Google Cloud

Community Day 2022

Building Semantic Search Engines

Shweta Bhatt
Senior Applied Scientist @ Jupiter
@shweta_bhatt8

Google Developer Groups
Cloud • Kolkata

About Myself

- **Current Role**: Senior Applied Scientist @ Jupiter
- **Experience**: Conversational AI, Information Retrieval, Information Extraction, Recommendation Engines
- **Education**: MSc. Artificial Intelligence, King's College London, MSc. Tech. Information Systems, BITS Pilani
- **Beyond work**: Travel, music, tea enthusiast!

Google Developer Groups
Cloud • Kolkata

Agenda

1. Search engines vs semantic search engines
2. Science behind semantic search engines
3. Framework for building a semantic search engine
4. Why is it such a hard problem in the real world?
5. How to improve such systems?

Haven’t used a search engine?

Aol.com

Ask

YAHOO!

DuckDuckGo

Google

Baidu 百度

WolframAlpha

Bing

What are search engines?

Search Engine

Helps people find the information they are looking for online using keywords or phrases.

What is semantic search?

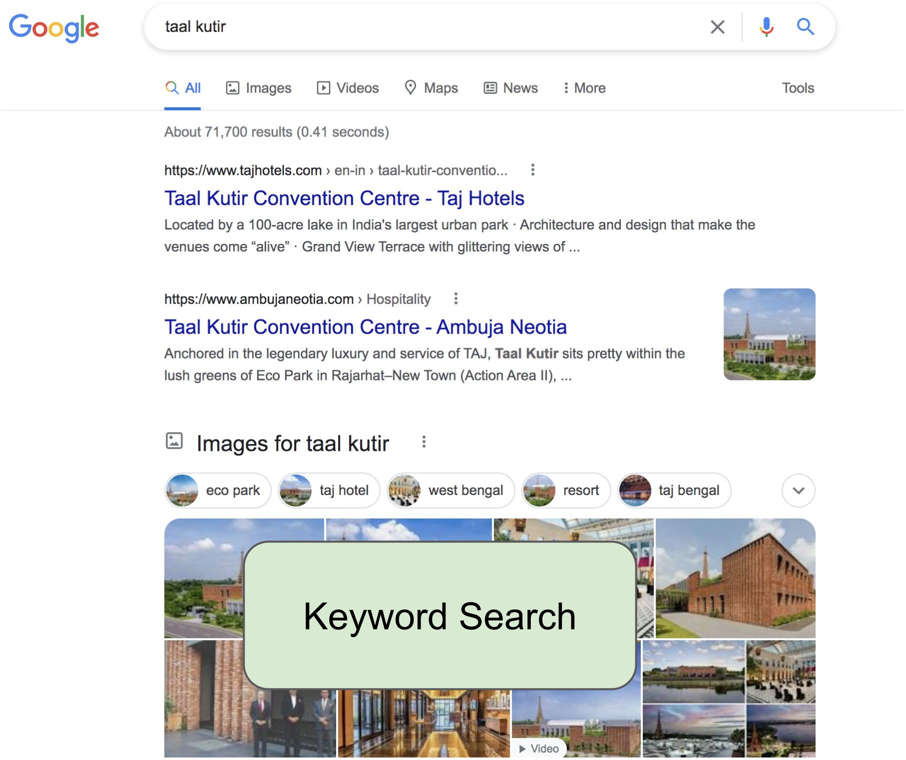
Semantic Search

- Understands the intent of the user / searcher via the contextual meaning of search text to generate more relevant results.

- Maximises the possibility of users getting the best search experience possible.

Does Google use semantic search?

Example



taal kutir

About 71,700 results (0.41 seconds)

https://www.tajhotels.com › en-in › taal-kutir-conventio...

Taal Kutir Convention Centre - Taj Hotels

Located by a 100-acre lake in India's largest urban park · Architecture and design that make the venues come “alive” · Grand View Terrace with glittering views of ...

