RAG + IR Applications

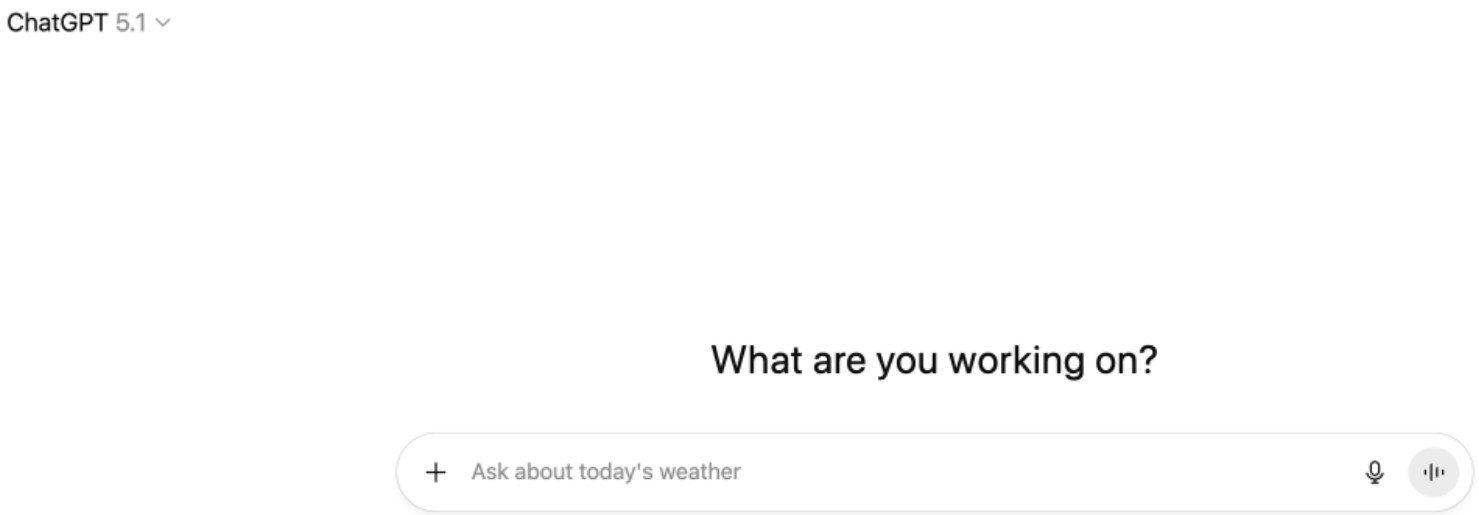
SI 650 / EECS 549
November 19, 2025

Slides based on Gao et al., 2023

Lecture Plan

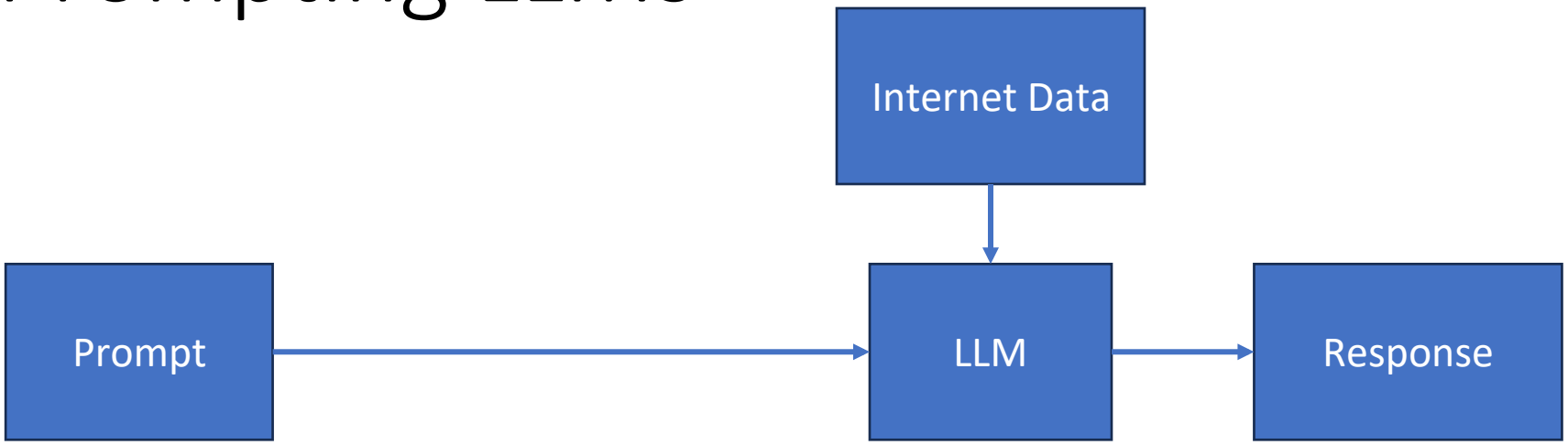
- RAG
- IR Applications + Tie-ins
  - How to apply what you know to new settings
  - Or, where other topics intersect in interesting ways

Think of ChatGPT



What could go wrong with you asking ChatGPT about today’s weather?

Prompting LLMs

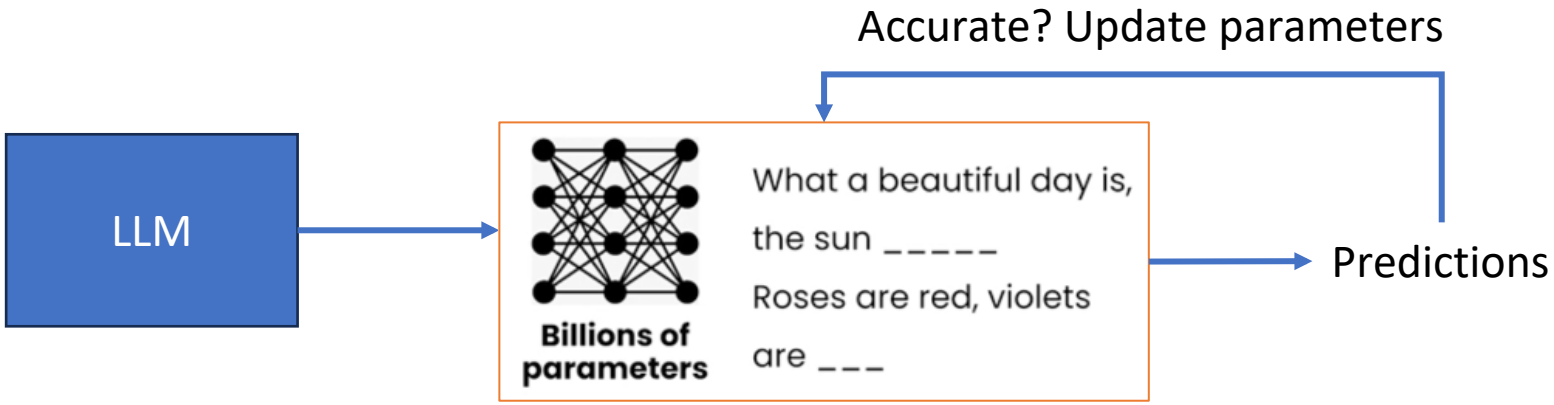


“What is the weather today?”

”Today the weather is ….”

This is an auto-regressive task.
The model is trying to predict the next token. And then the next …
The model is trained on large amounts of data and has billions of parameters but it surely doesn’t include up-to-date data on today’s weather

Training LLMs



- **LLMs generate probable word sequences**

LLMs just reproduce statistical patterns from their training data.

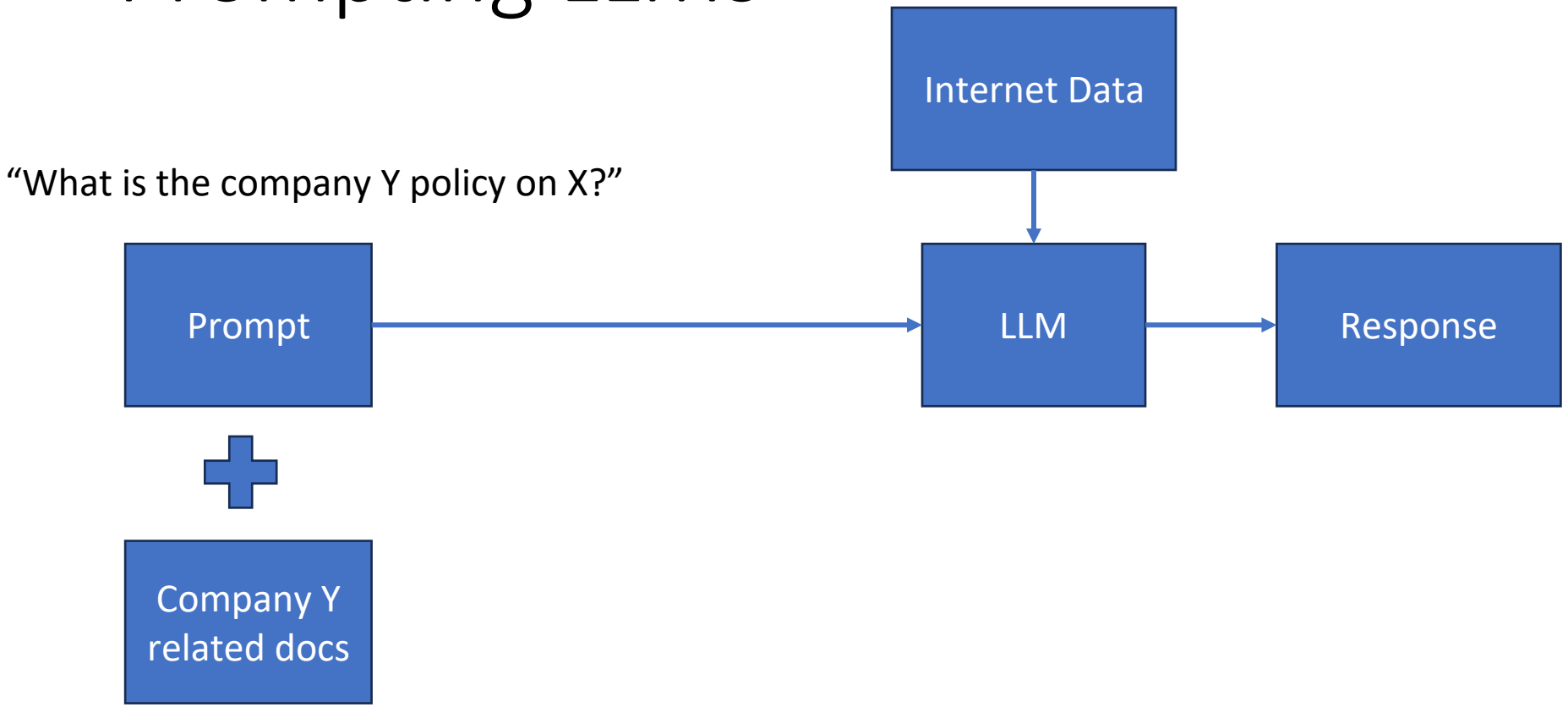
- **Knowledge gaps cause inaccurate responses**

Responses can “sound right” but aren’t true.

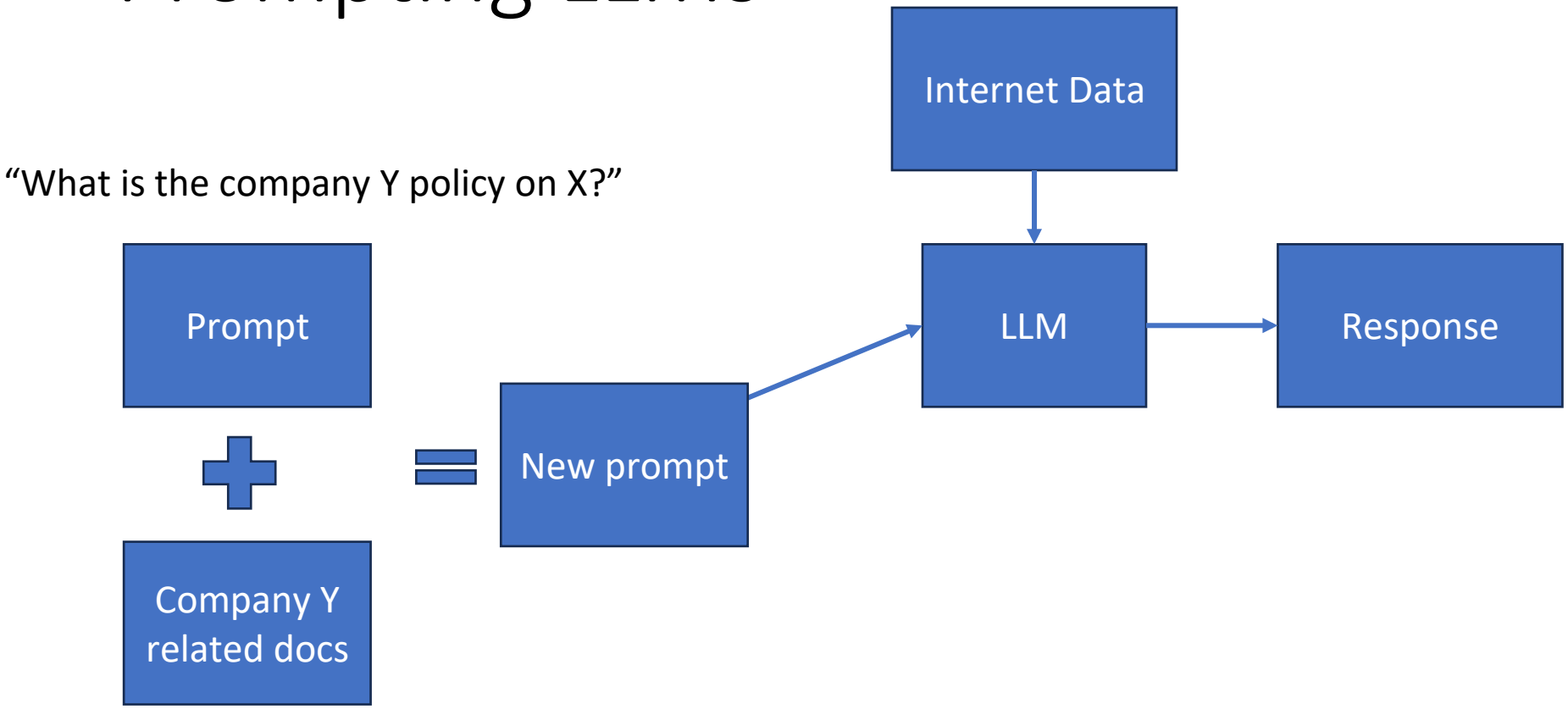
- LLMs are designed to generate “probable” text, not truthful text.

So, what can we do?

Prompting LLMs



Prompting LLMs



This is what happens under CAG (Cache Augmented Generation). What are some limitations of this technique?

Why not CAG?

