Series		IIT Madras	Video
Languages: Principles & Practice - Pune				



The image shows the Wikipedia search bar. It includes a user status indicator "Not logged in" with links for Talk, Contributions, Create account, and Log in. Below this are buttons for Read, Edit, and View history, followed by a search input field with the placeholder text "Search Wikipedia" and a search icon.



The image shows the Apollo Hospitals website header. It features the Apollo Hospitals logo on the left and the text "APOLLO LIFELINE" with a phone icon and "NATIONAL : 1860-5" on the right. Below this is a search bar with the placeholder text "Hospital Locations, Clinics and Pharmacies.." and a search icon.



The image shows the Amazon India website header. It features the Amazon logo on the left and a search bar with the placeholder text "All" and a search icon. Below the search bar are navigation links for Today's Deals, Customer Service, Gift Cards, Registry, and Sell.



### Senior Search Engineer I

McKinsey & Company · Gurugram, Haryana, India



### Staff Software Engineer - Search

GitHub · United Kingdom (Remote)



### Engineering Manager (Search Platform)

Databricks · Bengaluru, Karnataka, India (On-site)



### Software Development Engineer - SDE2 / AWS |

### Amazon OpenSearch, Amazon OpenSearch

Amazon Web Services (AWS) · Bengaluru, Karnataka, India



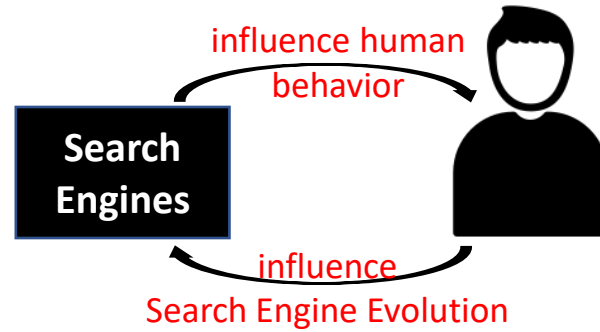
### Staff Software Engineer, Search

Google · Bengaluru, Karnataka, India (On-site)



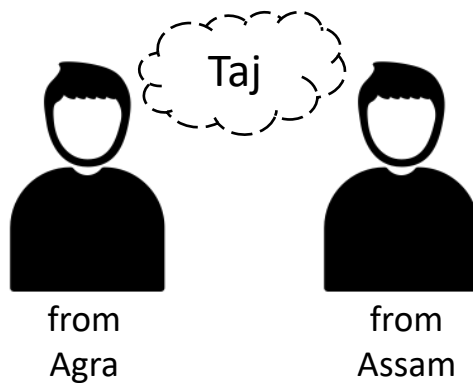
### Senior Principal Software Engineer - Search Platform

Atlassian · Greater Kolkata Area (Remote)

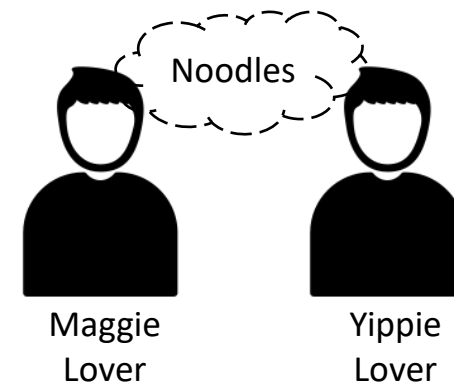


**We have amazing search products.  
Yet, an effective search is still a work of art!**

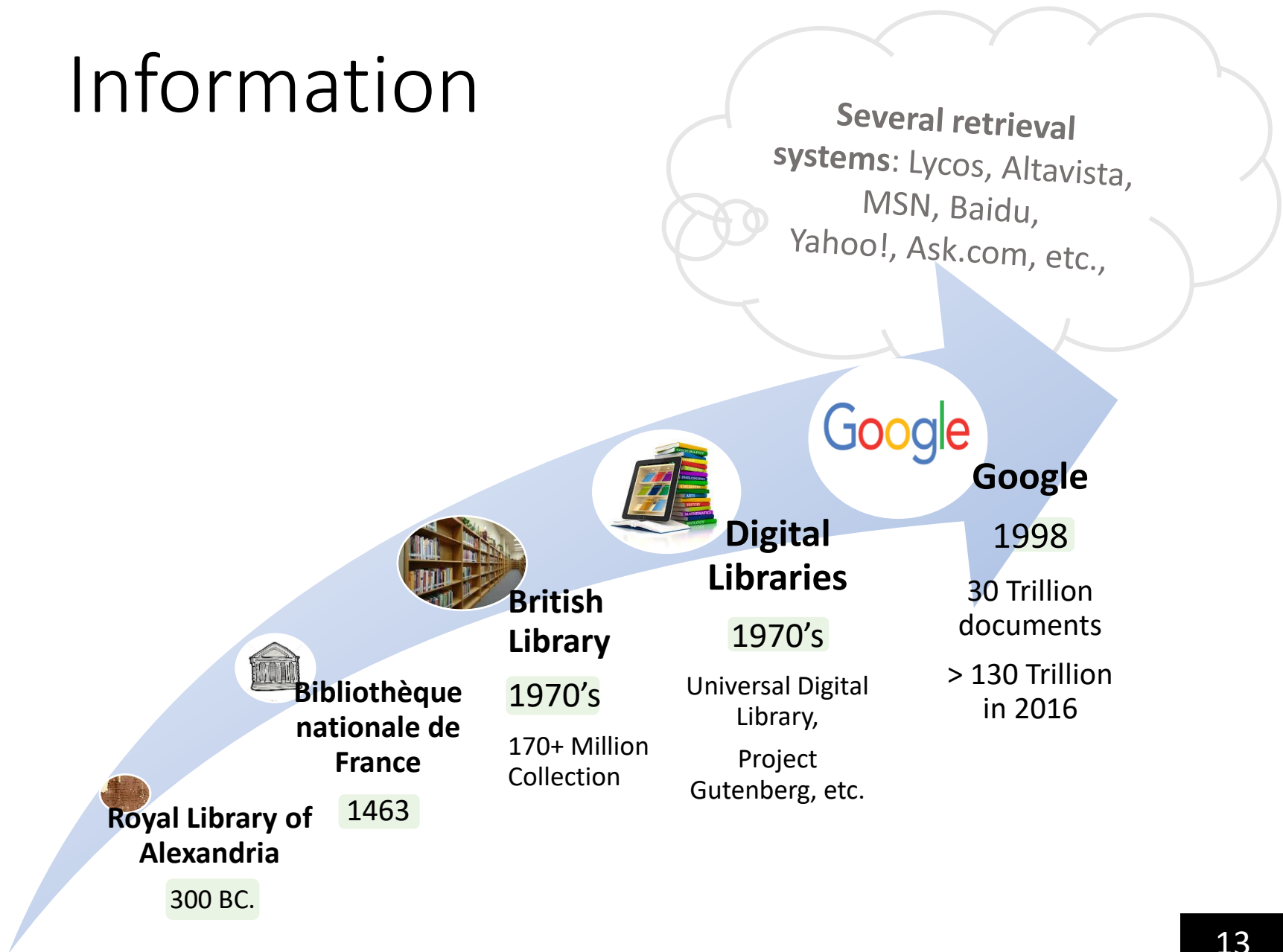
User needs to express the information need properly