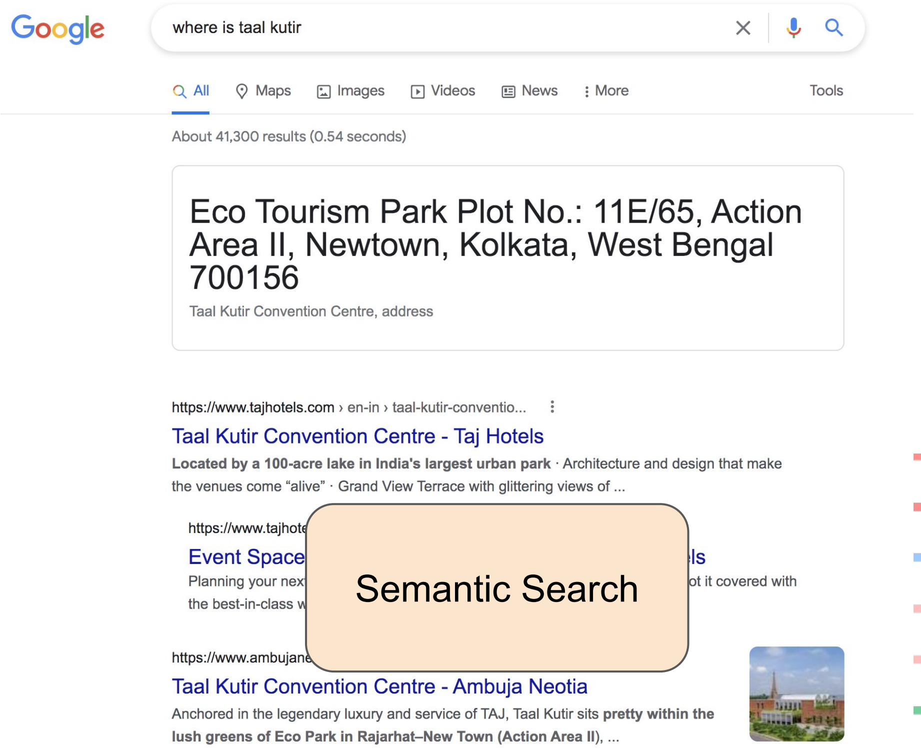
https://www.ambujaneotia.com › Hospitality

Taal Kutir Convention Centre - Ambuja Neotia

Anchored in the legendary luxury and service of TAJ, **Taal Kutir** sits pretty within the lush greens of Eco Park in Rajarhat–New Town (Action Area II), ...

Images for taal kutir

Keyword Search



where is taal kutir

About 41,300 results (0.54 seconds)

Eco Tourism Park Plot No.: 11E/65, Action Area II, Newtown, Kolkata, West Bengal 700156

Taal Kutir Convention Centre, address

https://www.tajhotels.com › en-in › taal-kutir-conventio...

Taal Kutir Convention Centre - Taj Hotels

Located by a 100-acre lake in India's largest urban park · Architecture and design that make the venues come “alive” · Grand View Terrace with glittering views of ...