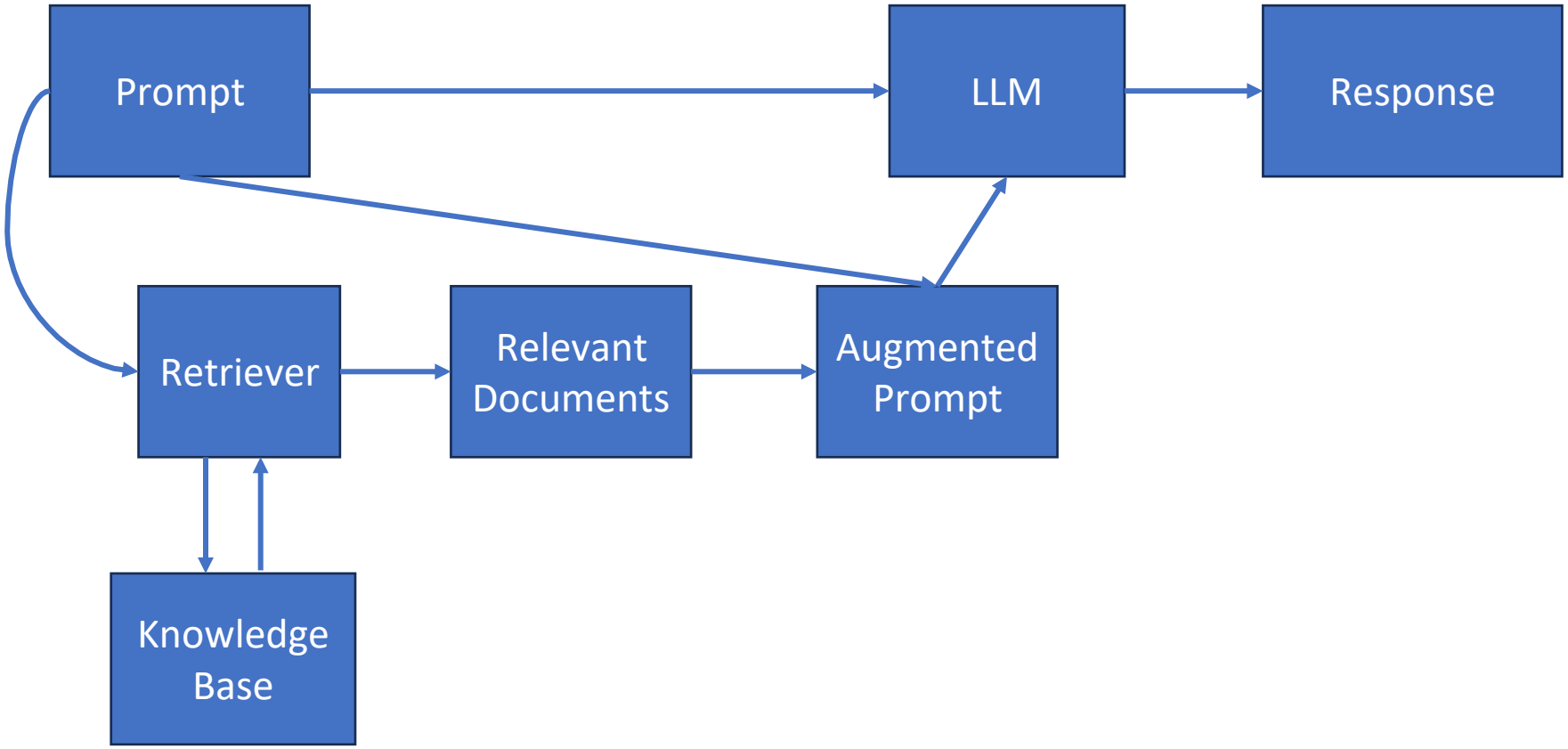
Higher Computational Cost

- Longer prompts take more computation to run
- Model performs computationally complex scan of every token
- Scan happens before generating each new token

Context Window Limit

- Eventually you hit the limit of LLM's context window
- Smaller models: only a few thousand tokens
- Largest models: millions

An Alternative: RAG



Advantages of RAGs

- **Injects missing knowledge**
  Adds info not in the training data (e.g., policies, updates)
- **Reduces hallucinations（幻觉）**
  Grounds answers with relevant context
- **Keeps models up to date**
  Reflects new info by updating the knowledge base
- **Enables source citation**
  Includes sources for verifiable answers
- **Focuses model on generation**
  Retriever finds facts, LLM writes responses

Applications of RAGs

Code Generation. Why?

- **LLM needs your project's context:** E.g. Classes, functions, definitions, and coding style

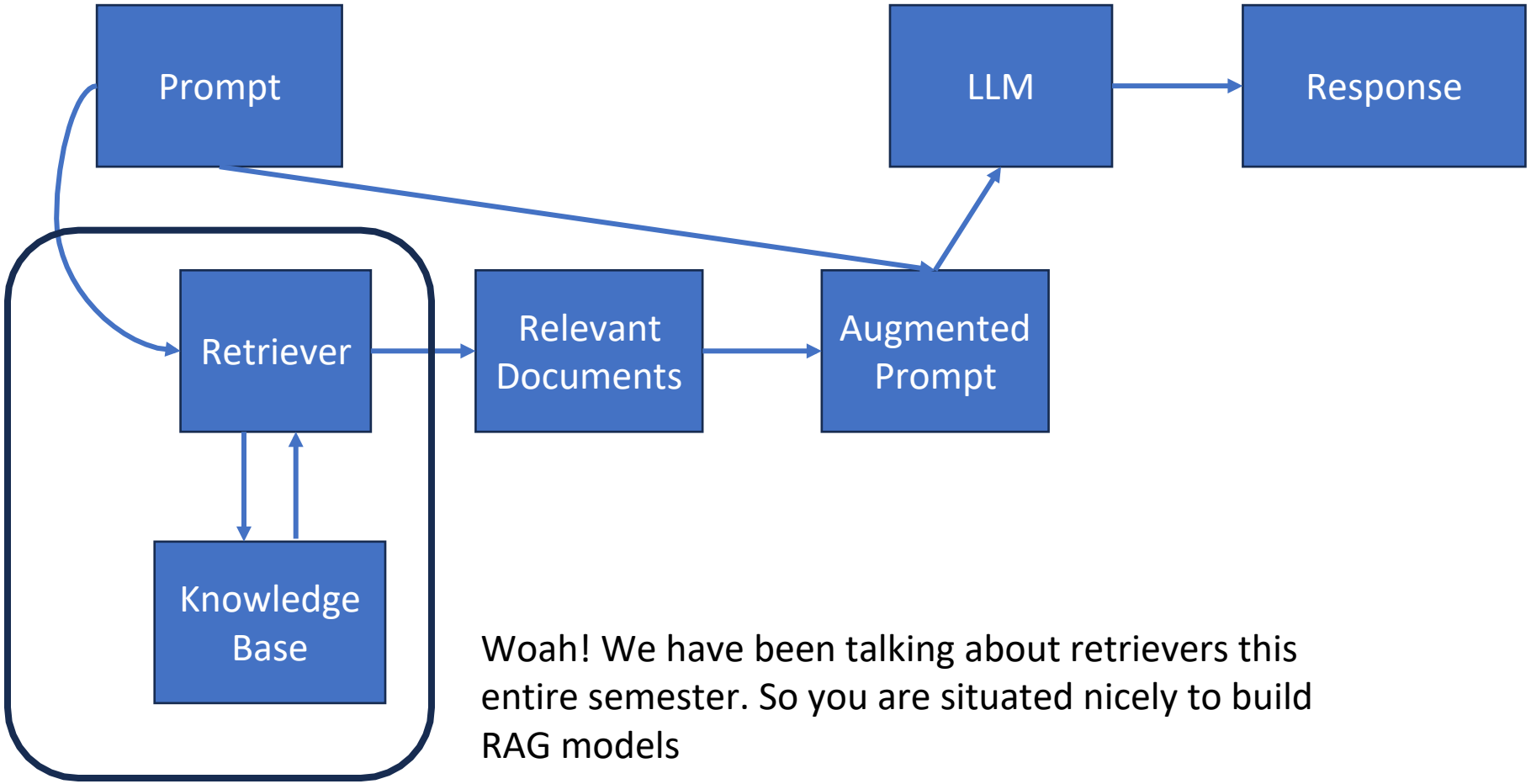
Company Chatbots. Why?

- **Tailored to your company** Every business has its own products, policies, and internal docs

Online Search. Why?

- Because people are lazy ☺

An Alternative: RAG



Woah! We have been talking about retrievers this entire semester. So you are situated nicely to build RAG models

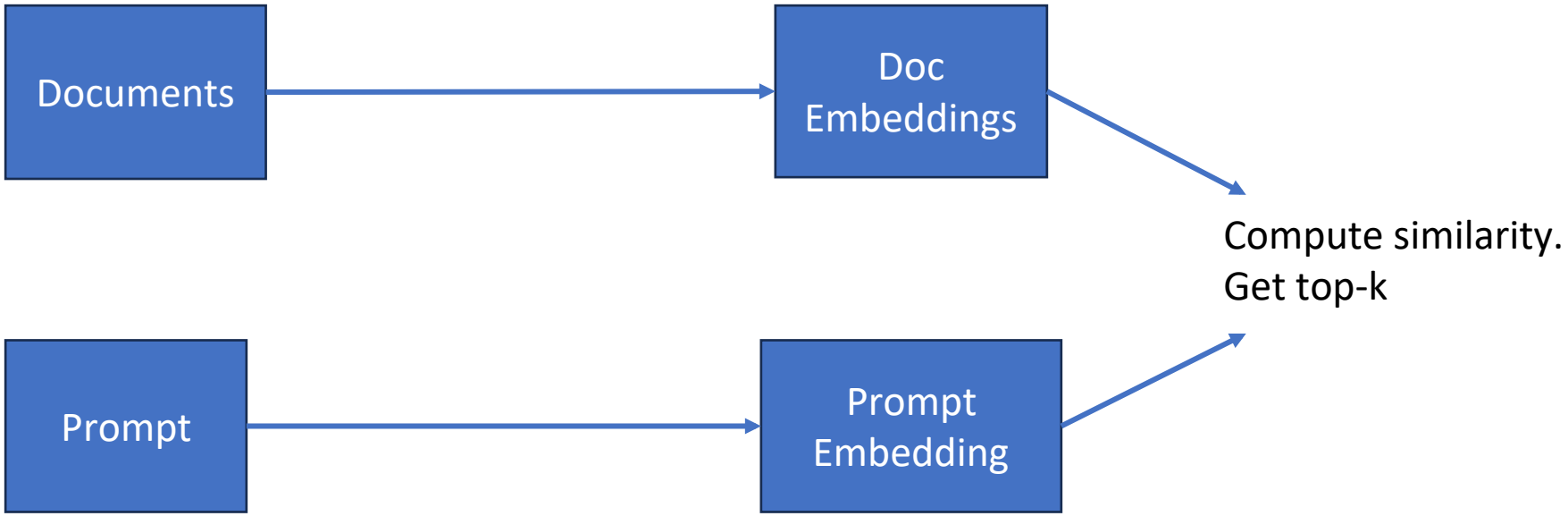
Ideas for the retriever based on your knowledge of SI650?

What kinds of retrievers (IR systems) have we covered so far?

- Vector Space Models (Sparse representations)
- Probabilistic Models
- Deep Neural Net models (Dense representations)

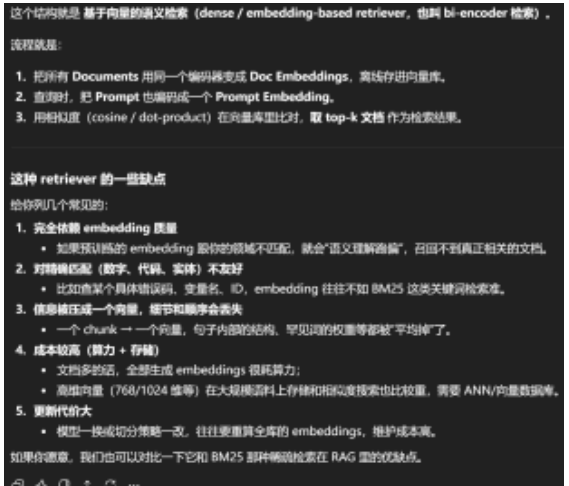
You could indeed build a simple BM25 based retriever. But as we know, these systems miss documents that are semantically relevant but don't include the original query words. So the state of the art uses embeddings (dense representations)

Retriever Ideas



We have done this before. What is this?

And what are some disadvantages?


这个结构就是 基于向量的语义检索（dense / embedding-based retriever，也叫 bi-encoder 检索）。

流程就是：

1. 把所有 Documents 用同一个编码器变成 Doc Embeddings，离线存进向量库。
2. 查询时，把 Prompt 也编码成一个 Prompt Embedding。
3. 用相似度（cosine / dot-product）在向量库里比对，取 top-k 文档 作为检索结果。

这种 retriever 的一些缺点

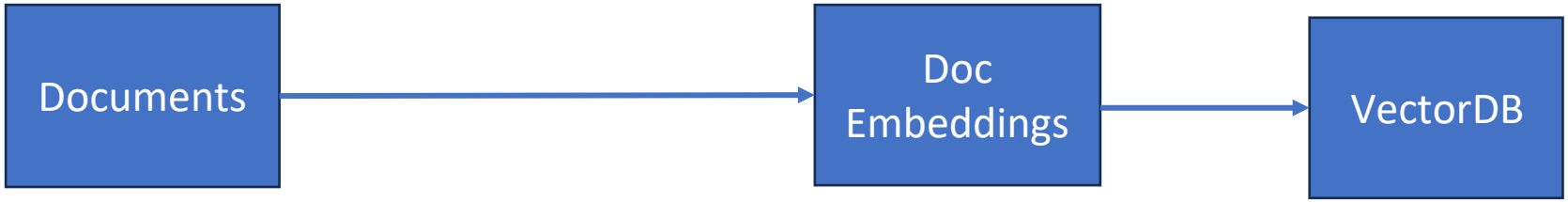
给你列几个常见的：

1. 完全依赖 embedding 质量
   - 如果预训练的 embedding 跟你的领域不匹配，就会"语义理解跑偏"，召回不到真正相关的文档。
2. 对精确匹配（数字、代码、实体）不友好
   - 比如查某个具体错误码、变量名、ID，embedding 往往不如 BM25 这类关键词检索准。
3. 信息被压成一个向量，细节和顺序会丢失
   - 一个 chunk → 一个向量，句子内部的结构、罕见词的权重等都被"平均掉"了。
4. 成本较高（算力 + 存储）
   - 文档多的话，全部生成 embeddings 很耗算力；
   - 高维向量（768/1024 维等）在大规模语料上存储和相似度搜索也比较重，需要 ANN/向量数据库。
5. 更新代价大
   - 模型一换或切分策略一改，往往要重算全库的 embeddings，维护成本高。

如果你愿意，我们也可以对比一下它和 BM25 那种稀疏检索在 RAG 里的优缺点。

Retriever Ideas

VectorDB is a *search engine* for vectors: optimized, structured, and engineered so you can find the most relevant vectors **fast**, **accurately**, and at **scale**.



VectorDBs store vectors in an index optimized for similarity search

- Builds specialized structures (IVF, **HNSW**, PQ, etc.)
- Fast approximate nearest neighbor (ANN) search in high-dimensional space

Retrieves the top-k closest vectors efficiently

If comparing the query to each document vector: O(n) per query

Vector DBs reduce this to roughly O(log n)

Supports production-grade features such as

- Persistence, scaling, replication
- Distributed search
- Hybrid search (dense + sparse)
- Monitoring
- Consistency guarantees

These are what make them “databases,” not just “vector stores.”