Relevance may be subjective



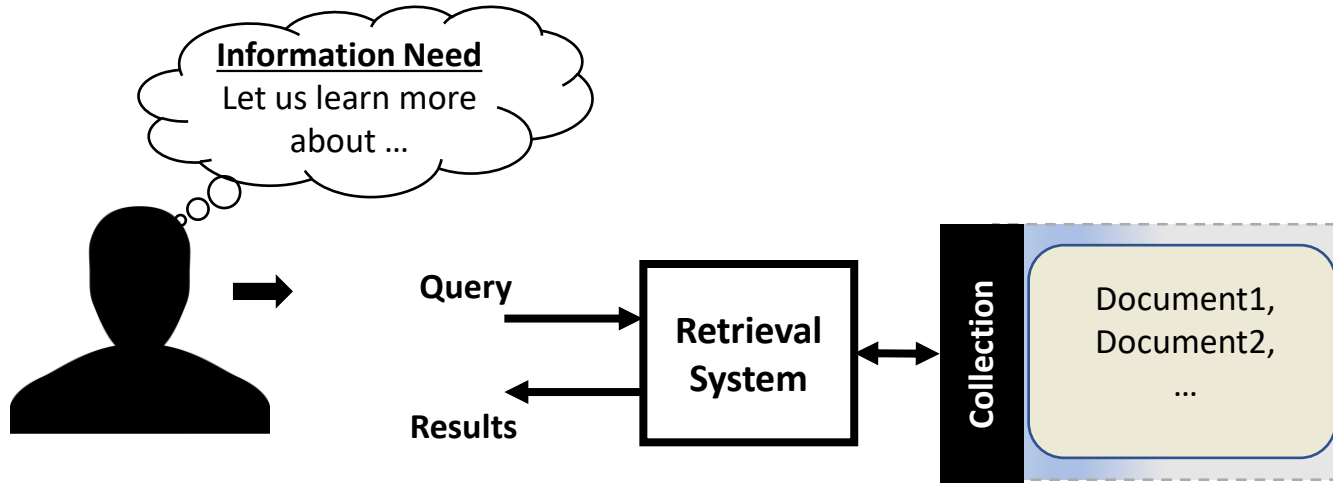
# Information



# The Science behind Search

**How do search engines work?**

# What is Information Retrieval?



Information Retrieval (IR) is finding material (usually documents) of an **unstructured** nature (usually text) that satisfies an **information need** from within **large collections**.

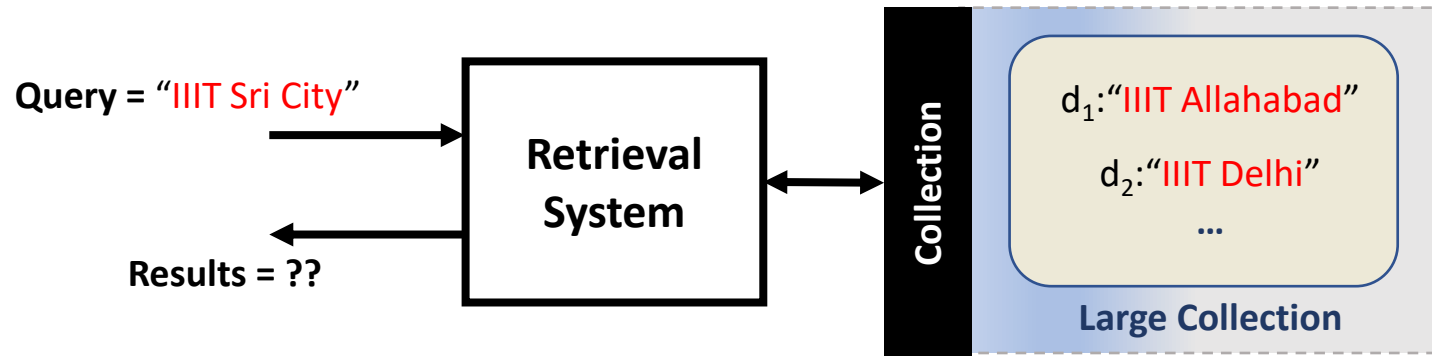
– From the Manning et al. IR Book.

# Simple Retrieval Problem

- Say, we have a **collection** with 5 **documents**, each having the following contents
  - d1: IIIT ALLAHABAD
  - d2: IIIT DELHI
  - d3: IIIT GUWAHATI
  - d4: IIIT KANCHIPURAM
  - d5: IIIT SRI CITY
- Assume, the **Query** is
  - IIIT SRI CITY
- Which **document** will you match and why?