https://www.tajhote

Event Space ls

Planning your nex ot it covered with

the best-in-class w

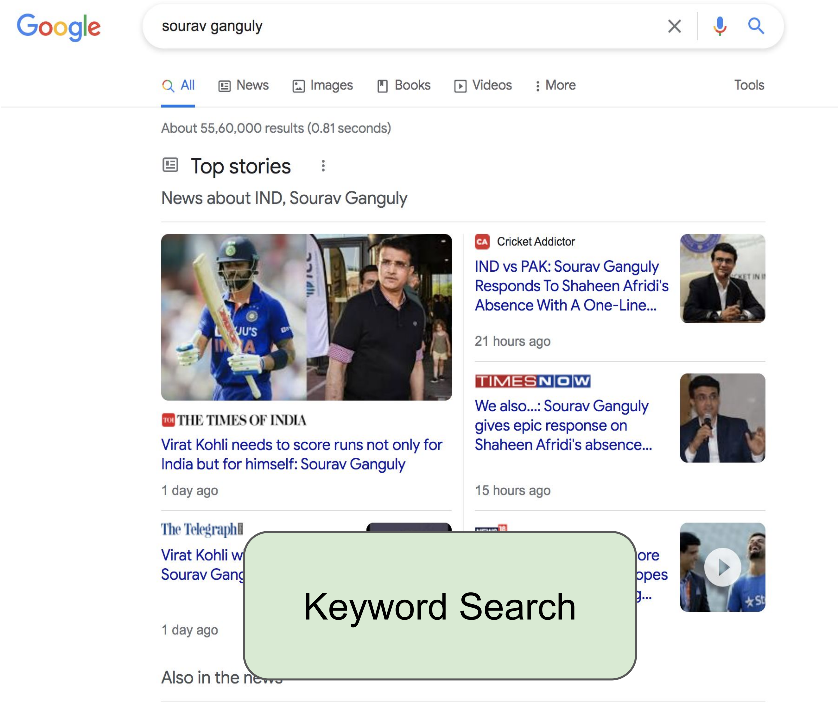
Semantic Search

https://www.ambujane

Taal Kutir Convention Centre - Ambuja Neotia

Anchored in the legendary luxury and service of TAJ, Taal Kutir sits **pretty within the lush greens of Eco Park in Rajarhat–New Town (Action Area II)**, ...

Another example



About 55,60,000 results (0.81 seconds)

Top stories

News about IND, Sourav Ganguly

THE TIMES OF INDIA

Virat Kohli needs to score runs not only for India but for himself: Sourav Ganguly

1 day ago

Cricket Addictor

IND vs PAK: Sourav Ganguly Responds To Shaheen Afridi's Absence With A One-Line...

21 hours ago

TIMES NOW

We also...: Sourav Ganguly gives epic response on Shaheen Afridi's absence...

15 hours ago

The Telegraph

Virat Kohli w

Sourav Gang

1 day ago

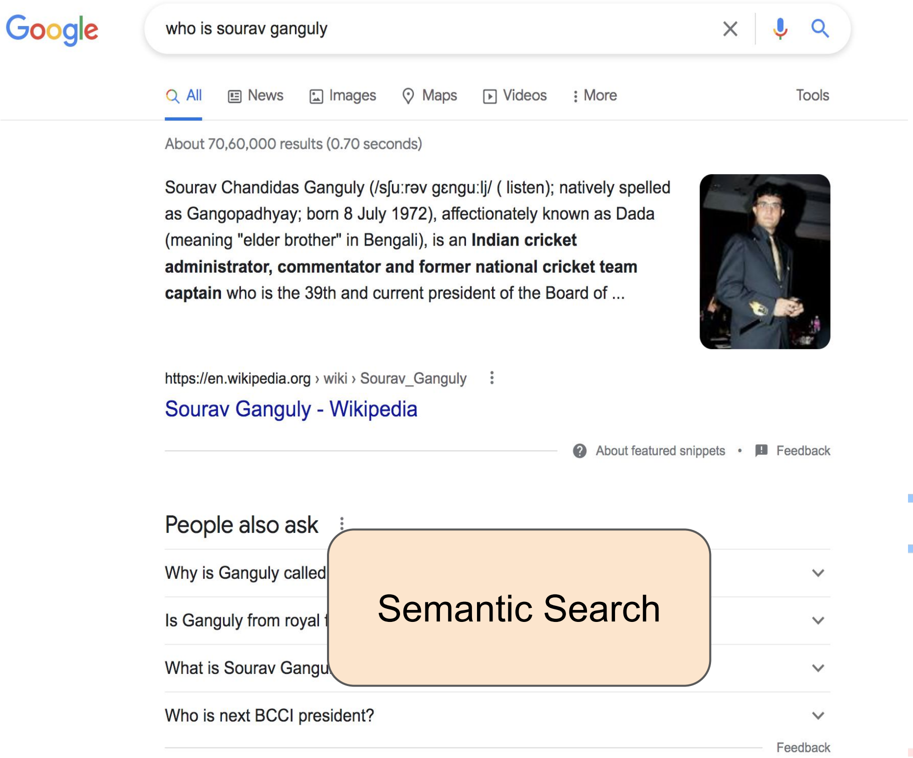
ore

opes

g...