Hierarchical Navigable Small World Graph

- Graph-based index
- Organizes vectors into multiple layers of a small-worl
  - Every ve
    appear in
    on.
- Extremel
- The defa

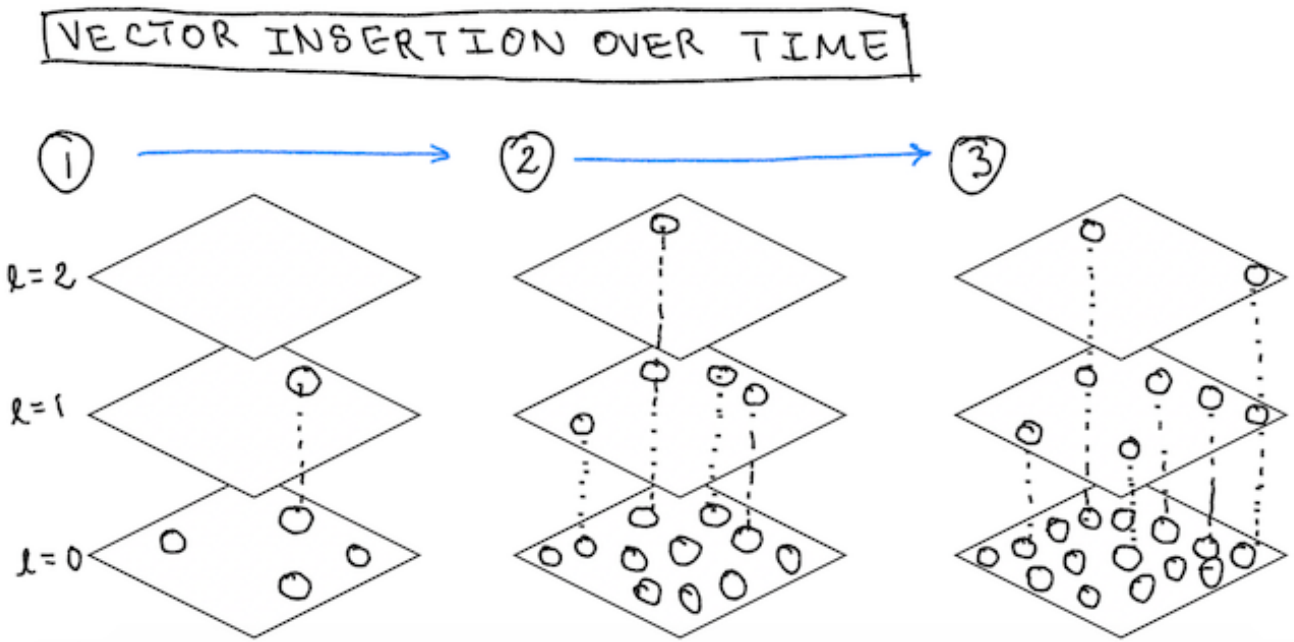


Image credit: https://medium.com/@amallya0523/hnsw-algorithm-efficiently-searching-vector-databases-9276a934393d

要点就是：

- 它是一种 **基于图的向量索引结构。**
- 把所有向量组织成一个 **多层的小世界网络**：
  - 每个向量一定在 **level 0**（最底层）；
  - 只有一小部分会被“提升”到 **level 1**；
  - 更少的一部分出现在 **level 2**，再往上就更稀疏。
- 这样层次结构让搜索时可以：
  1. 先在高层“粗略导航”找到大概区域；
  2. 再往下层一层层细找最近邻，
     所以 **又快、召回率又高。**
- 因为效果好、速度快，**现在大多数向量数据库的默认索引都是 HNSW 或类似结构。**

小世界网络（Small-World Network）大概就是：
大多数点彼此之间只隔很少几步，但每个点又只连着少数邻居 的那种网络结构。

Hierarchical Navigable Small World Graph

- Graph-based index
- Organi
small-wc
  - Every
  appear
  on.
- Extrem
- The de

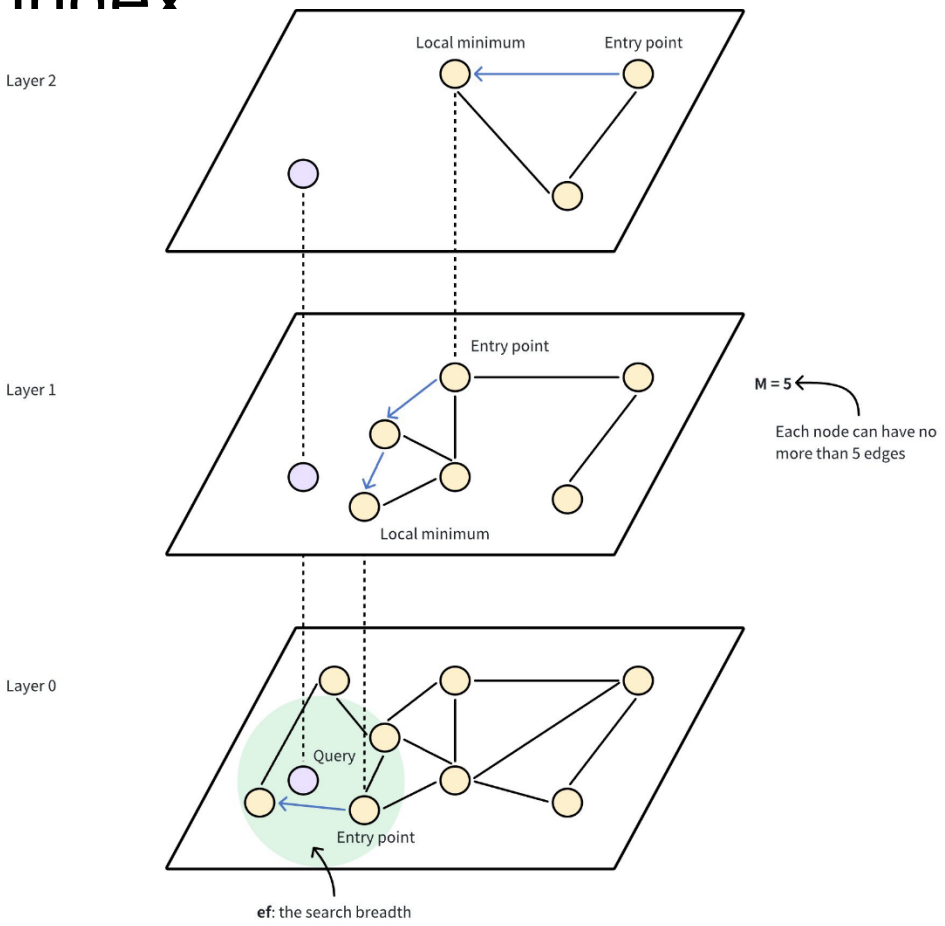
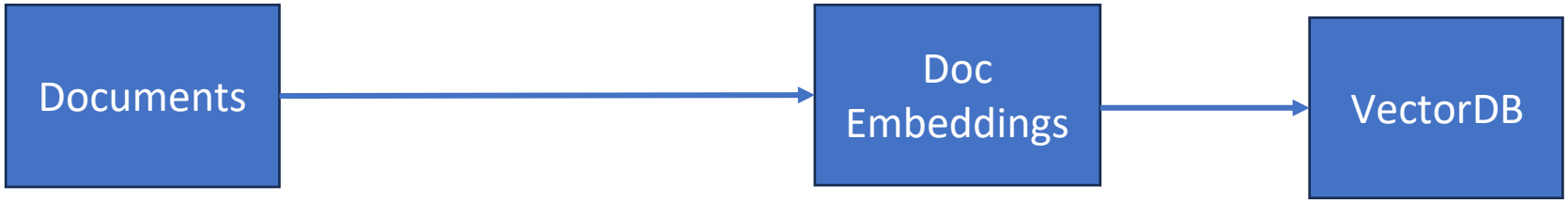


Image credit: https://milvus.io/docs/hnsw.md

Retriever Ideas

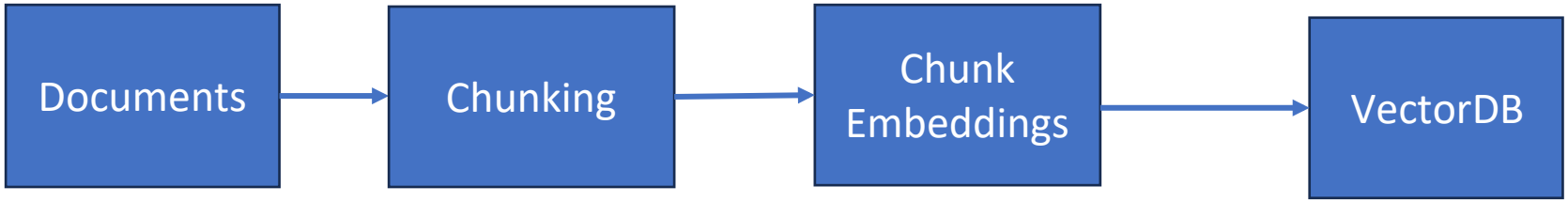


Any issues with creating embeddings for entire documents?

- Imagine a single embedding of a 20 page PDF. It will be fuzzy
- Embeddings behave best for semantically coherent units
- Scale issue: Even with 200k+ context windows, you *don't* want to retrieve entire long documents.

Solution: Chunking

Retriever Ideas



Any issues with creating embeddings for entire documents?

- Imagine a single embedding of a 20 page PDF. It will be fuzzy
- Embeddings behave best for semantically coherent units
- Scale issue: Even with 200k+ context windows, you *don't* want to retrieve entire long documents.

Solution: Chunking

Chunking

How to choose chunk sizes?

- **If too large:** many relevant sentences get bundled together
- **If too small:** the system retrieves lots of fragments that *sound like* the query but don't contain the real answer.

Most practical RAG systems tune chunk size around 200–500 tokens, with some form of *overlap;*

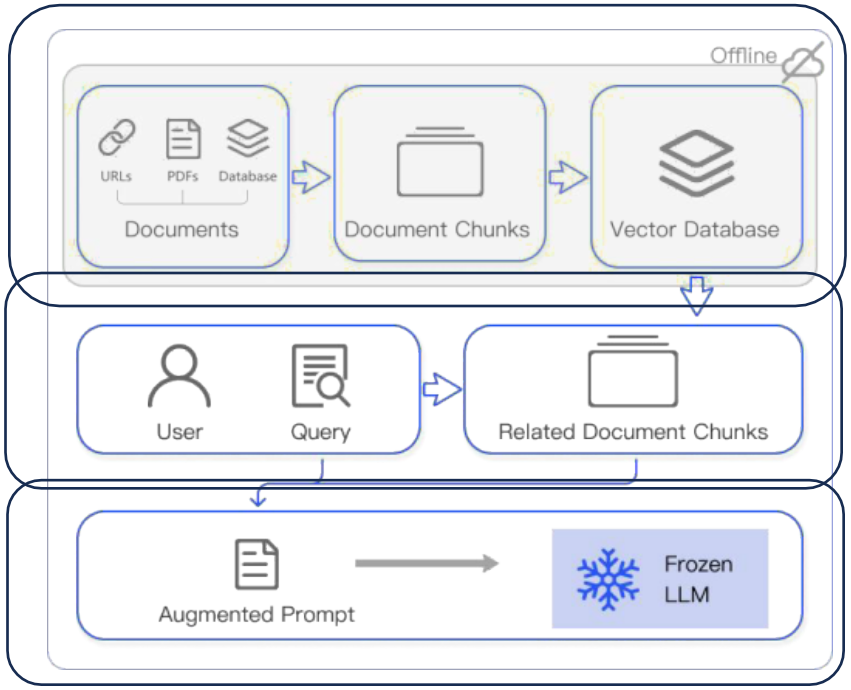
Overlap (e.g., 20–30%) ensures: Explanations that span two paragraphs don't get severed. Retrieval remains robust to where you cut.

Chunk-level retrieval also lets you log which evidence was used and helps you debug errors

Naïve RAG

Step1 Indexing

1. Divide the document into even chunks, each chunk being a piece of the original text.
2. Using the encoding model to generate an embedding for each chunk.
3. Store the Embedding of each block in the vector database.



Step2 Retrieval

Retrieve the k most relevant using vector similarity search.

Step3 Generation

The original query and the retrieved text are combined and input into a LLM to get the final answer

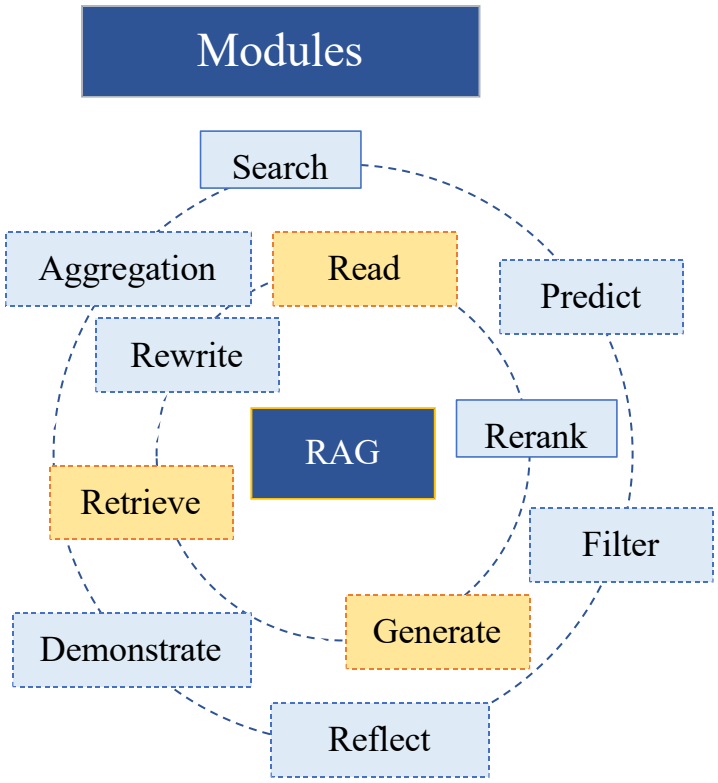
Slide from: https://github.com/Tongji-KGLLM/RAG-Survey

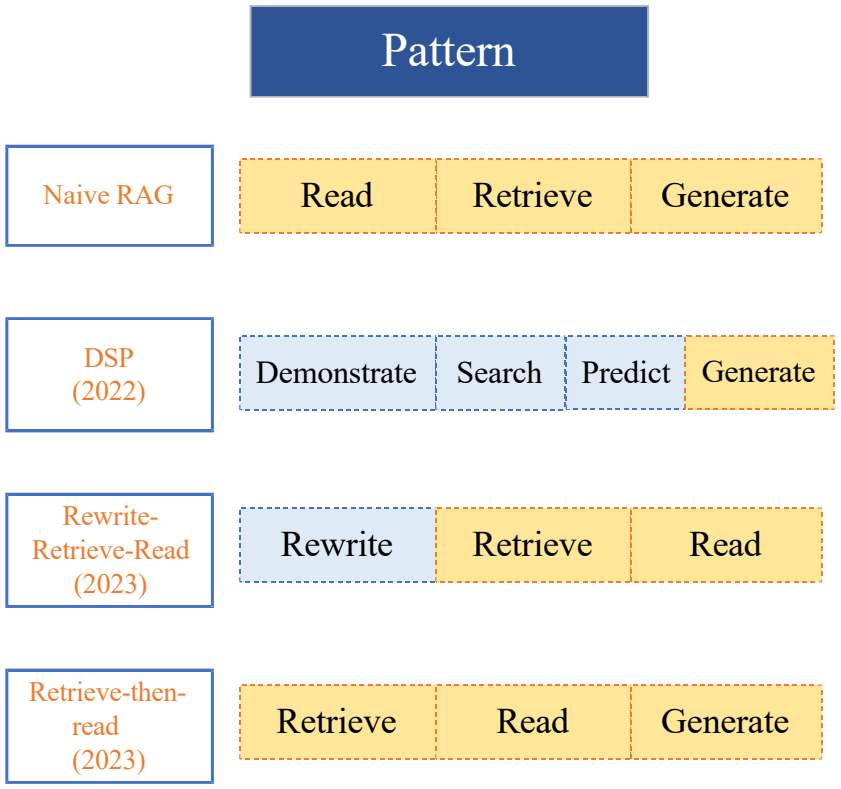
Advanced RAGs

Advanced RAG = optimize before and after retrieval so that the LLM sees only the exact evidence needed and the retrieval is faster. A few examples:

- **Better Indexing:** smarter chunking, multi-index structures
- **Pre-Retrieval Query Optimization:** query rewrite. query classification (e.g., lookup vs. reasoning)
- **Retrieval Optimization:** hybrid search (dense + sparse)
- **Post-Retrieval Optimization:** cross-encoder reranking, chunk summarization
- **Generation Optimization:** chain-of-thought with citations

Modular RAG

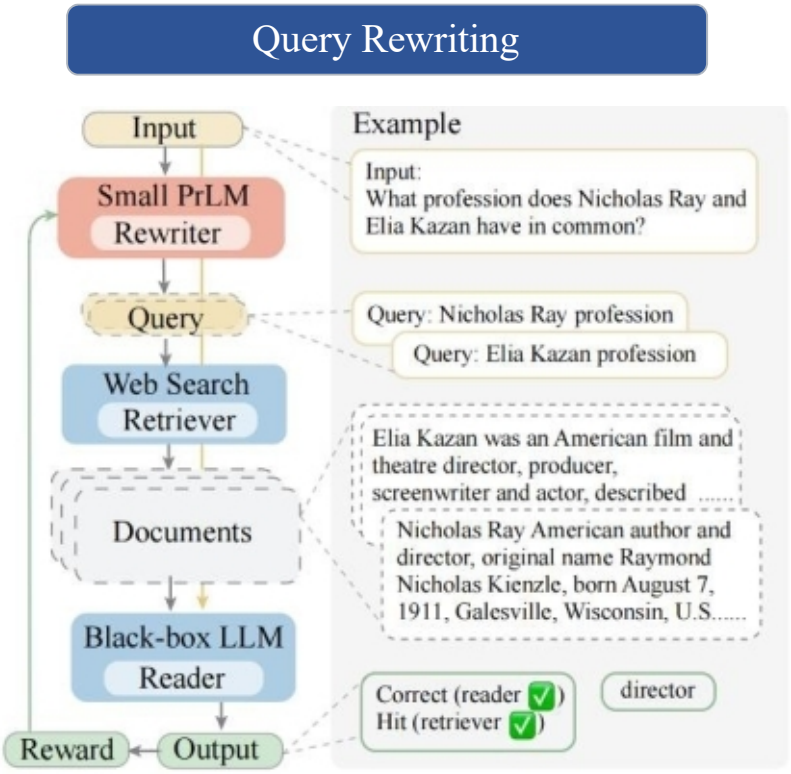




Slide from: https://github.com/Tongji-KGLLM/RAG-Survey

Techniques for better RAG: Query Rewriting

Questions and answers do not always possess high semantic similarity; adjusting the Query can yield better retrieval results.