# The Problem: How to Build a Retrieval System?



# One (bad) Approach

- First match the **term** IIT.
  - Filter out documents that contain this term.
- Next match the **term** Sri.
  - Filter out documents that contain this term.
- Next match the **term** City.
  - Filter out documents that contain this term.

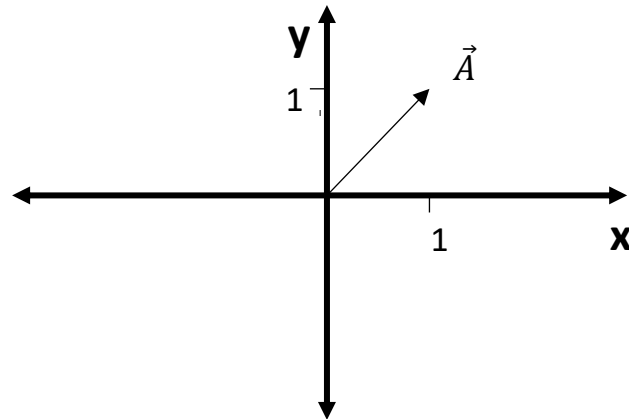
**Three iterations!**  
**Quiz: Can we do better?**

# A Better Approach

**Revisiting  
Linear Algebra**

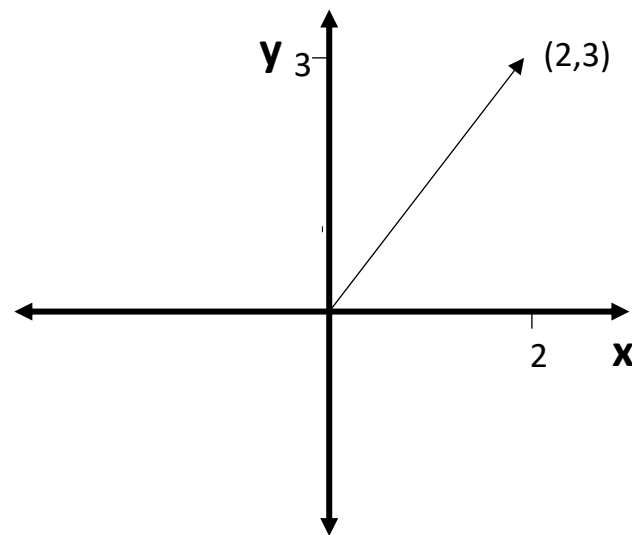
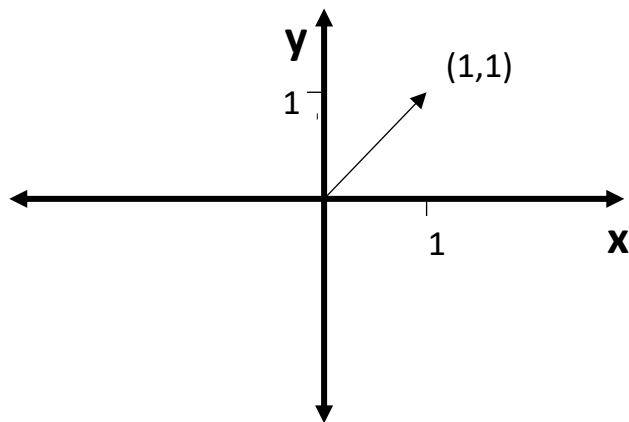
# Vector

- Geometric entity which has magnitude and direction

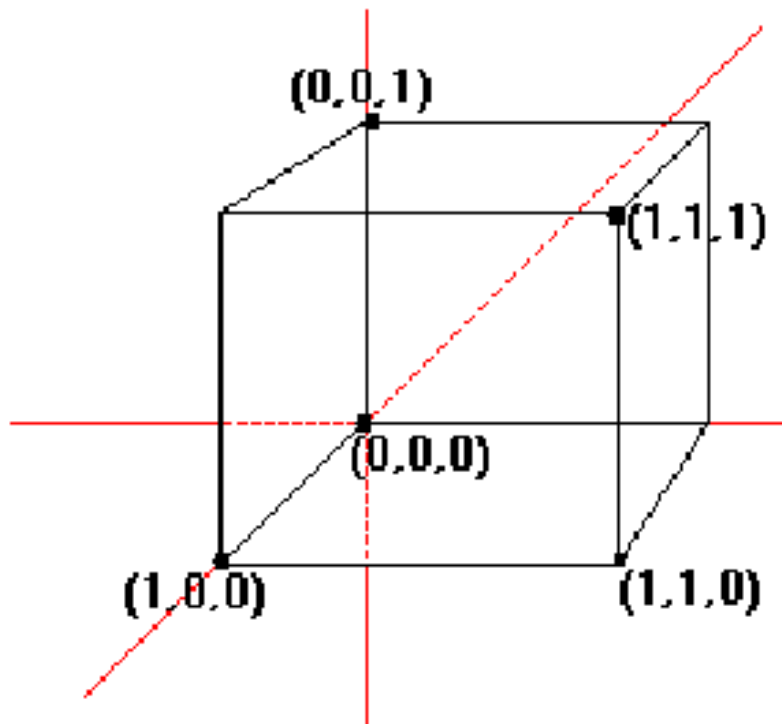


$\vec{A}$  is fixed at (0,0)

# How is (2,3) Different?



What is  $(1,1,1)$  ?