Also in the news

Keyword Search



About 70,60,000 results (0.70 seconds)

Sourav Chandidas Ganguly (/sʃuːrəv gɛnguːlj/ (listen); natively spelled as Gangopadhyay; born 8 July 1972), affectionately known as Dada (meaning "elder brother" in Bengali), is an **Indian cricket administrator, commentator and former national cricket team captain** who is the 39th and current president of the Board of ...

https://en.wikipedia.org › wiki › Sourav_Ganguly

Sourav Ganguly - Wikipedia

About featured snippets • Feedback

People also ask

Why is Ganguly called

Is Ganguly from royal f

What is Sourav Gangu

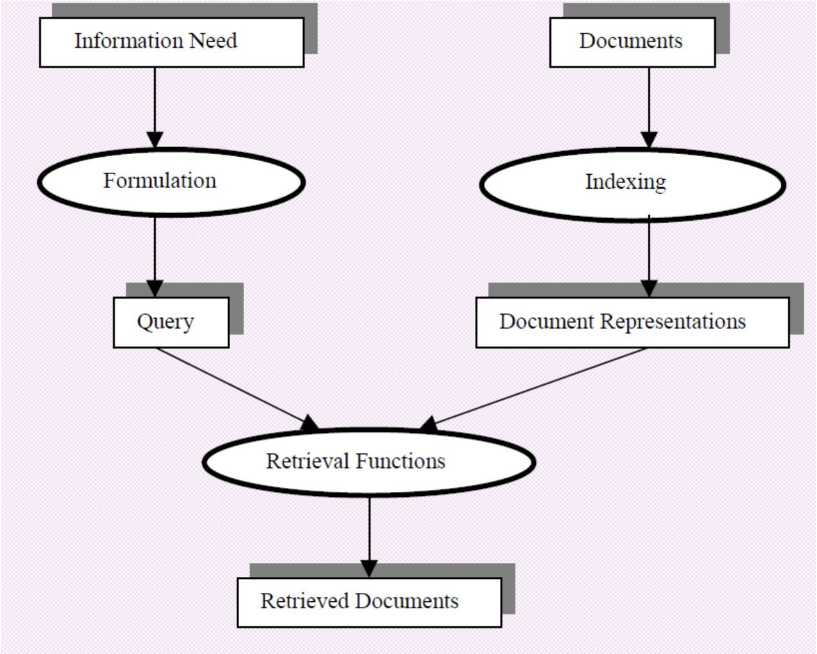
Who is next BCCI president?

Feedback

Semantic Search

Underlying problem?