Rewrite-Retrieve-Read [Ma et al., 2023]

Slide from: https://github.com/Tongji-KGLLM/RAG-Survey

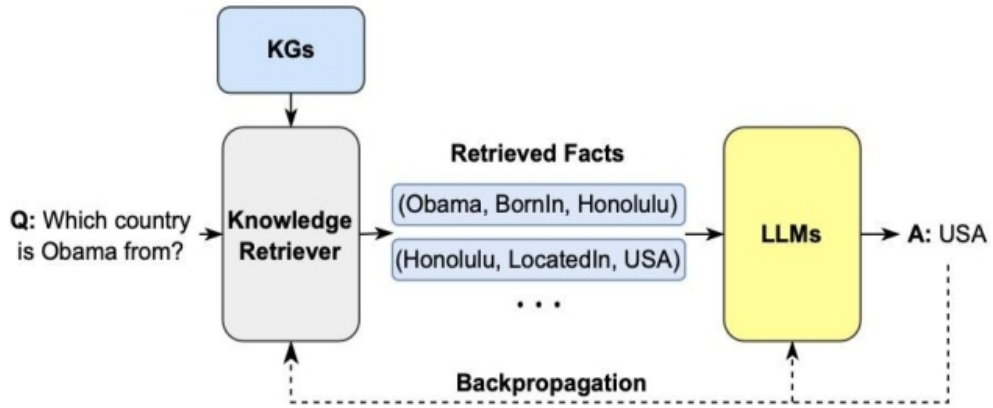
Techniques for better RAG: Knowledge Graphs

GraphRAG

- Extract entities from the user's input query, then construct a subgraph to form context, and finally feed it into the large model for generation.

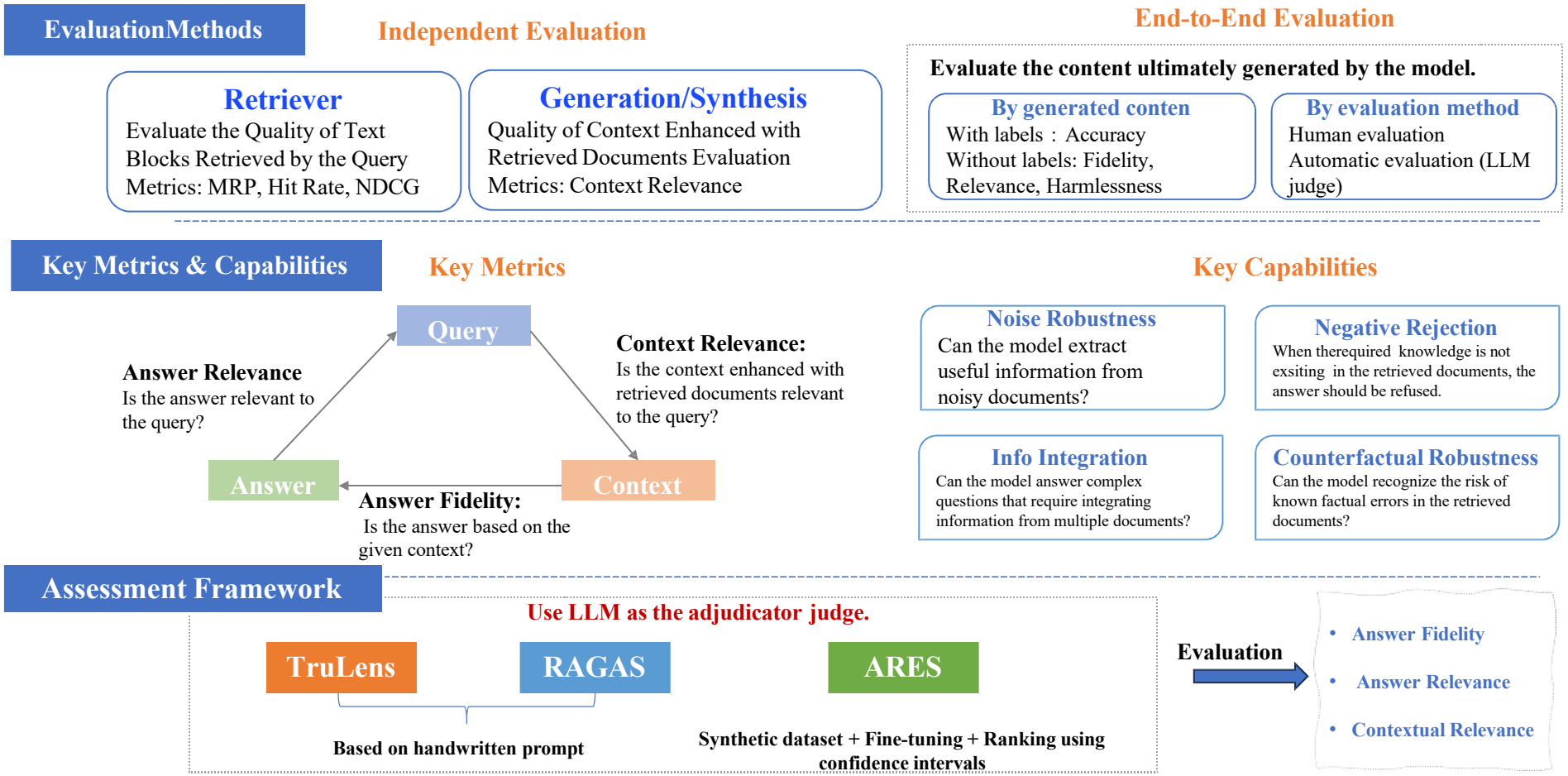
Implementation

- Use LLM (or other models) to extract key entities from the question.
- Retrieve subgraphs based on entities, delving to a certain depth, such as 2 hops or even more.
- Utilize the obtained context to generate answers through LLM.



Slide from: https://github.com/Tongji-KGLLM/RAG-Survey

Evaluating RAGs

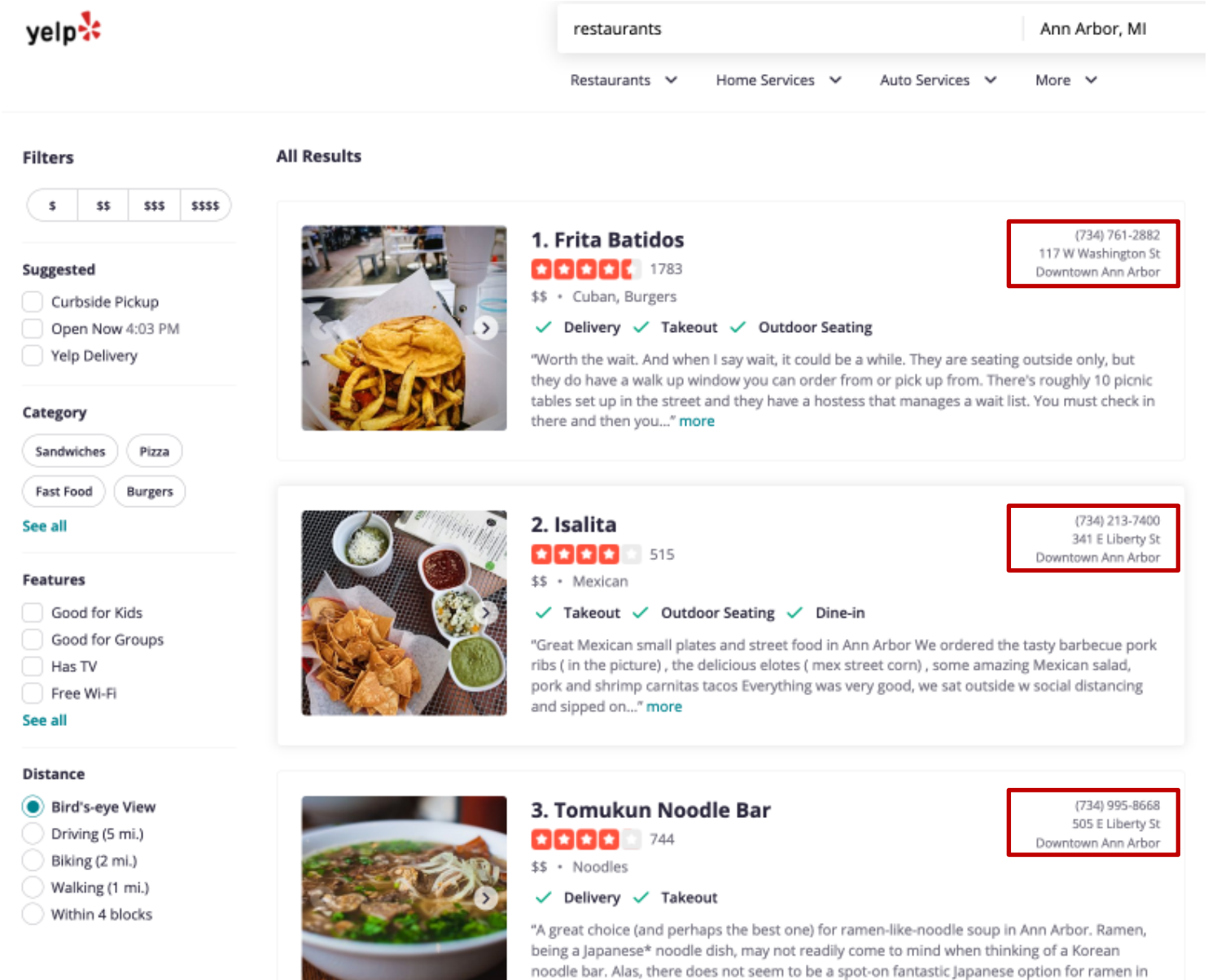


Slide from: https://github.com/Tongji-KGLLM/RAG-Survey

Quiz Time

IR applications and Tie-Ins

Information Extraction



Information Extraction

- Extracting database records from unstructured and semi-structured inputs
- Examples:
  - Recognizing names of people in text
  - Extracting prices from tables
  - Linking companies with products
  - Extracting addresses from emails
- Main steps:
  - Segmentation
  - Classification
  - Association
  - Clustering

FDA expands pet food recall

The nationwide pet food recall was expanded Wednesday to include products containing rice protein laced with melamine, a toxic agent, the Food and Drug Administration said.

Before this latest announcement, the FDA attributed pet illness and deaths to recalled pet food with wheat gluten found to contain melamine, a component of fertilizers and plastic utensils.

Also on Wednesday, Menu Foods, the company that recalled more than 60 million cans and pouches of wet cat and dog food on March 15, added one of its Natural Life brand products to its recall list. It added two product dates to eight of its already recalled pet foods.

The FDA has recorded 16 animal deaths related to the wheat gluten-pet food recall. However, other organizations have put the death toll in the thousands. After consumer complaints to Natural Balance of Pacoima, California, reporting kidney failure in several cats and dogs after eating the company's venison products, the firm issued a nationwide recall of its venison and brown rice canned and bagged dog foods and treats, and venison and green pea dry cat food, the FDA said.

FDA – organization
Food and Drug Administration - organization
Menu Foods – company
Natural Life – brand
Natural Balance – company
Pacoima – location
California - location

Landscape of IE Techniques

Lexicons

Abraham Lincoln was born in Kentucky.

Classify Pre-segmented Candidates

Abraham Lincoln was born in Kentucky.

Sliding Window

Abraham Lincoln was born in Kentucky.

Try alternate window sizes:

Boundary

BEGIN

Classifier

BEGIN END BEGIN END

ee Grammars

as born in Kentucky.

V P NP PP VP VP S

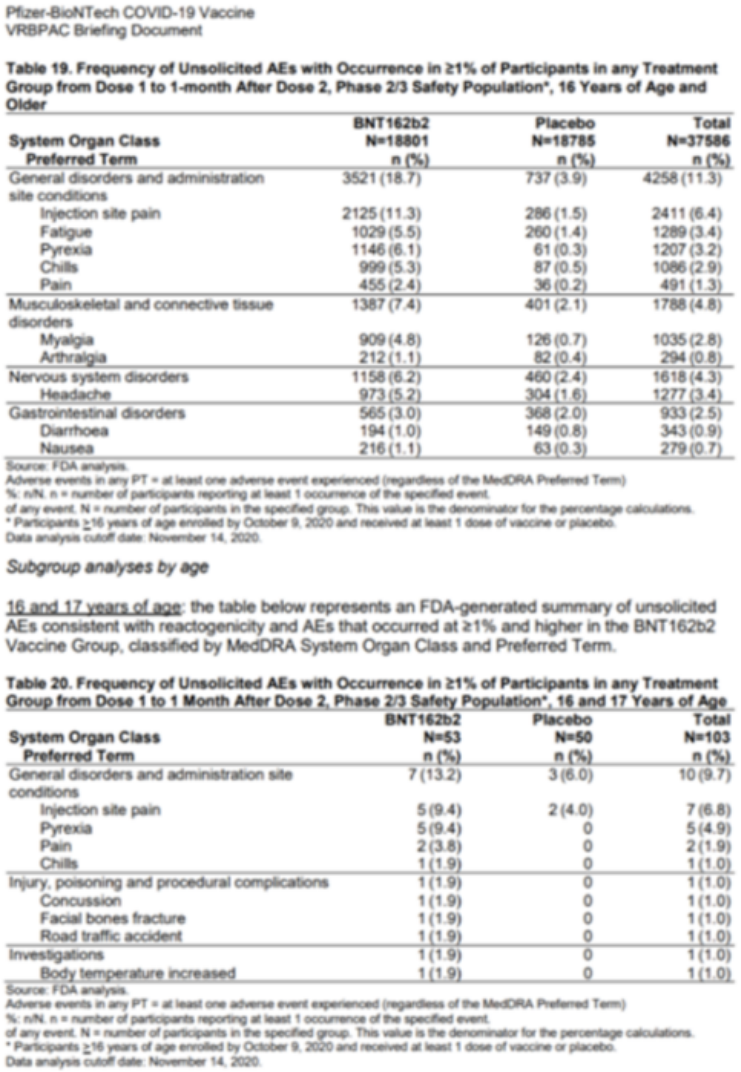
Most likely parse?

So much deep learning now

Slide by William Cohen

Information Extraction Example

- How do we extract relevant content and categorize them?