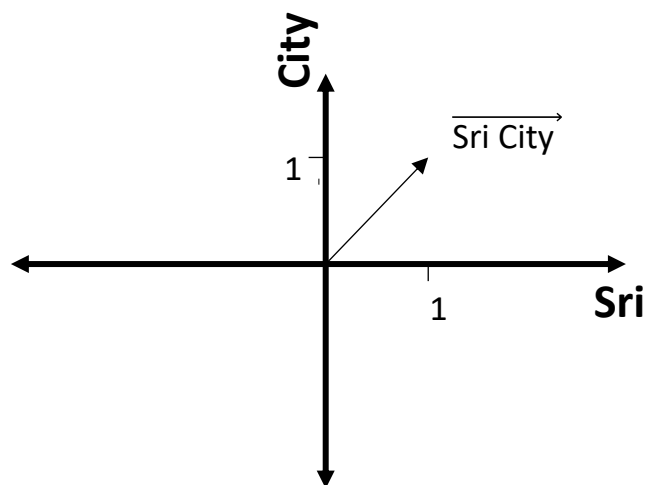


# Remember!

**A number is just a mathematical object. We  
give meaning to it!**

# Sentences are Vectors

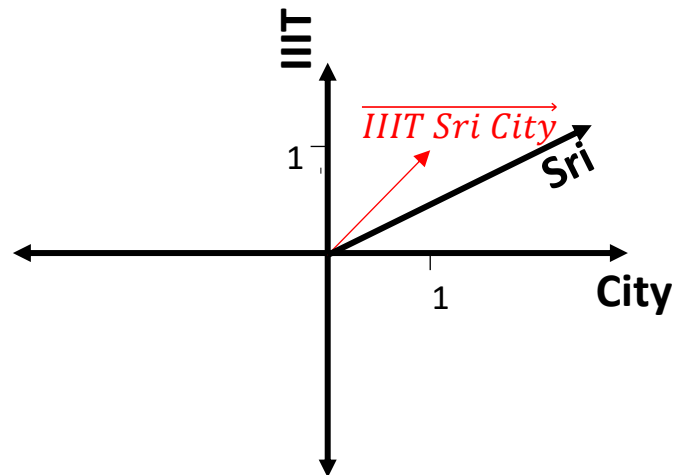
- “Sri City” as a vector





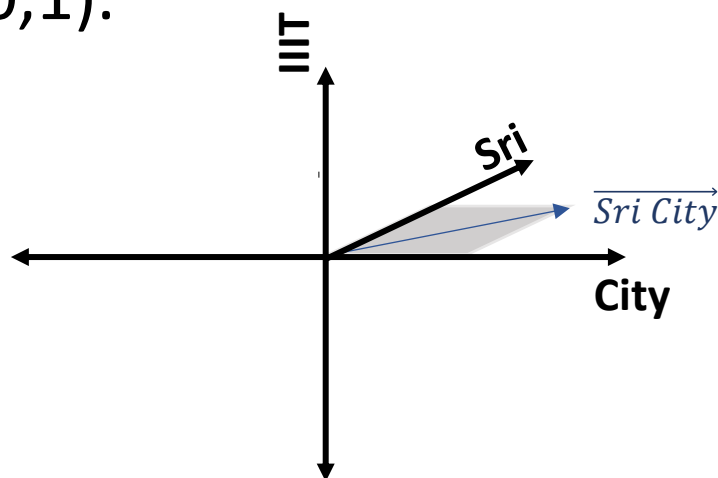
# Sentences are Vectors

- “**IIIT Sri City**” is a 3-dimensional vector



# Sentences are Vectors

- On this 3D space, “Sri City” vector will lie on the x (City) and z (Sri) plane. If  $(x,y,z)$  denotes the vector, “Sri City” is  $(1,0,1)$ .



# Natural Language Phrases as Vectors

Let query  $q = \text{"IIIT Sri City"}$ .

Let document,  $d_1 = \text{"IIIT Sri City"}$  and  $d_2 = \text{"IIIT Delhi"}$ .

	IIIT	Sri	City	Delhi
q	1	1	1	0
$d_1$	1	1	1	0
$d_2$	1	0	0	1

$q = (1,1,1,0)$ ,  $d_1 = (1,1,1,0)$  and  $d_2 = (1,0,0,1)$

# Quiz

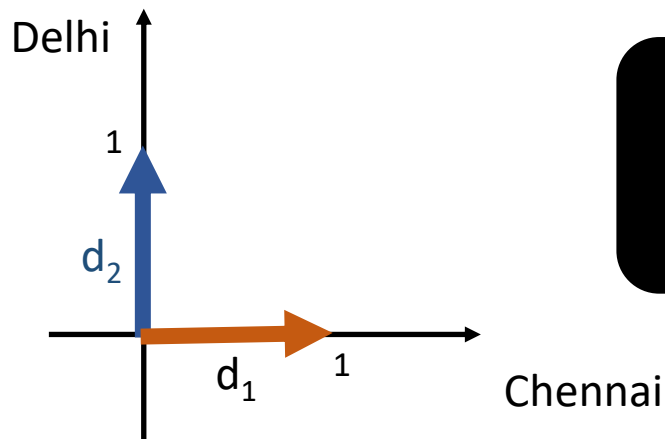
- Considering the following vectors:

	IIIT	Sri	City	Delhi
q	1	1	1	0
d <sub>1</sub>	1	1	1	0
d <sub>2</sub>	1	0	0	1

- What is the Natural Language (NL) equivalent of  $(0,1,1,0)$  ?
- What is the NL equivalent of  $(1,0,0,1)$  ?
- What is the vector for Delhi?
- What is the NL equivalent of q here?

# Similarity Score

- Assume, we have the following two documents:
  - $d_1 = \text{“Chennai”}$
  - $d_2 = \text{“Delhi”}$
- On a scale of 0 – 1, similarity of  $d_1$  and  $d_2$  must be 0



Can we express similarity as a function of the angles?

0 – 90 to 1 – 0: How?

	0°	30°	45°	60°	90°
sin	0	1/2	1/√2	√3/2	1
cos	1	√3/2	1/√2	1/2	0
tan	0	1/√3	1	√3	Not defined

# Back to Trigonometry: Dot Product

- If  $\mathbf{a}$  and  $\mathbf{b}$  are non-unit vectors, what is the cosine of angle between them ( $\cos \theta$ )?