Information Retrieval



A systematic approach to information retrieval, Lalmas et al. 2001

AI systems = Code + Data

Data for ad hoc IR

1. Corpus of search queries
2. Corpus of candidate documents
3. Ground truth (explicit or implicit)

Solution Architecture

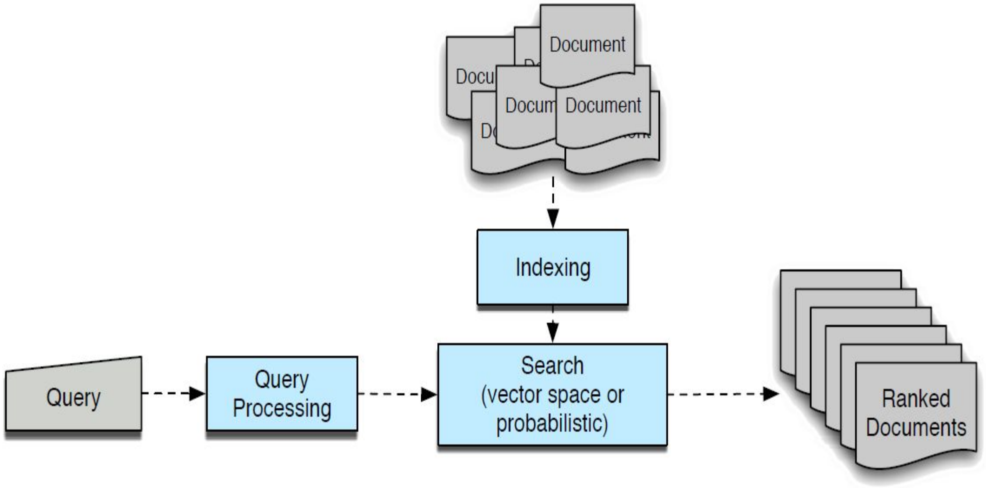
Semantic Search Architecture

Offline steps:

1. Document representation
2. Document indexing

Online steps:

1. Query representation
2. Candidate retrieval
3. Candidate ranking



Text representation

- Sparse features: 1 hot representations of characters (TFIDF), character n-grams
- Dense representations: Term or pre-trained embeddings

How to choose a representation ?

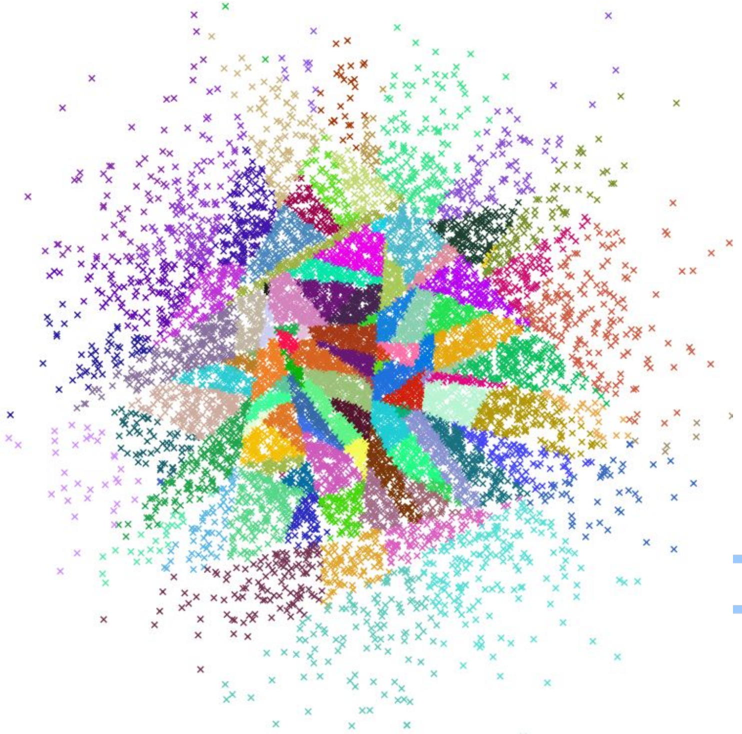
How to choose the right representation?

- Depends on size, property and domain of the data available
- DNNs for IR can learn text representations in situ, or use pre-trained embeddings.
- For pre-trained embeddings, make sure the representations are suitable for the task.

Embedding based models often make different errors than exact matching models and the combination of two may be more effective.