Pfizer-BioNTech COVID-19 Vaccine
VRBPAC Briefing Document

Table 19. Frequency of Unsolicited AEs with Occurrence in ≥1% of Participants in any Treatment Group from Dose 1 to 1-month After Dose 2, Phase 2/3 Safety Population*, 16 Years of Age and Older

System Organ Class Preferred Term	BNT162b2 N=18801 n (%)	Placebo N=18785 n (%)	Total N=37586 n (%)
General disorders and administration site conditions	3521 (18.7)	737 (3.9)	4258 (11.3)
Injection site pain	2125 (11.3)	286 (1.5)	2411 (6.4)
Fatigue	1029 (5.5)	260 (1.4)	1289 (3.4)
Pyrexia	1146 (6.1)	61 (0.3)	1207 (3.2)
Chills	999 (5.3)	87 (0.5)	1086 (2.9)
Pain	455 (2.4)	36 (0.2)	491 (1.3)
Musculoskeletal and connective tissue disorders	1387 (7.4)	401 (2.1)	1788 (4.8)
Myalgia	909 (4.8)	126 (0.7)	1035 (2.8)
Arthralgia	212 (1.1)	82 (0.4)	294 (0.8)
Nervous system disorders	1158 (6.2)	460 (2.4)	1618 (4.3)
Headache	973 (5.2)	304 (1.6)	1277 (3.4)
Gastrointestinal disorders	565 (3.0)	368 (2.0)	933 (2.5)
Diarrhoea	194 (1.0)	149 (0.8)	343 (0.9)
Nausea	216 (1.1)	63 (0.3)	279 (0.7)

Source: FDA analysis.
Adverse events in any PT = at least one adverse event experienced (regardless of the MedDRA Preferred Term)
%: n/N. n = number of participants reporting at least 1 occurrence of the specified event.
of any event. N = number of participants in the specified group. This value is the denominator for the percentage calculations.
* Participants ≥16 years of age enrolled by October 9, 2020 and received at least 1 dose of vaccine or placebo.
Data analysis cutoff date: November 14, 2020.

Subgroup analyses by age

16 and 17 years of age: the table below represents an FDA-generated summary of unsolicited AEs consistent with reactogenicity and AEs that occurred at ≥1% and higher in the BNT162b2 Vaccine Group, classified by MedDRA System Organ Class and Preferred Term.

Table 20. Frequency of Unsolicited AEs with Occurrence in ≥1% of Participants in any Treatment Group from Dose 1 to 1 Month After Dose 2, Phase 2/3 Safety Population*, 16 and 17 Years of Age

System Organ Class Preferred Term	BNT162b2 N=53 n (%)	Placebo N=50 n (%)	Total N=103 n (%)
General disorders and administration site conditions	7 (13.2)	3 (6.0)	10 (9.7)
Injection site pain	5 (9.4)	2 (4.0)	7 (6.8)
Pyrexia	5 (9.4)	0	5 (4.9)
Pain	2 (3.8)	0	2 (1.9)
Chills	1 (1.9)	0	1 (1.0)
Injury, poisoning and procedural complications	1 (1.9)	0	1 (1.0)
Concussion	1 (1.9)	0	1 (1.0)
Facial bones fracture	1 (1.9)	0	1 (1.0)
Road traffic accident	1 (1.9)	0	1 (1.0)
Investigations	1 (1.9)	0	1 (1.0)
Body temperature increased	1 (1.9)	0	1 (1.0)

Source: FDA analysis.
Adverse events in any PT = at least one adverse event experienced (regardless of the MedDRA Preferred Term)
%: n/N. n = number of participants reporting at least 1 occurrence of the specified event.
of any event. N = number of participants in the specified group. This value is the denominator for the percentage calculations.
* Participants ≥16 years of age enrolled by October 9, 2020 and received at least 1 dose of vaccine or placebo.
Data analysis cutoff date: November 14, 2020.

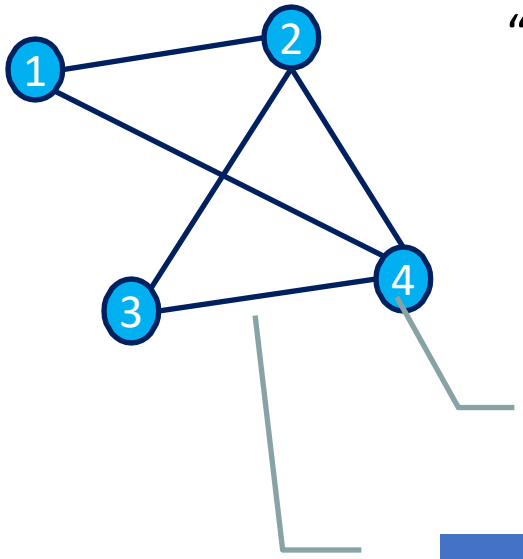
Information Extraction Pipeline

- Preprocessing:
  - Encode documents as graphs
- ML algorithm
  - Graph Neural Networks

Graphs

- Networks are collections of points joined by lines.

Mathematically (Topologically):
"Network" ≡ "Graph"



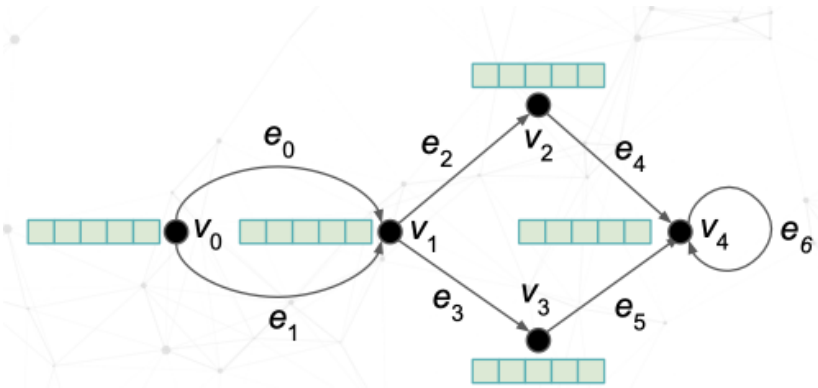
$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

Vertex, node, site, actor, …

Edge, arc, link, bond, tie, relation …

Graph Neural Networks

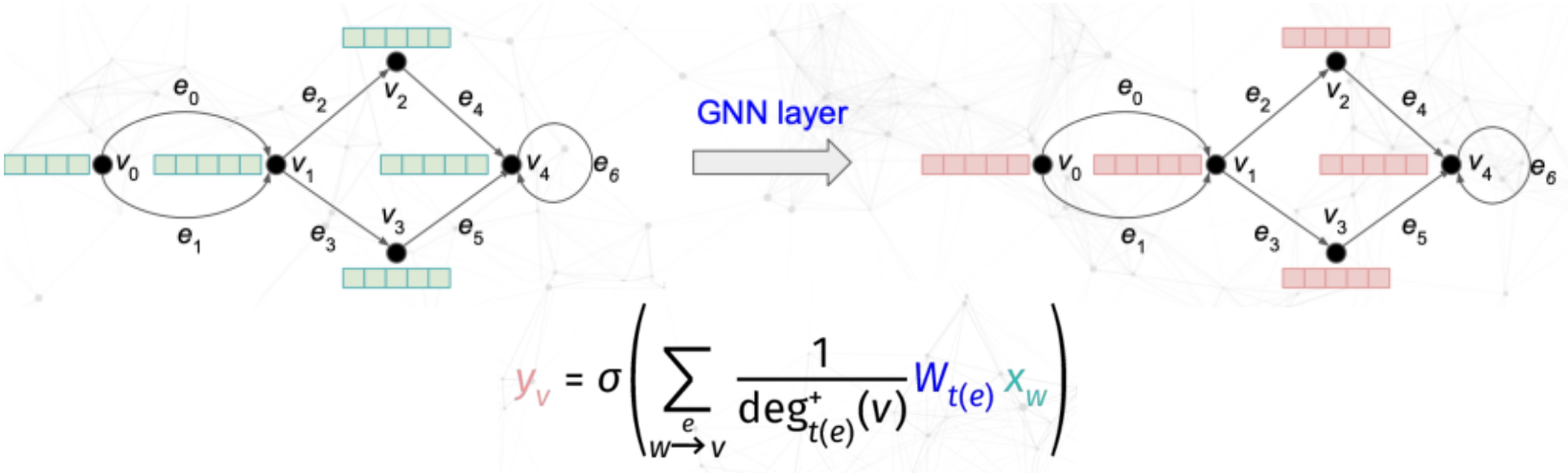
- Let's consider the node classification problem:
  - Input: a graph G of type Graph
  - Output: a labeling of its vertices, L: V → {0, 1}



GNN

```
labels = [
    0,   # v_0
    1,   # v_1
    0,   # v_2
    0,   # v_3
    1    # v_4
]
```

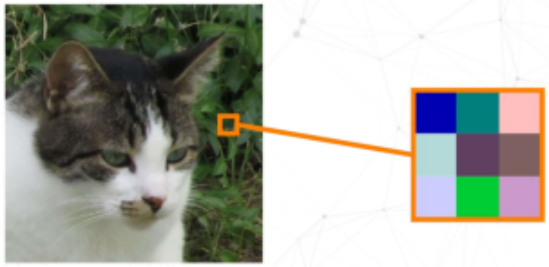
Graph Neural Networks



- y_v is the output feature vector of node v
- x_w is the input feature vector of node w
- t(e) is the type of the edge e: v → w
- W_t is a learned weight matrix corresponding to edge type t
- $\deg^+_t(v)$ is the number of incoming edges of type t for node v
- σ is an activation function

Schlichtkrull et at, “Modeling Relational Data with Graph Convolutional Networks”, arxiv:1703.06103

Useful parallel: Convolutional Networks



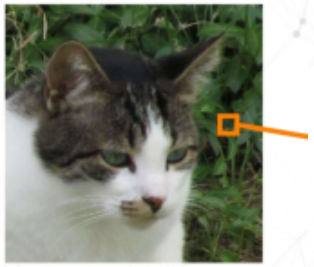
Original image A 3 × 3 patch

$$y_{i,j} = \sigma\left(\sum_{\substack{k=-1,0,1 \\ l=-1,0,1}} W_{k,l}\, x_{i+k,j+l}\right)$$

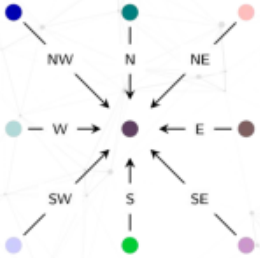
- yi, j is the output feature vector of pixel (i, j)
- xi, j is the input feature vector of pixel (i, j)
- Wk, l are learned weight matrices
- σ is an activation function

Schlichtkrull et at, “Modeling Relational Data with Graph Convolutional Networks”, arxiv:1703.06103

Useful parallel: Convolutional Networks



Original image



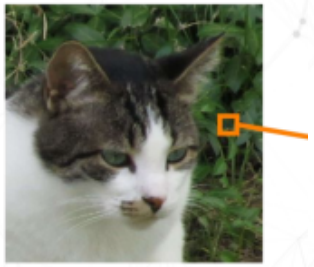
A graph representing the 3 × 3 patch

$$y_{i,j} = \sigma\left(\sum_{\substack{k=-1,0,1 \\ l=-1,0,1}} W_{k,l}\, x_{i+k,j+l}\right)$$

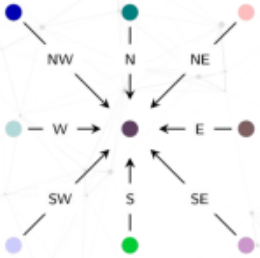
- $y_{i,j}$ is the output feature vector of pixel (i, j)
- $x_{i,j}$ is the input feature vector of pixel (i, j)
- $W_{k,l}$ are learned weight matrices
- σ is an activation function

Schlichtkrull et at, “Modeling Relational Data with Graph Convolutional Networks”, arxiv:1703.06103

Useful parallel: Convolutional Networks



Original image



A graph representing the 3 × 3 patch

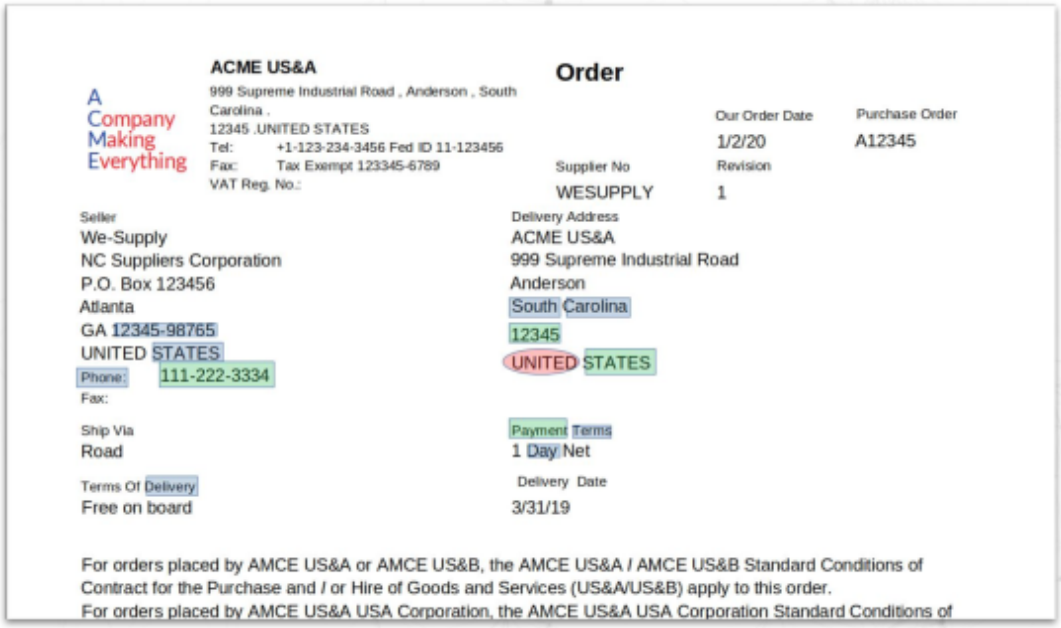
$$y_p = \sigma\left(W_0 x_p + \sum_{\substack{t \\ q \to p}} W_t x_q\right)$$

- x_p is the input feature vector of pixel p
- y_p is the output feature vector of pixel p
- $t \in \{N, NW, W, \ldots\}$ is a cardinal direction
- W_t are learned weight matrices
- σ is an activation function

Schlichtkrull et at, “Modeling Relational Data with Graph Convolutional Networks”, arxiv:1703.06103

What do graphs have to do with tables?