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos(\theta)$$

(or)

$$\cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}$$

# Matching Documents to Queries

- Document as a vector of term-occurrence

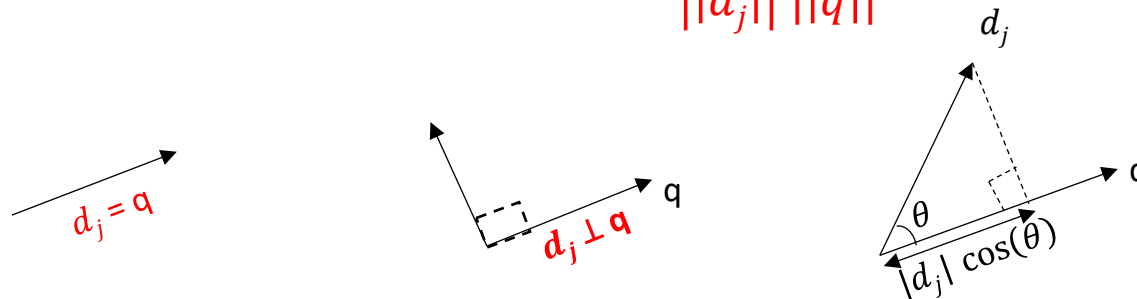
$$d_j = (w_{1j}, w_{2j}, \dots, w_{nj})$$

- Query as a vector of term-occurrence

$$q = (w_{1q}, w_{2q}, \dots, w_{mq})$$

- Similarity between these vectors can be represented as

$$\text{Cosine Similarity} = \cos(\theta) = \frac{d_j \cdot q}{\|d_j\| \|q\|}$$





# Example

Let query  $q = \text{"BITS Pilani"}$ .

Let document,  $d_1 = \text{"BITS Pilani Goa Campus"}$  and  $d_2 = \text{"IIT Delhi"}$ .

	BITS	Pilani	Goa	Campus	IIT	Delhi
q	1	1	0	0	0	0
$d_1$	1	1	1	1	0	0
$d_2$	0	0	0	0	1	1

In our VSM,  $q = (1,1,0,0,0,0)$ ,  $d_1 = (1,1,1,1,0,0)$  and  $d_2 = (0,0,0,0,1,1)$

$$\text{similarity}(d_1, q) = \frac{d_1 \cdot q}{\|d_1\| \|q\|} = \frac{1.1 + 1.1}{\sqrt{1^2+1^2+1^2+1^2} \sqrt{1^2+1^2}} = 0.71.$$

$$\text{similarity}(d_2, q) = \frac{d_2 \cdot q}{\|d_2\| \|q\|} = 0.$$

# Standard Measures

- Just like cosine similarity, we have more measures:
  - Euclidean distance,
  - Manhattan distance,
  - Chebyshev distance,
  - Canberra distance,
  - Bray–Curtis dissimilarity,
  - Pearson correlation coefficient
  - ...

# Which of the Following are Sets?

- ~~{1, 2, 3, 4, 5, 6, 5, 7, 8, 9, 10, 11, 12, 13}~~
- ~~{A, B, C, D, E, F, G, H, I, I, J, K, L, M, N, O}~~
- ~~{apple, banana, orange, apple, banana, orange}~~

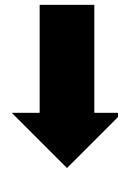


# Bag

- {1, 2, 3, 4, 5, 6, 5, 7, 8, 9, 10, 11, 12, 13}
- {A, B, C, D, E, F, G, H, I, I, J, K, L, M, N, O}
- {apple, banana, orange, apple, banana, orange}

# Set of Words Representation

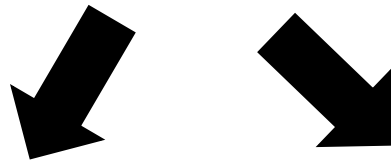
- “IIIT Sri City”  $\rightarrow$  {IIIT, Sri, City}
  - “IIIT Sri City, Sri City”  $\rightarrow$  {IIIT, Sri, City}
- (Assuming, we ignore the punctuations)



	IIIT	Sri	City
q	1	1	1

# Bag of Words Representation

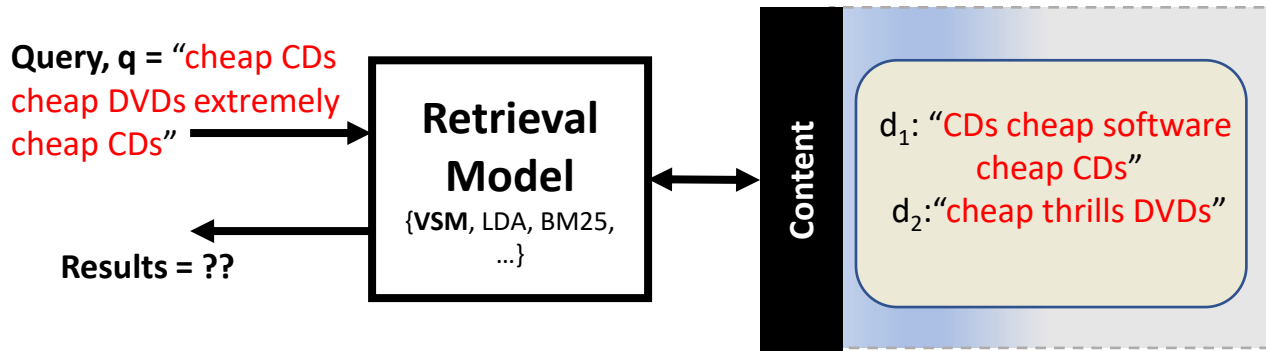
- “IIIT Sri City” → {IIIT, Sri, City}
- “IIIT Sri City, Sri City” → [IIIT, Sri, Sri, City, City]



	IIIT Sri City			IIIT Sri City, Sri City			
	IIIT	Sri	City		IIIT	Sri	City
q	1	1	1	q	1	2	2