Document Indexing

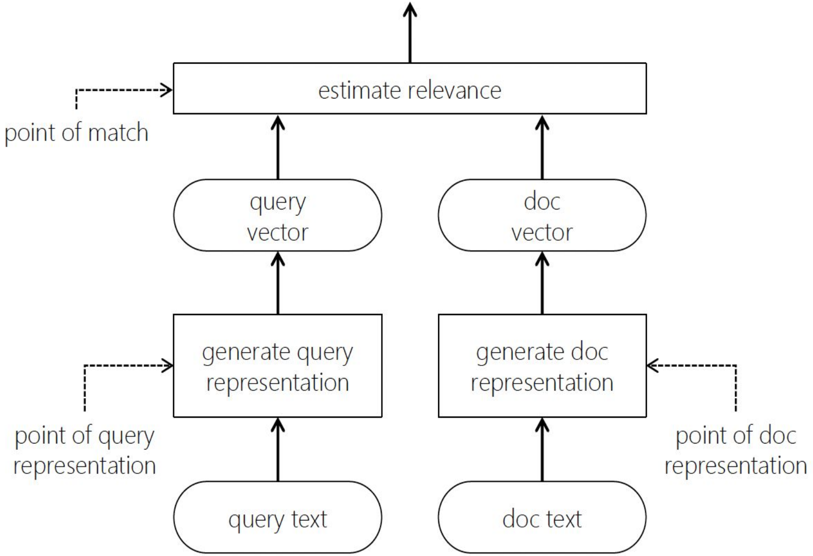
Linking or tagging documents with certain attributes which makes discovering them easier.



Candidate retrieval

- Retrieve matching candidates to the query on the basis of some measure of relevance.

- Rank the retrieved documents by the relevance score.


estimate relevance
point of match
query vector
doc vector
generate query representation
generate doc representation
point of query representation
point of doc representation
query text
doc text

Bhaskar Mitra and Nick Craswell (2018), “An Introduction to Neural Information Retrieval”,

No one size fits all!

Retrieval Approaches

Unsupervised

- No labelled data available
- Use existing retrieval model or query-document similarity metric (edit distance)