- Nodes: words, Edges: neighboring words



ACME US&A

A Company Making Everything

999 Supreme Industrial Road , Anderson , South Carolina .
12345 .UNITED STATES
Tel: +1-123-234-3456 Fed ID 11-123456
Fax: Tax Exempt 123345-6789
VAT Reg. No.:

Order

Our Order Date	Purchase Order
1/2/20	A12345

Supplier No	Revision
WESUPPLY	1

Seller
We-Supply
NC Suppliers Corporation
P.O. Box 123456
Atlanta
GA 12345-98765
UNITED STATES
Phone: 111-222-3334
Fax:

Delivery Address
ACME US&A
999 Supreme Industrial Road
Anderson
South Carolina
12345
UNITED STATES

Ship Via
Road

Payment Terms
1 Day Net

Terms Of Delivery
Free on board

Delivery Date
3/31/19

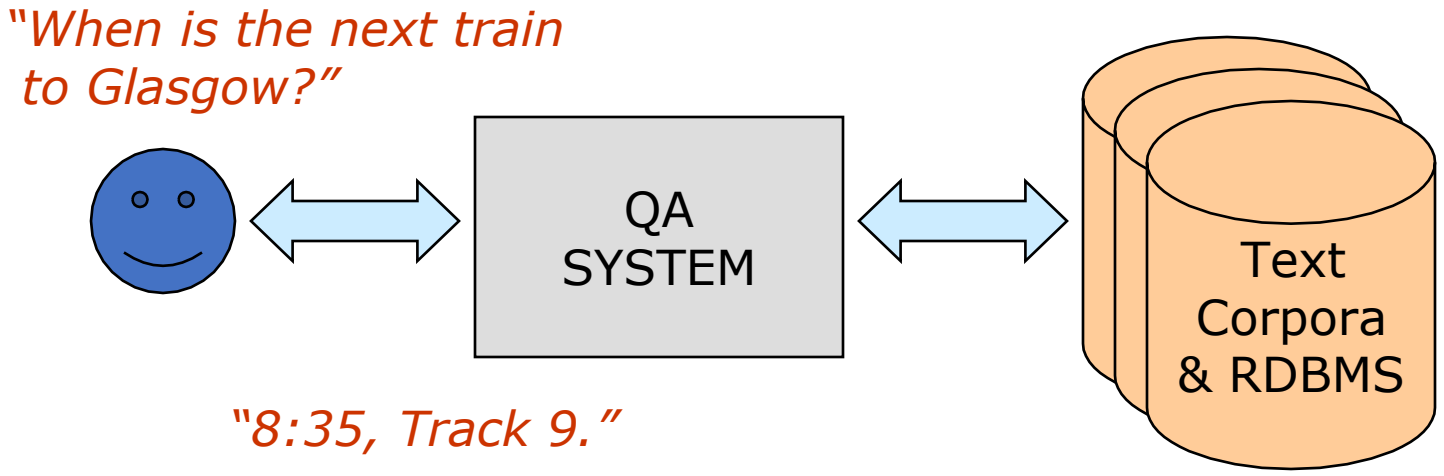
For orders placed by AMCE US&A or AMCE US&B, the AMCE US&A / AMCE US&B Standard Conditions of Contract for the Purchase and / or Hire of Goods and Services (US&A/US&B) apply to this order.
For orders placed by AMCE US&A USA Corporation, the AMCE US&A USA Corporation Standard Conditions of

Question Answering

Question Answering

Question Answering

- **Inputs**: a question in English; a set of text and database resources
- **Output**: a set of possible answers drawn from the resources

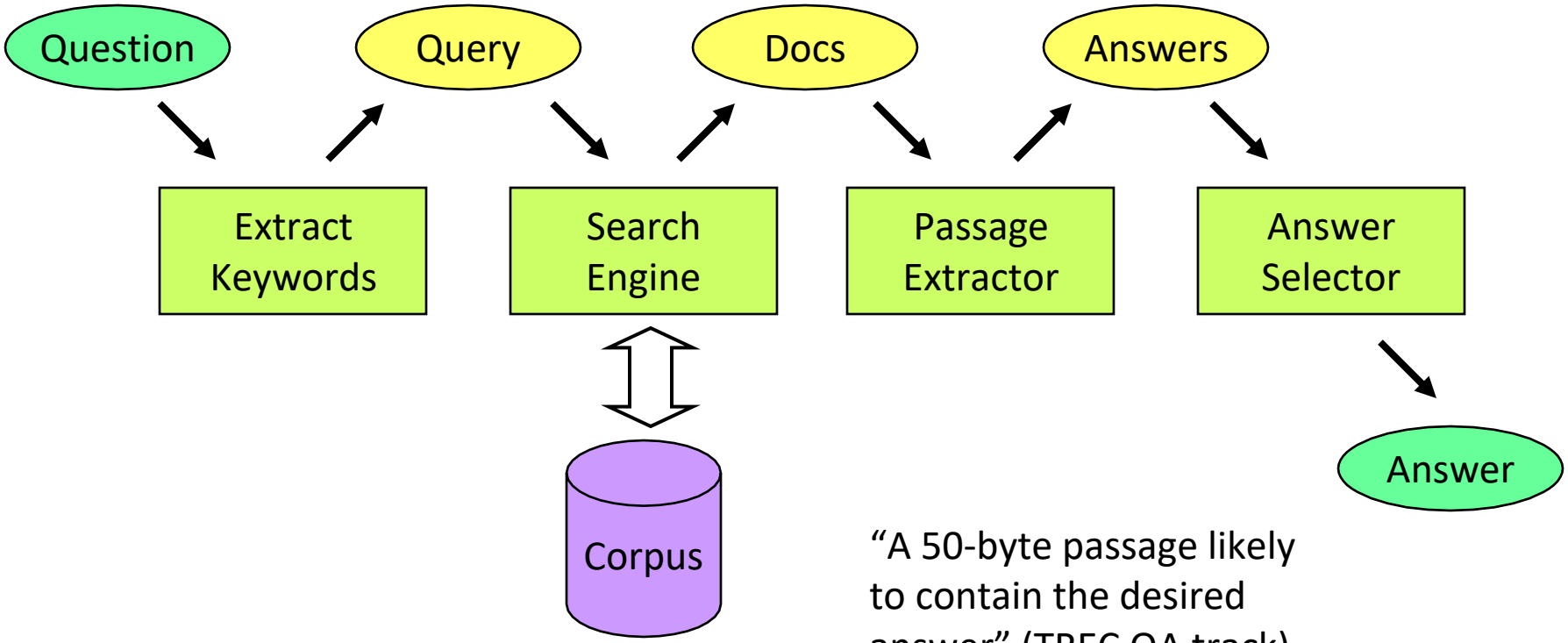


Ancestors of Modern QA

- Information Retrieval
  - Retrieve relevant documents from a set of keywords; search engines
- Information Extraction
  - Template filling from text (e.g. event detection); e.g. TIPSTER, MUC
- Relational QA
  - Translate question to relational DB query; e.g. LUNAR, FRED

Typical TREC QA Pipeline

"A simple factoid question"



"A 50-byte passage likely to contain the desired answer" (TREC QA track)

What's Doable and What's Not?

- Q: What year did the Titanic sink?

  A: 1912

- Q: Why did Titanic sink?

  A: ?

Why did the Titanic sink? The immediate cause of RMS **Titanic's** demise was a collision with an iceberg that caused the ocean liner to **sink** on April 14–15, 1912. While the ship could reportedly stay afloat if as many as 4 of its 16 compartments were breached, the impact had affected at least 5 compartments. Oct 15, 2020

www.britannica.com › ... › Accidents & Disasters

Titanic | History, Sinking, Rescue, Survivors, & Facts | Britannica

who discovered clouds

All Images News V

About 54,600,000 results (0.61 seconds)

Luke Howard

who was the first person to land on the sun

All News Images Videos Shoppi

About 511,000,000 results (0.56 seconds)

Hung Il Gong

when did canada become part of the united states

All News Shopping Maps Images

About 2,940,000,000 results (0.81 seconds)

July 1, 1867

On **July 1, 1867**, with passage of the British North A

what year did tom hanks land on the m

All News Images Vide

About 1,560,000 results (0.73 seconds)

1970

Reading Comprehension

SQuAD2.0
The Stanford Question Answering Dataset

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	**88.592**	**90.859**
2 Jul 19, 2019	XLNet + SG-Net Verifier (ensemble) *Shanghai Jiao Tong University & CloudWalk*	88.050	90.645
3 Jul 23, 2019	XLNet + SG-Net Verifier (single model) *Shanghai Jiao Tong University & CloudWalk*	87.046	89.899
3 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	87.147	89.474
3 Jul 20, 2019	RoBERTa (single model) *Facebook AI*	86.820	89.795
4 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
5 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147
6 May 21, 2019	XLNet (single model) *Google Brain & CMU*	86.346	89.133

https://rajpurkar.github.io/SQuAD-explorer/

HotpotQA

A Dataset for Diverse, Explainable Multi-hop Question Answering

Paragraph A, Return to Olympus:
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
Supporting facts: 1, 2, 4, 6, 7

Entity Relation Network

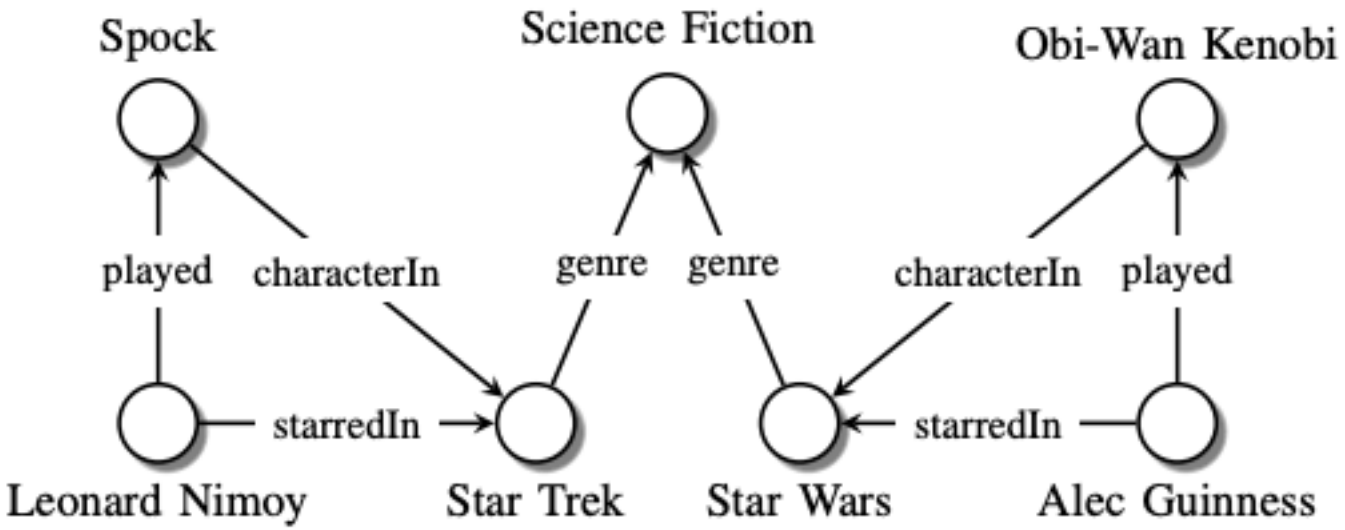
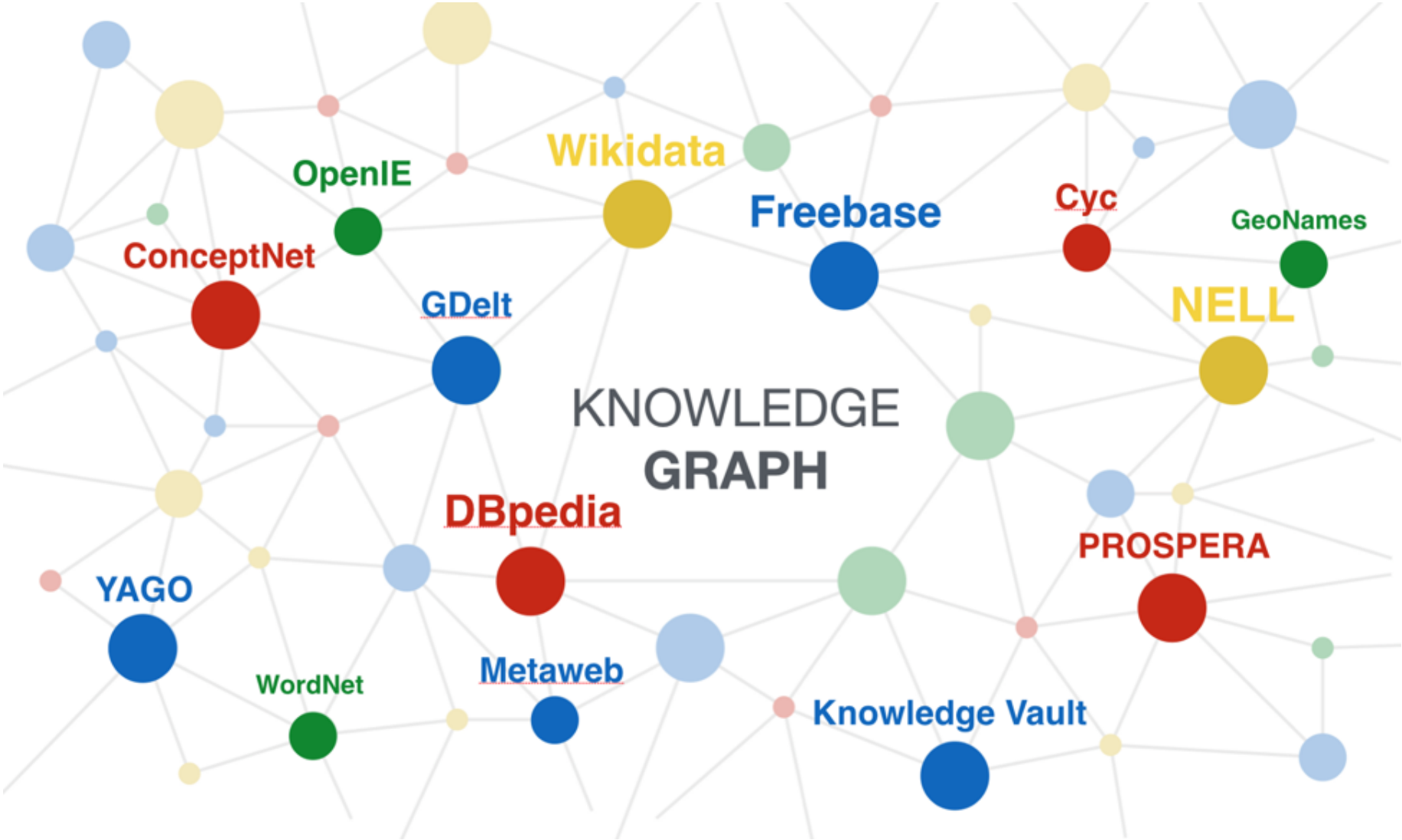


Fig. 1. Sample knowledge graph. Nodes represent entities, edge labels represent types of relations, edges represent existing relationships.

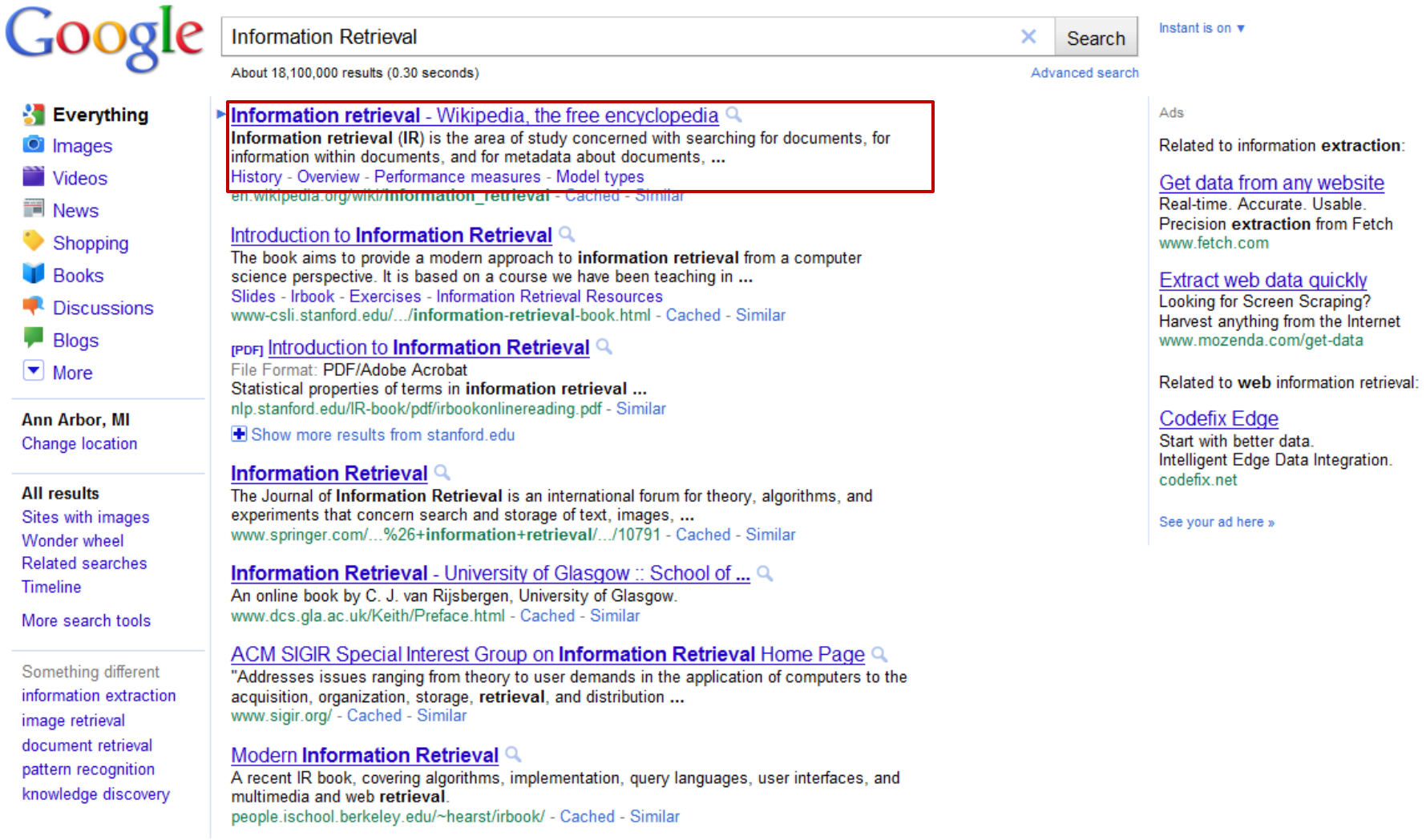
Knowledge Graphs



https://wso2.com/blog/research/probabilistic-error-detection-model-for-knowledge-graph-refinement

Text Summarization

Text Summarization



The Summarization Problem

- Text summarization:
  - creation of a shortened version of a document, or multiple documents which contains the most important contents of the original text.
  - Example: snippets of search results; abstracts of a paper;
- Text summarization as sentence extraction:
  - Pick the most important k sentences in a document/multiple documents
  - There are other criteria (e.g., redundancy, coherence), but in this lecture we only talk about importance

The Summarization Problem (Cont.)