Leads to different vectors

# Which Document to Retrieve?

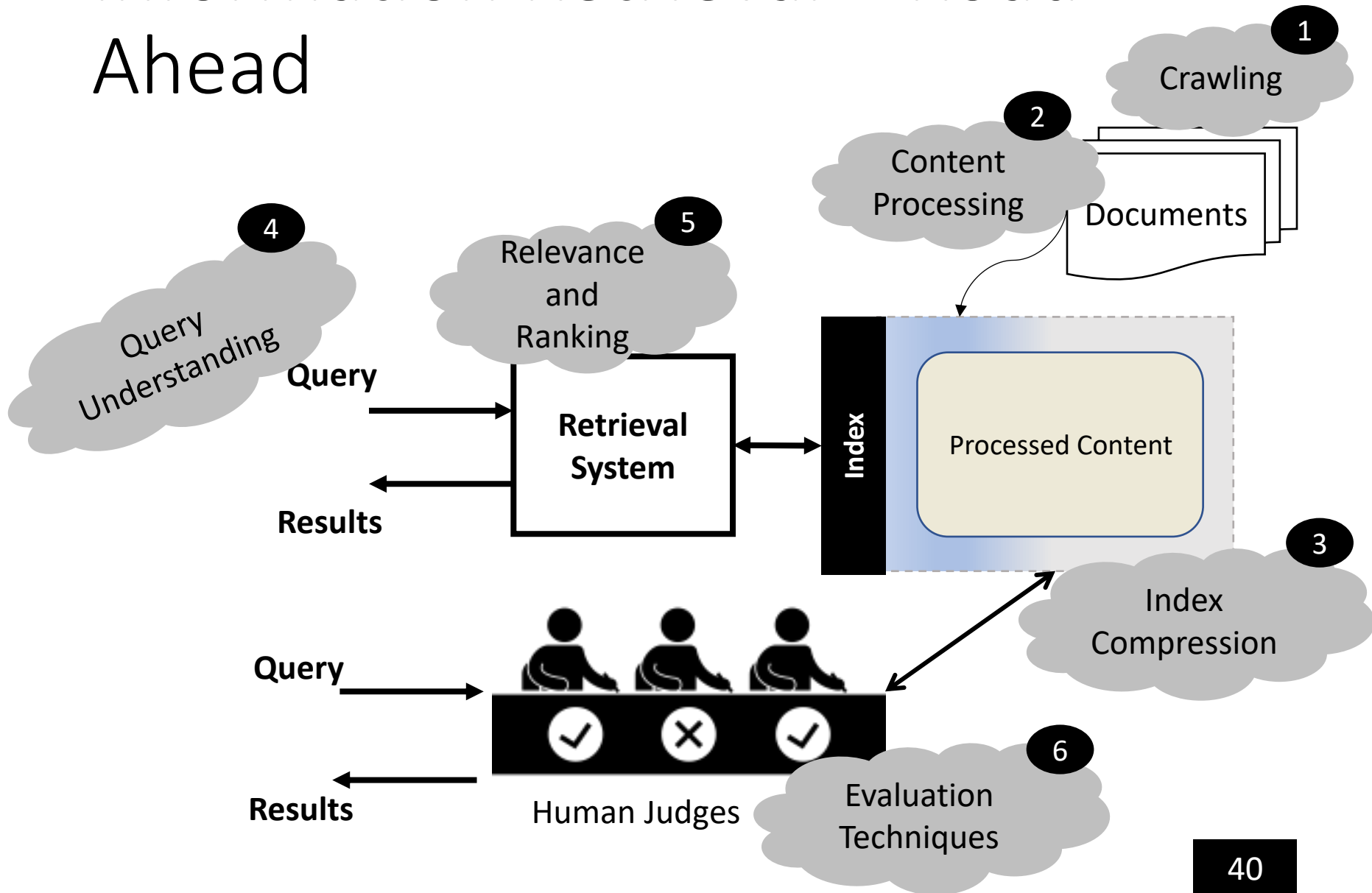


	cheap	CDs	DVDs	extremely	software	thrills
q	3	2	1	1	0	0
$d_1$	2	2	0	0	1	0
$d_2$	1	0	1	0	0	1