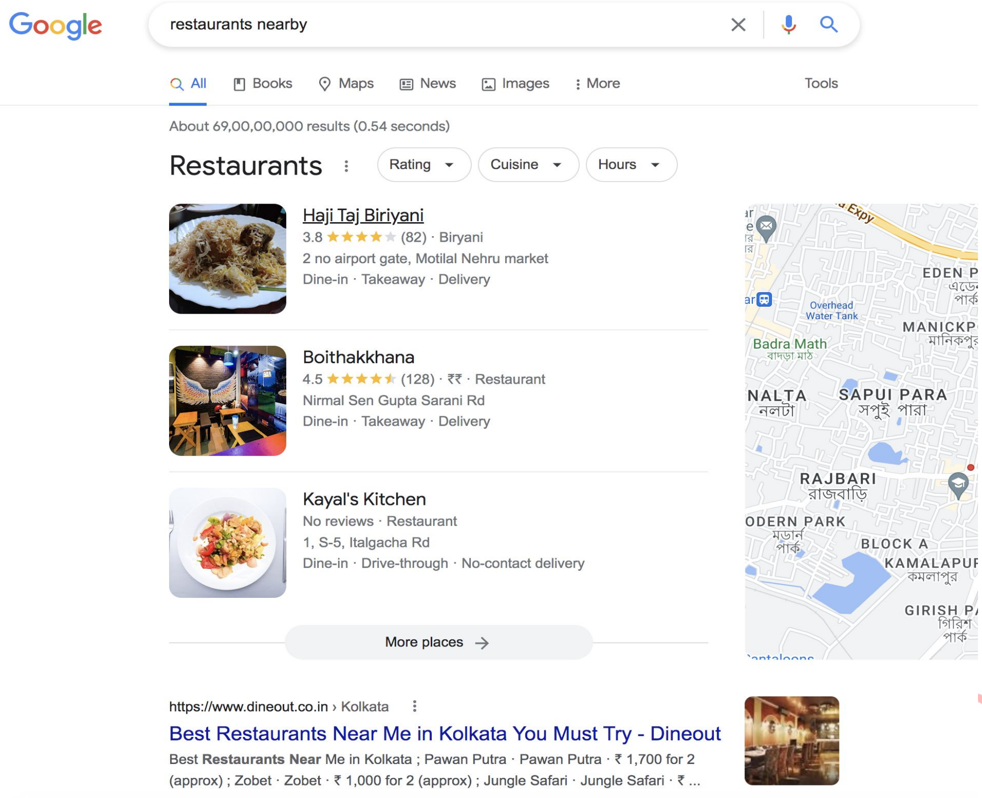
Semi-supervised

- Small amount of labelled data available
- Train a retrieval model with few parameters

Fully Supervised

- Large amount of labelled data available
- Train a model that optimizes directly for the target task

Ranking



restaurants nearby

About 69,00,00,000 results (0.54 seconds)

Restaurants

Haji Taj Biriyani
3.8 ★★★★ (82) · Biryani
2 no airport gate, Motilal Nehru market
Dine-in · Takeaway · Delivery

Boithakkhana
4.5 ★★★★★ (128) · ₹₹ · Restaurant
Nirmal Sen Gupta Sarani Rd
Dine-in · Takeaway · Delivery

Kayal's Kitchen
No reviews · Restaurant
1, S-5, Italgacha Rd
Dine-in · Drive-through · No-contact delivery

https://www.dineout.co.in › Kolkata

Best Restaurants Near Me in Kolkata You Must Try - Dineout

Best **Restaurants Near** Me in Kolkata ; Pawan Putra · Pawan Putra · ₹ 1,700 for 2 (approx) ; Zobet · Zobet · ₹ 1,000 for 2 (approx) ; Jungle Safari · Jungle Safari · ₹ ...

Ranking Approaches

- No labelled data available ⇒ Unsupervised
- Labelled data (ranked ground truth) available ⇒ Supervised Learning to rank (LTR)

You can’t improve what you can’t measure!

Metrics

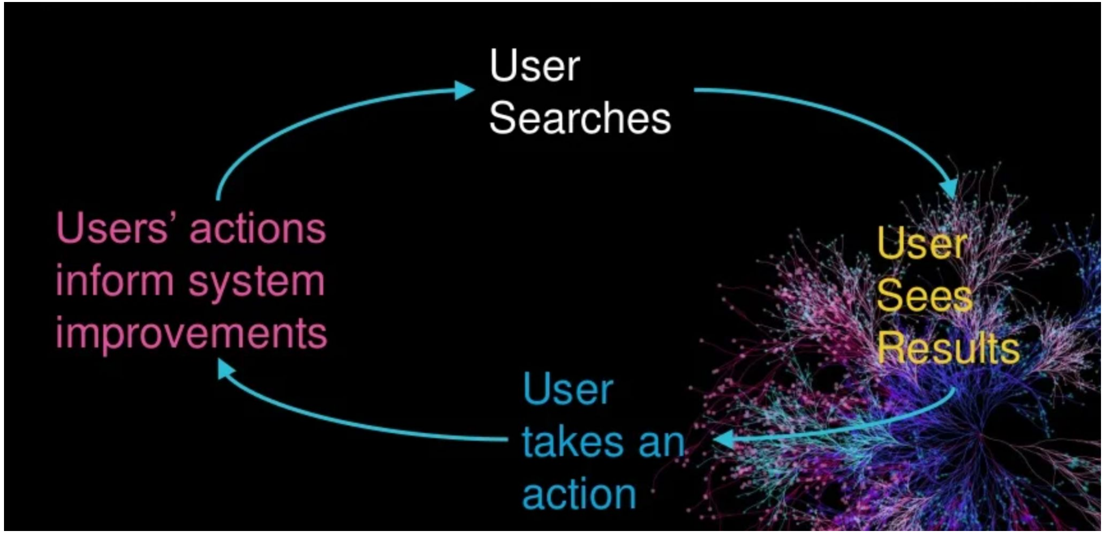
Offline

- MAP (Mean Average Precision)
- MRR (Mean Reciprocal Rank)
- NDCG (Normalized Discounted Cumulative Gain)

Online

- A/B testing
- Click through rate
- View rate
- Any other business metric

Importance of feedback loops



Google Developer Groups
Cloud • Kolkata

Why is search so difficult?