- Essentially “semantic compression” of text
- Extraction-based vs. generation-based summary
- In general, we need a purpose for summarization, but it’s hard to define it

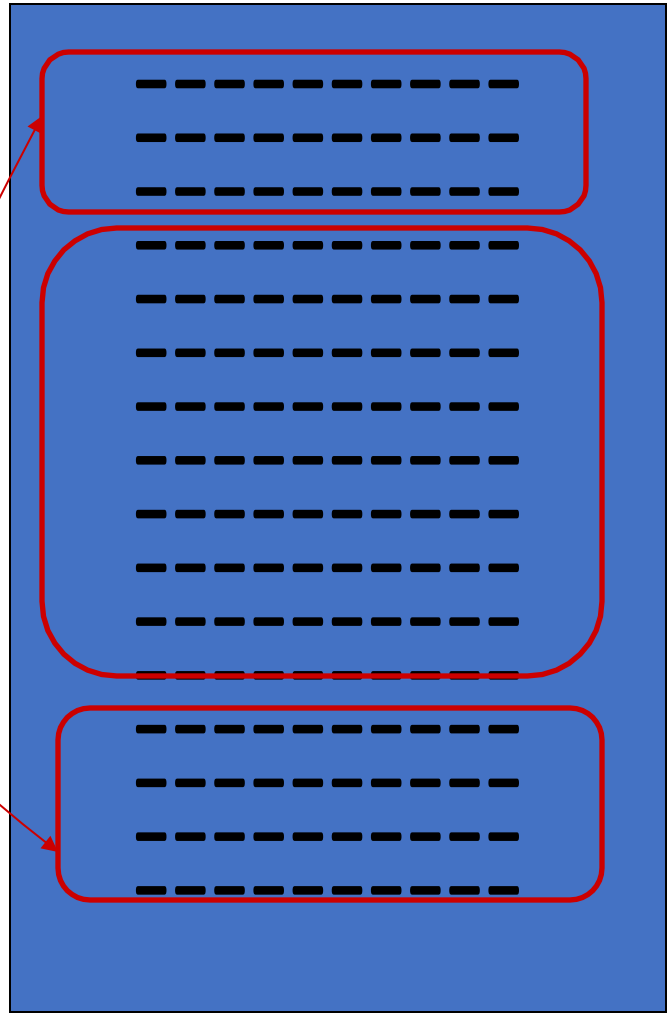
Examples of Summarization

- News summary
- Summarize retrieval results
  - Single document summary
  - Multi-document summary
- Summarize a cluster of documents (automatic label creation for clusters)

“Retrieval-based” Summarization

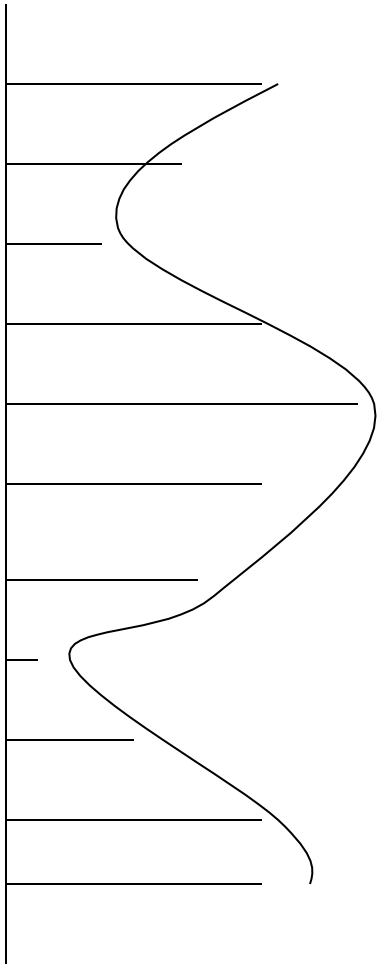
- Basic approach
  - Rank “sentences”, and select top N as a summary
- Observation: term vector ≈ summary?
- Methods for ranking sentences
  - Based on term weights
  - Based on position of sentences
  - Based on the similarity of sentence and document vector

Simple Discourse Analysis

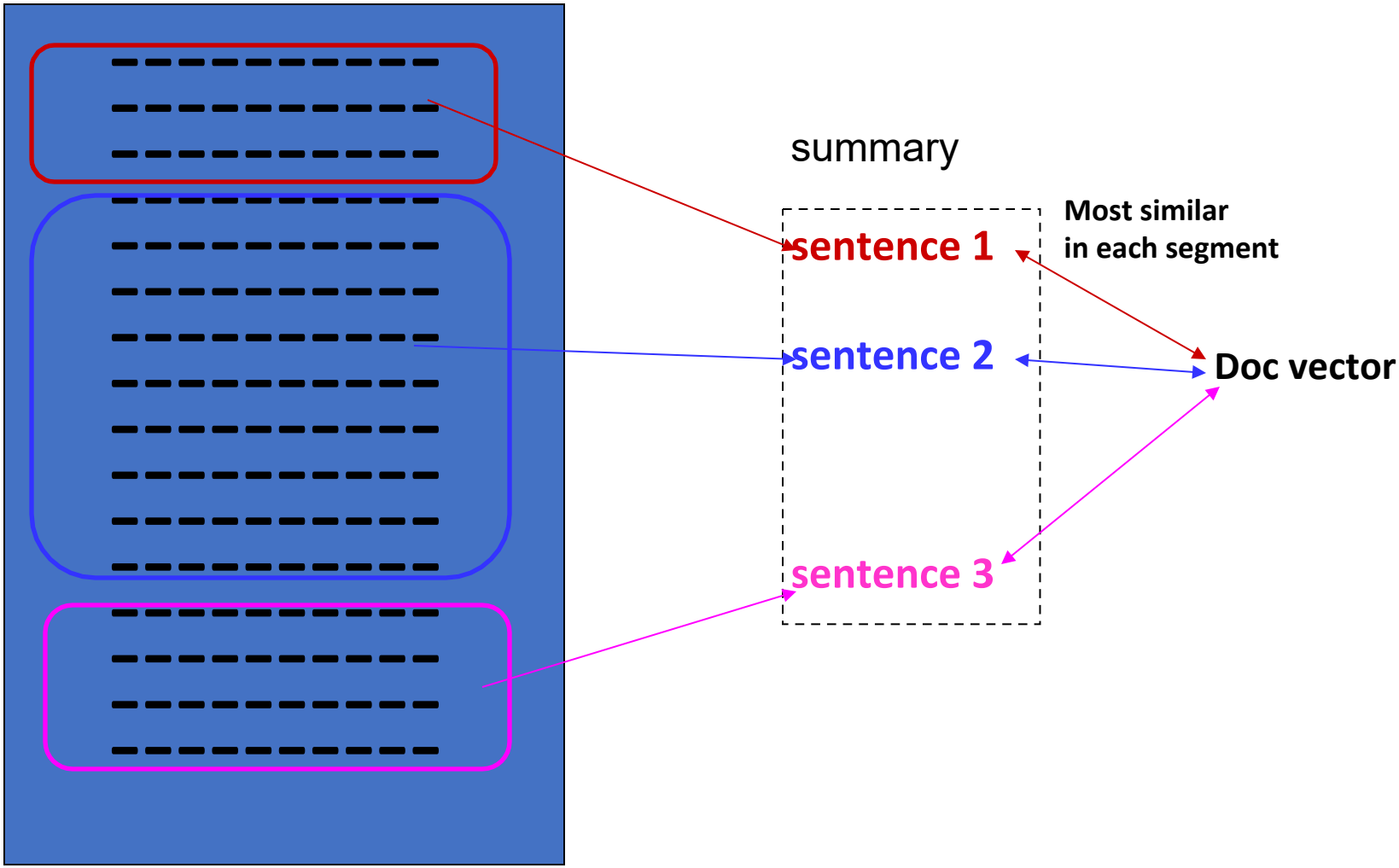


vector 1
vector 2 — similarity
vector 3 — similarity
…
…

vector n-1
vector n — similarity



A Simple Summarization Method



Cosine Similarity Between Sentences

- Let s_1 and s_2 be two sentences.
- Let x and y be their representations in an n-dimensional vector space
- The cosine between is then computed based on the inner product of the two.

$$\cos(x, y) = \frac{\sum_{i=1,n} x_i y_i}{\|x\| \|y\|}$$

- The cosine ranges from 0 to 1.

slide from: Dragomir Radev

LexRank (Cosine Centrality)

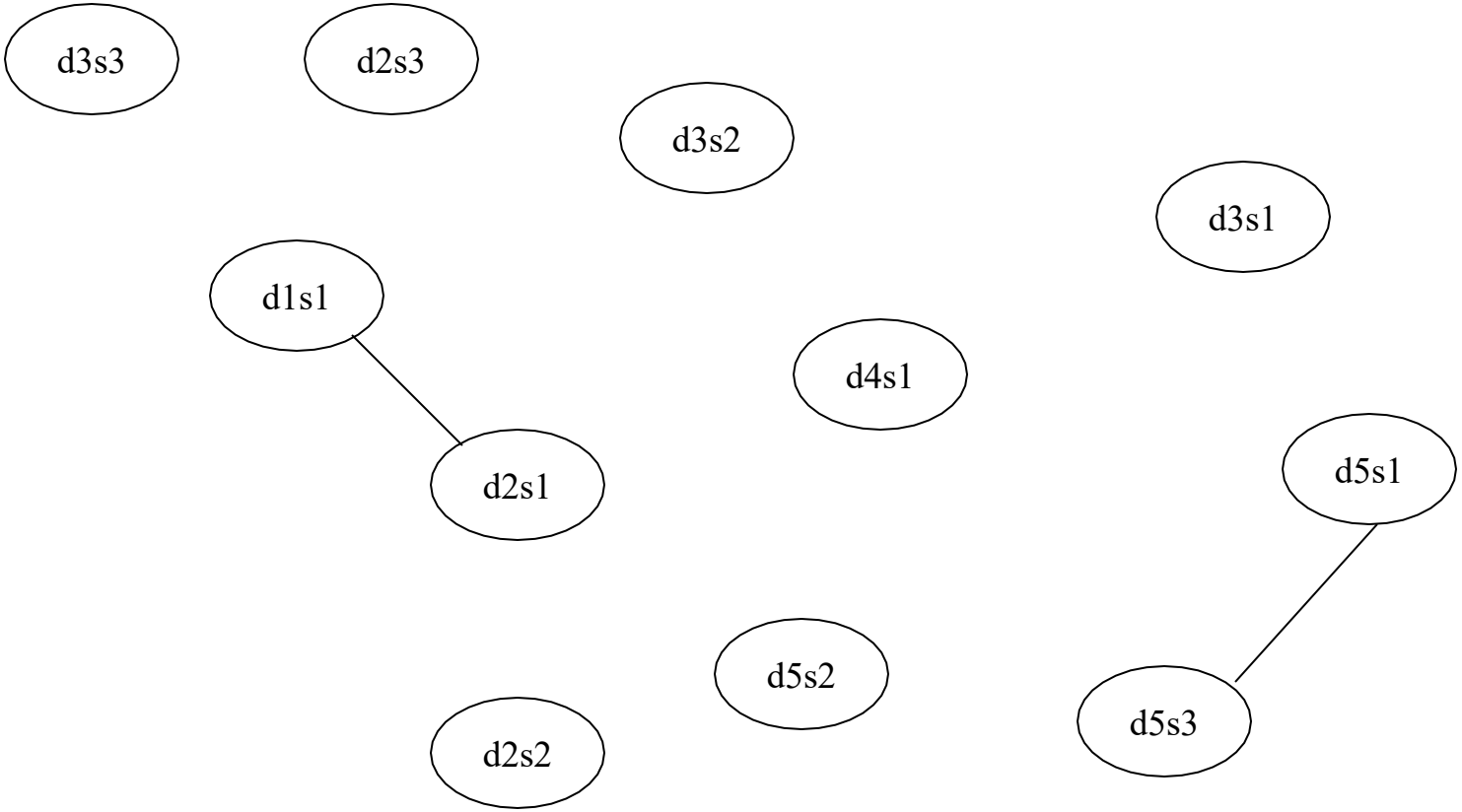
	1	2	3	4	5	6	7	8	9	10	11
1	1.00	0.45	0.02	0.17	0.03	0.22	0.03	0.28	0.06	0.06	0.00
2	0.45	1.00	0.16	0.27	0.03	0.19	0.03	0.21	0.03	0.15	0.00
3	0.02	0.16	1.00	0.03	0.00	0.01	0.03	0.04	0.00	0.01	0.00
4	0.17	0.27	0.03	1.00	0.01	0.16	0.28	0.17	0.00	0.09	0.01
5	0.03	0.03	0.00	0.01	1.00	0.29	0.05	0.15	0.20	0.04	0.18
6	0.22	0.19	0.01	0.16	0.29	1.00	0.05	0.29	0.04	0.20	0.03
7	0.03	0.03	0.03	0.28	0.05	0.05	1.00	0.06	0.00	0.00	0.01
8	0.28	0.21	0.04	0.17	0.15	0.29	0.06	1.00	0.25	0.20	0.17
9	0.06	0.03	0.00	0.00	0.20	0.04	0.00	0.25	1.00	0.26	0.38
10	0.06	0.15	0.01	0.09	0.04	0.20	0.00	0.20	0.26	1.00	0.12
11	0.00	0.00	0.00	0.01	0.18	0.03	0.01	0.17	0.38	0.12	1.00