$\text{sim}(q, d_1) = 0.86$

$\text{sim}(q, d_2) = 0.59$

# Information Retrieval – Road Ahead





# Technologies & Frameworks



Apache



Apache



Apache



Univ. of Glasgow

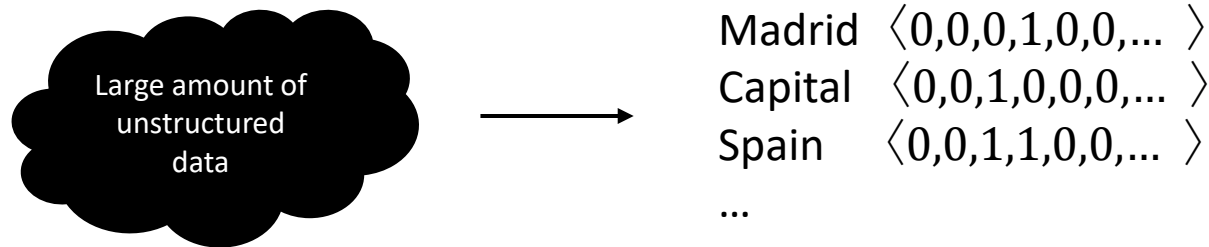


Galago, Indri

UMass & CMU

**Thanks to these... We can now focus on  
more complex problems.**

# Distributed Representations

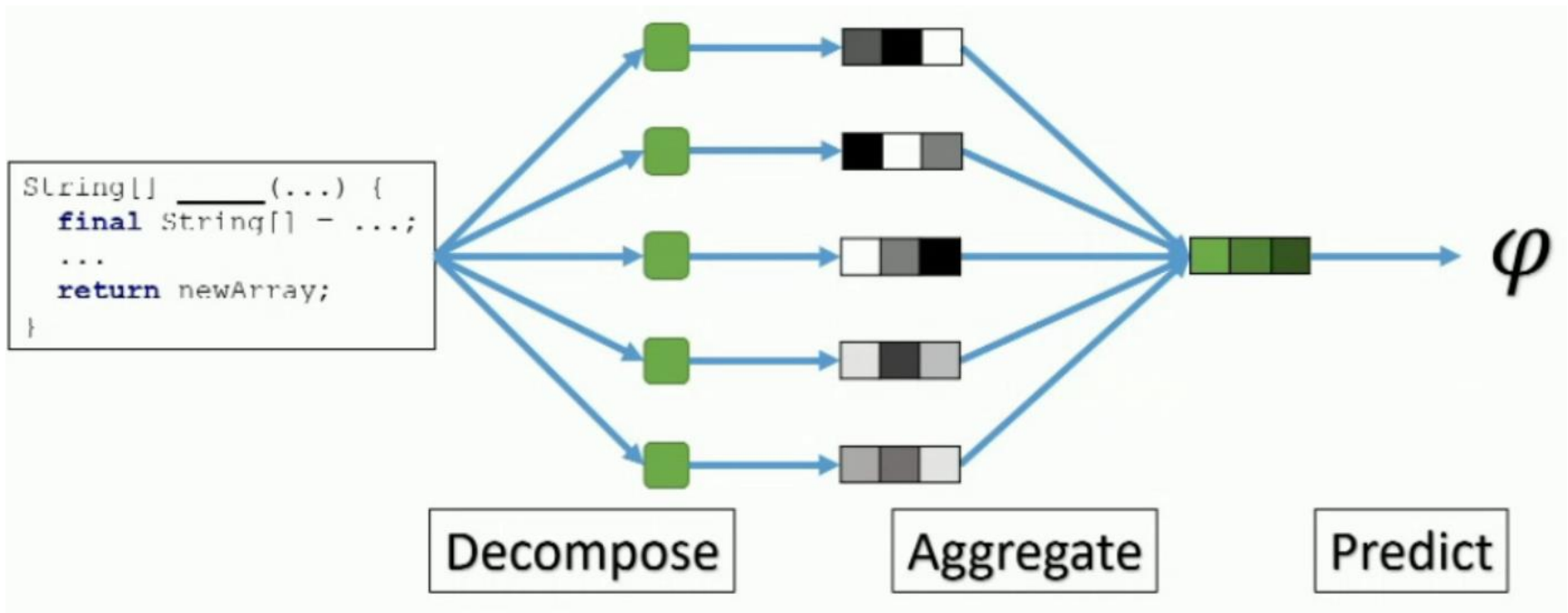


Can we encode the vector representations in a way that they capture semantic relationships?

For example, Madrid + Capital = Spain

# Code2Vec

- Predict properties of code



# Entity Search

Google search for "Amitabh Bachchan". The search bar contains "Amitabh Bachchan" and the search results show a knowledge panel for the actor. The panel includes a profile picture, a link to his Tumblr page, and a list of music services (Gaana, YouTube, Spotify). A detailed biographical paragraph follows, mentioning his birth date (11 October 1942), height (1.83 m), spouse (Jaya Bachchan), and upcoming movies (Hera Pheri 3, Brahmāstra, Chehre). A "Movies" section lists "Sholay" (1975), "Kabhi Khushi Kabhie..." (2001), "Mard" (1985), "Gulabo Sitabo" (2020), and "Coolie" (1983).

Search result for "Amitabh Bachchan". The result features a carousel of images, a "See all images" button, and a summary section. The summary identifies him as an "Indian Film Actor" and provides a biographical overview, including his birth date (11 Oct 1942) and location (Allahabad, India). It also lists his height (6' 0"), spouse (Jaya Bachchan), children (Abhishek Bachchan, Shweta Bachchan-Nanda), upcoming films (Chehre, Wisdom for Heroes, Brahmastra), and parents (Harivansh Rai Bachchan, Teji Bachchan). Social media links for Wikipedia, Twitter, Facebook, Official site, and Instagram are provided.

# We Need Answers!

where was albert einstein born



All

News

Images

Maps

Shopping

More

Tools

About 2,33,00,000 results (0.82 seconds)

Albert Einstein / Place of birth



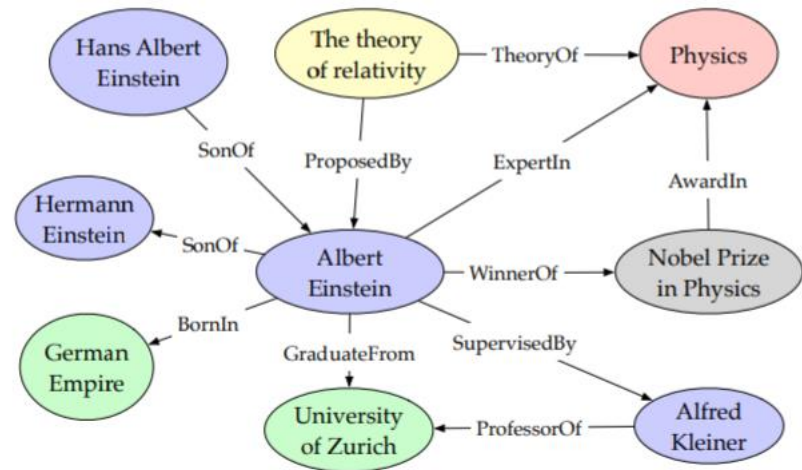
Ulm, Germany

# Knowledge Graphs

## Subject – Predicate – Object Triples

(Albert Einstein, **BornIn**, German Empire)  
(Albert Einstein, **SonOf**, Hermann Einstein)  
(Albert Einstein, **GraduateFrom**, University of Zurich)  
(Albert Einstein, **WinnerOf**, Nobel Prize in Physics)  
(Albert Einstein, **ExpertIn**, Physics)  
(Nobel Prize in Physics, **AwardIn**, Physics)  
(The theory of relativity, **TheoryOf**, Physics)  
(Albert Einstein, **SupervisedBy**, Alfred Kleiner)  
(Alfred Kleiner, **ProfessorOf**, University of Zurich)  
(The theory of relativity, **ProposedBy**, Albert Einstein)  
(Hans Albert Einstein, **SonOf**, Albert Einstein)

## Knowledge Graph



These facts are stored in a “knowledge base”.

# Recent Trends

- LLMs and Search
- Reasoning and Knowledge Graphs
- Multimodal IR
- Evaluating IR Systems
- Handling Fairness and Sensitivity
- Domain Specific IR
- Multilingual IR
- ...



## Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations

Krisztian Balog  
Google  
London, UK  
krisztianb@google.com

Filip Radlinski  
Google  
London, UK

## Query Rewriting for Voice Shopping Null Queries

Iftah Gamzu  
Amazon  
Tel-Aviv, Israel

Marina Haikin  
Amazon  
Tel-Aviv, Israel  
@amazon.com

Nissim Halabi  
Amazon  
Tel-Aviv, Israel  
nissimh@amazon.com

## Operationalizing the Legal Principle of Data Minimization for Personalization

Asia J. Biega  
Microsoft Research  
Montréal

Peter Potash  
Microsoft Research  
Montréal

Hal Daumé III  
Microsoft Research NYC  
University of Maryland

Fernando Diaz  
Microsoft Research  
Montréal

Michèle Finck  
Max Planck Institute for Innovation  
and Competition

Research contributions from  
leading corporates in SIGIR 2020

**In spite of all these developments,  
“search”ing effectively has been an art.**

(This is why) The **academia**, **research labs**, and the **software industry** needs you.  
Let us strive to build better search experiences.



Search Relevance Engineer  
Freshworks · Hyderabad, IN  
Posted 1 month ago · 301 views



Search COE Engineer  
Morgan Stanley · Mumbai, IN  
Posted 2 days ago · 32 views



Technical Architect / Sr SDE -  
Amazon Search Engine  
Technologies  
Amazon · Bangalore, IN  
Posted 3 weeks ago · 271 views

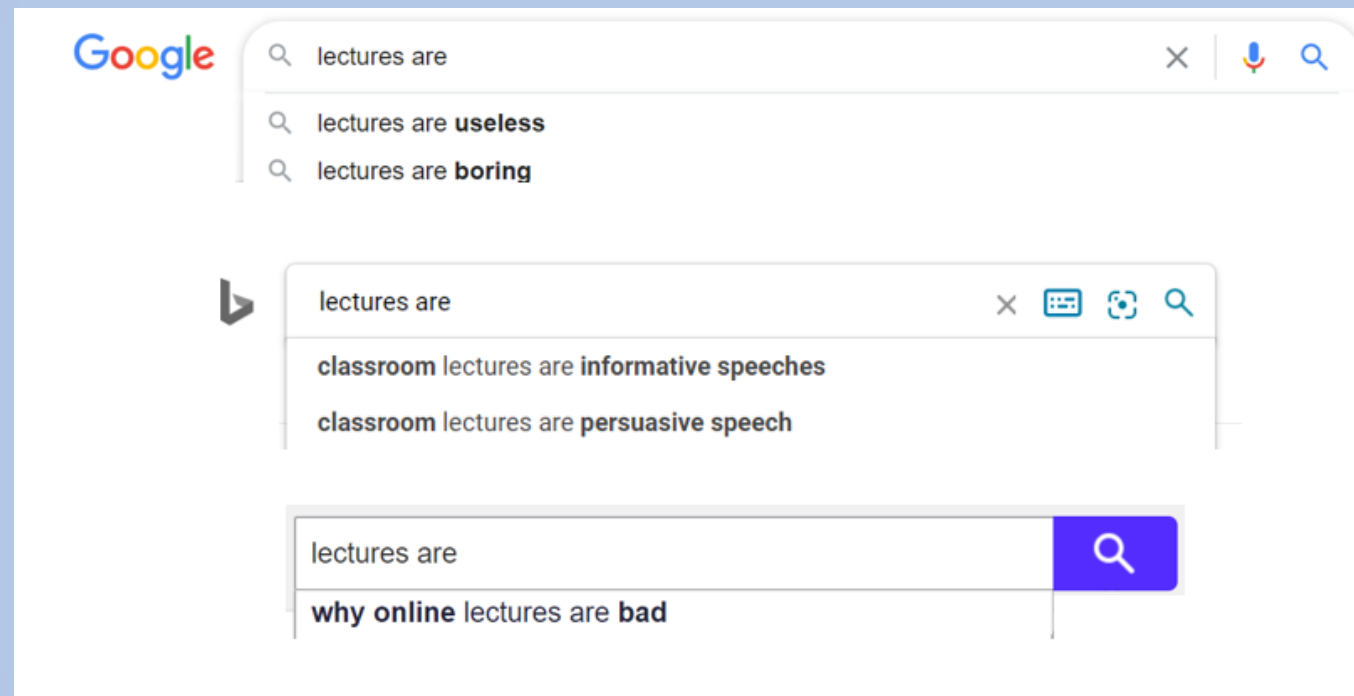


Search Engi  
Kubric · Bangal  
Posted 3 months ag



# References

- Proceedings of SIGIR 2024 - <https://sigir-2024.github.io/proceedings.html>
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.  
<https://nlp.stanford.edu/IR-book/information-retrieval-book.html>
- A Survey on Knowledge Graphs: Representation, Acquisition and Applications - Shaoxiong Ji, Shirui Pan – 2021.
- code2vec: learning distributed representations of code - Uri Alon, Meital Zilberstein, Omer Levy, Eran Yahav – 2019.



# Thank You

[venkateshv@cmi.ac.in](mailto:venkateshv@cmi.ac.in)

This slide deck is available at <http://vvtesh.co.in/>.

# Real-World Applications

- VSM dominates the retrieval methods
  - BM25 in Elasticsearch
  - Dense retrieval in Redis

# Alternatives to VSM

- Set-based Approaches with Jaccard Similarity
- Formal Concept Analysis
- Latent Semantic Indexing
- Probabilistic Retrieval (BIM, PRP)
- Language Models
- Program Synthesis
  - Read (Vector) Space is Not the Final Frontier: Product Search as Program Synthesis – Tagliabue & Greco, SIGIR ecom 2023.

# Demerits of VSM

- Dimensionality
  - How to reduce the vector size?
- Documents are represented as a linear combination of terms
  - May be unreal in most contexts
- All elements are conveniently assumed as positive
- Terms may not be independent