Unstructured Data

- Data drift
- Spelling errors
- Domain adaptation
- Multilingual search queries
- Lack of data (labelled / unlabelled)
- Out-of-scope queries
- No match found

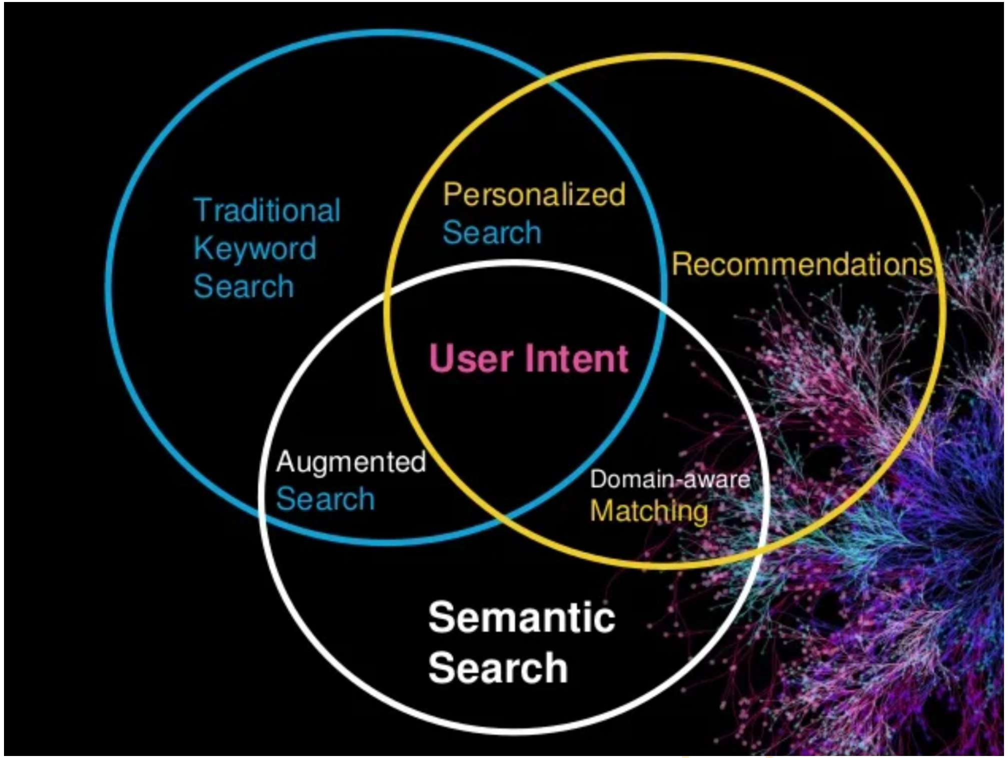
Tips to improve search

- **Monitor data drift** and use **active learning** feedback loops
- **Autocomplete feature** to prevent spelling mistakes
- Use **transfer learning** to fine tune large pre-trained models with unlabelled domain specific data
- **Analyse search query data** to gauge the need of a multilingual embedding model (Use 80/20 principle!)
- Explore **data augmentation/synthetic data** creation techniques
- Handle **no match found/out-of-scope** queries gracefully

Tips to improve search

- **Diagnose/improve content** when users aren't clicking on search results
- **Feature sought after content prominently** for everyone to see (most searched content)
- Analyze the page in the app the user was visiting just before they started searching, did the page provide the information a user searched for, **poor UX design**?

Beyond Semantic Search



Thank you!

Questions?

@shweta_bhatt8