slide from: Dragomir Radev

Building the Sentence Network

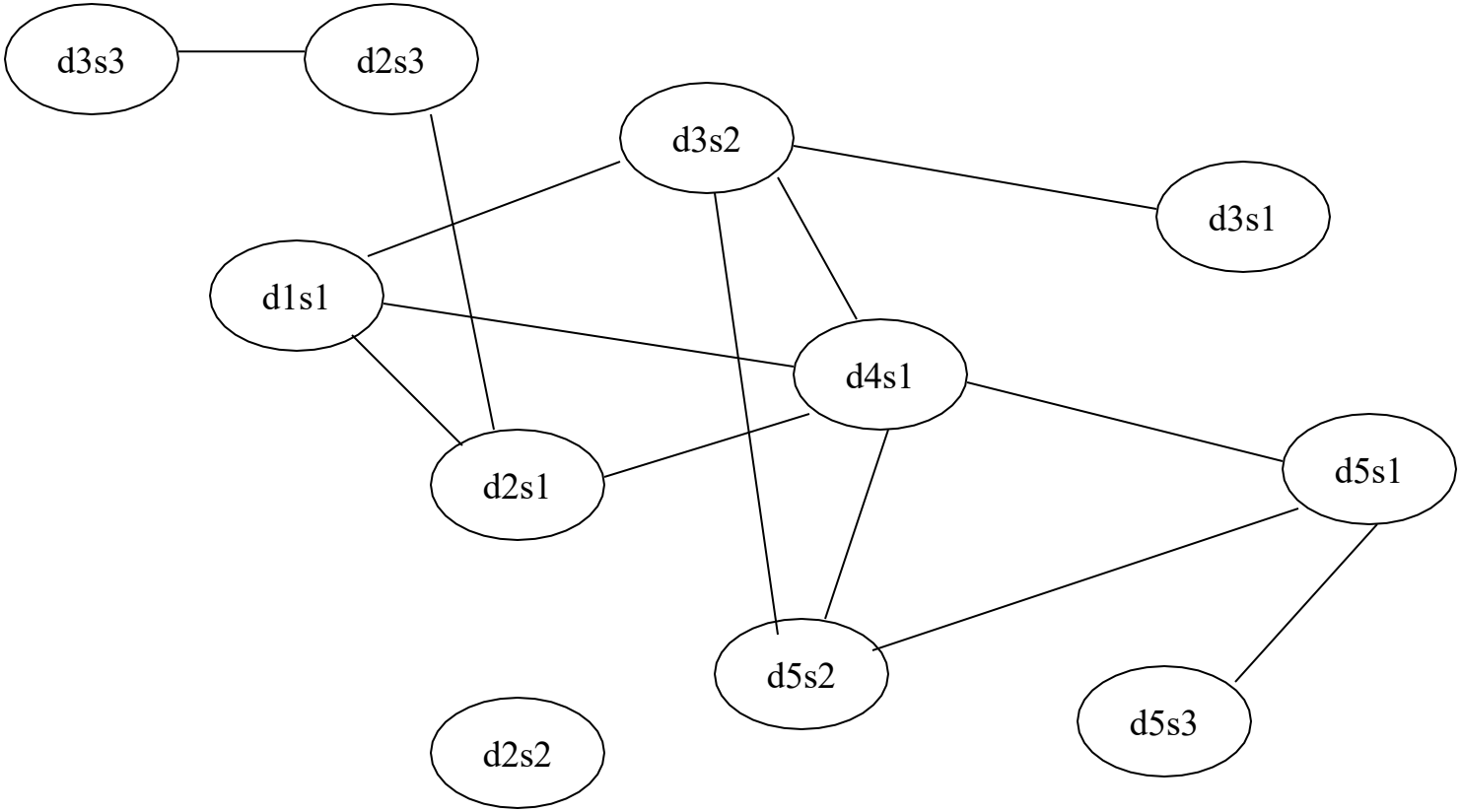
- Every vertex is a sentence
- Add an undirected edge between two vertices (sentences) if the cosine similarity of the two sentences is above a threshold t.

Lexical Centrality (t=0.3)



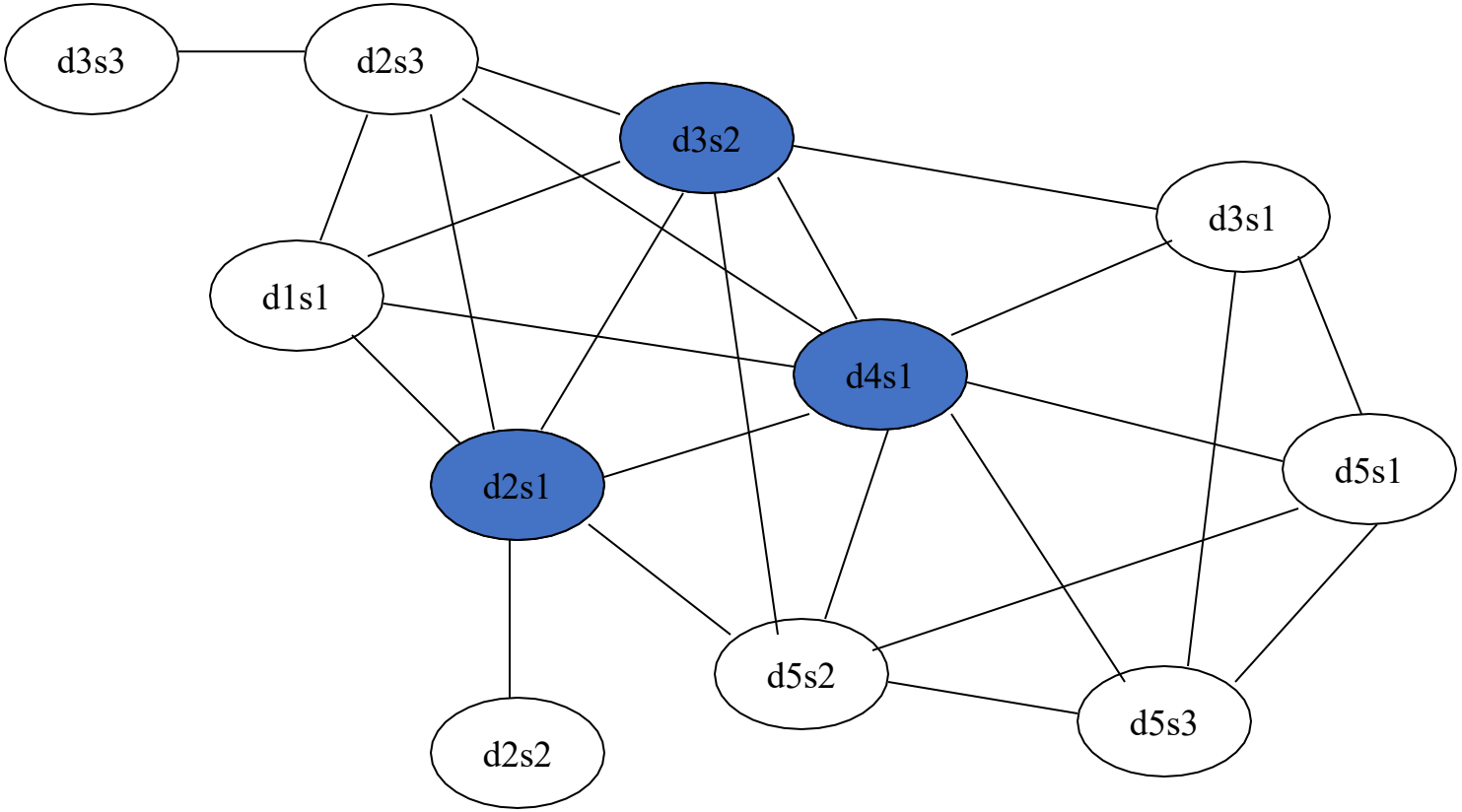
slide from: Dragomir Radev

Lexical Centrality (t=0.2)



slide from: Dragomir Radev

Lexical Centrality (t=0.1)



Sentences vote for the most central sentence…

slide from: Dragomir Radev

LexRank

$$p(Ti) = \frac{d}{c(T_1)} p(T_1)E(T_1, Ti) + \ldots + \frac{d}{c(T_n)} p(T_n)E(T_n, Ti) + \frac{1-d}{N}$$

- $T_1 \ldots T_n$ are pages that link to A, $c(T_i)$ is the outdegree of page T_i, and N is the total number of pages.
- d is the "damping factor", or the probability that we "jump" to a far-away node during the random walk. It accounts for disconnected components or periodic graphs.
- When $d = 0$, we have a strict uniform distribution.
  When $d = 1$, the method is not guaranteed to converge to a unique solution.
- Typical value for d is between [0.1,0.2] (Brin and Page, 1998).

This is essentially computing PageRank on the sentence network

slide from: Dragomir Radev

Generative Text Summarization

- Document encoding + Text generation
- GAN is commonly used
- Related to other types of text generation:
  - Review generation
  - Title generation
  - Machine translation
  - Text simplification
  - …

Sentiment and Opinion Analysis

Sentiment and Opinion Analysis

- Computational study of subjectivity in text
  - opinions, sentiments, subjectivity, evaluations, attitudes, appraisal, affects, views, emotions, etc.
  - Reviews, blogs, tweets, discussions, news, comments, feedback, or any other documents
- Terminology:
  - Sentiment analysis is more widely used in industry.
  - Both are widely used in academia
  - Sometimes used interchangeably.

- Slide partially from Prof. Bing Liu's tutorial

Why Bother with this in IR?

- “Opinions” are key influencers of our behaviors.
- Our beliefs and perceptions of reality are conditioned on how others see the world.
- Whenever we need to make a decision, we often seek out the opinions of others. In the past,
  - Individuals: seek opinions from friends and family
  - Organizations: use surveys, focus groups, opinion polls, consultants.

- Slide partially from Prof. Bing Liu’s tutorial

Applications of Sentiment Analysis

- Businesses and organizations
  - Benchmark products and services; market intelligence.
  - Businesses spend a huge amount of money to find consumer opinions using consultants, surveys and focus groups, etc.
- Individuals
  - Make decisions to buy products or to use services
  - Find public opinions about political candidates and issues
- Search engines
  - Provide opinion polarized summary of products
- Social science, political science, health informatics, …
  - Understanding public opinions, concerns, …
- Finance, …

Different Levels of Sentiment Analysis

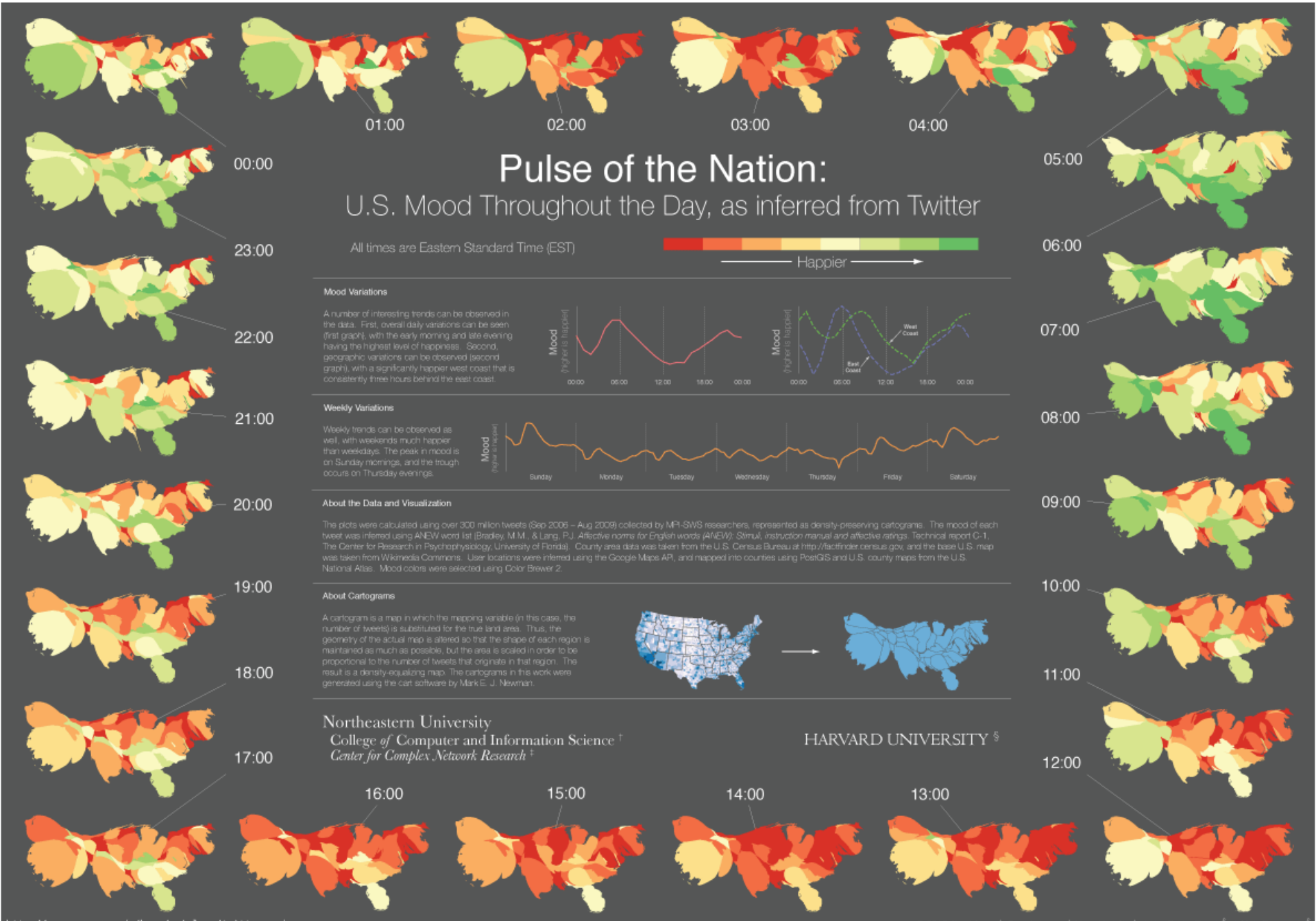
- Sentiment classification
  - Whether a document/sentence conveys a sentiment
- Sentiment polarities
  - Positive, negative, neutral sentiments
- Sentiment in real scales
- Mood, emotion
  - Multi-class classification
- Attitude
  - Sentiment towards particular subjects
- Topic-sentiment mixture
  - Latent aspects of sentiments/opinions
- …

Sentiment Classification

- Yet another text classification problem
- But much more challenging than topic-based classification!
- In general, it touches almost every aspect of NLP
- Human agreement is around 70% or lower
- For human agreed examples, performance is usually between 80% to 90%.
- Evaluation needs to be done in real applications

Features Matter

- Keyword features
  - Most useful features in topic classification
  - Presence vs. frequency
  - Dictionary match performs reasonably well.
- Part-of-speech
  - Very effective in sentiment classification
- Syntax vs. word proximity
- Negation tends to be important
- Topic-related features
- Specific features – e.g., emoticons
- Coreference, sarcasm, metaphor, …



Pulse of the Nation:

U.S. Mood Throughout the Day, as inferred from Twitter

All times are Eastern Standard Time (EST)

Happier

Mood Variations

A number of interesting trends can be observed in the data. First, overall daily variations can be seen (first graph), with the early morning and late evening having the highest level of happiness. Second, geographic variations can be observed (second graph), with a significantly happier west coast that is consistently three hours behind the east coast.

Weekly Variations

Weekly trends can be observed as well, with weekends much happier than weekdays. The peak in mood is on Sunday mornings, and the trough occurs on Thursday evenings.

About the Data and Visualization

The plots were calculated using over 300 million tweets (Sep 2006 – Aug 2009) collected by MPI-SWS researchers, represented as density-preserving cartograms. The mood of each tweet was inferred using ANEW word list (Bradley, M.M., & Lang, P.J. *Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings*. Technical report C-1, The Center for Research in Psychophysiology, University of Florida). County area data was taken from the U.S. Census Bureau at http://factfinder.census.gov, and the base U.S. map was taken from Wikimedia Commons. User locations were inferred using the Google Maps API, and mapped into counties using PostGIS and U.S. county maps from the U.S. National Atlas. Mood colors were selected using Color Brewer 2.

About Cartograms

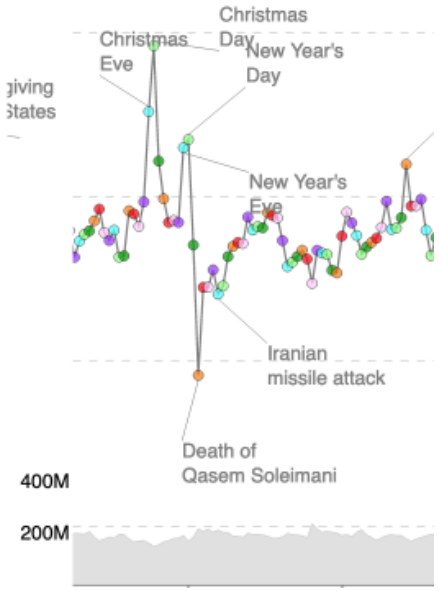
A cartogram is a map in which the mapping variable (in this case, the number of tweets) is substituted for the true land area. Thus, the geometry of the actual map is altered so that the shape of each region is maintained as much as possible, but the area is scaled in order to be proportional to the number of tweets that originate in that region. The result is a density-equalizing map. The cartograms in this work were generated using the cart software by Mark E. J. Newman.

Northeastern University
College *of* Computer and Information Science †
Center for Complex Network Research ‡

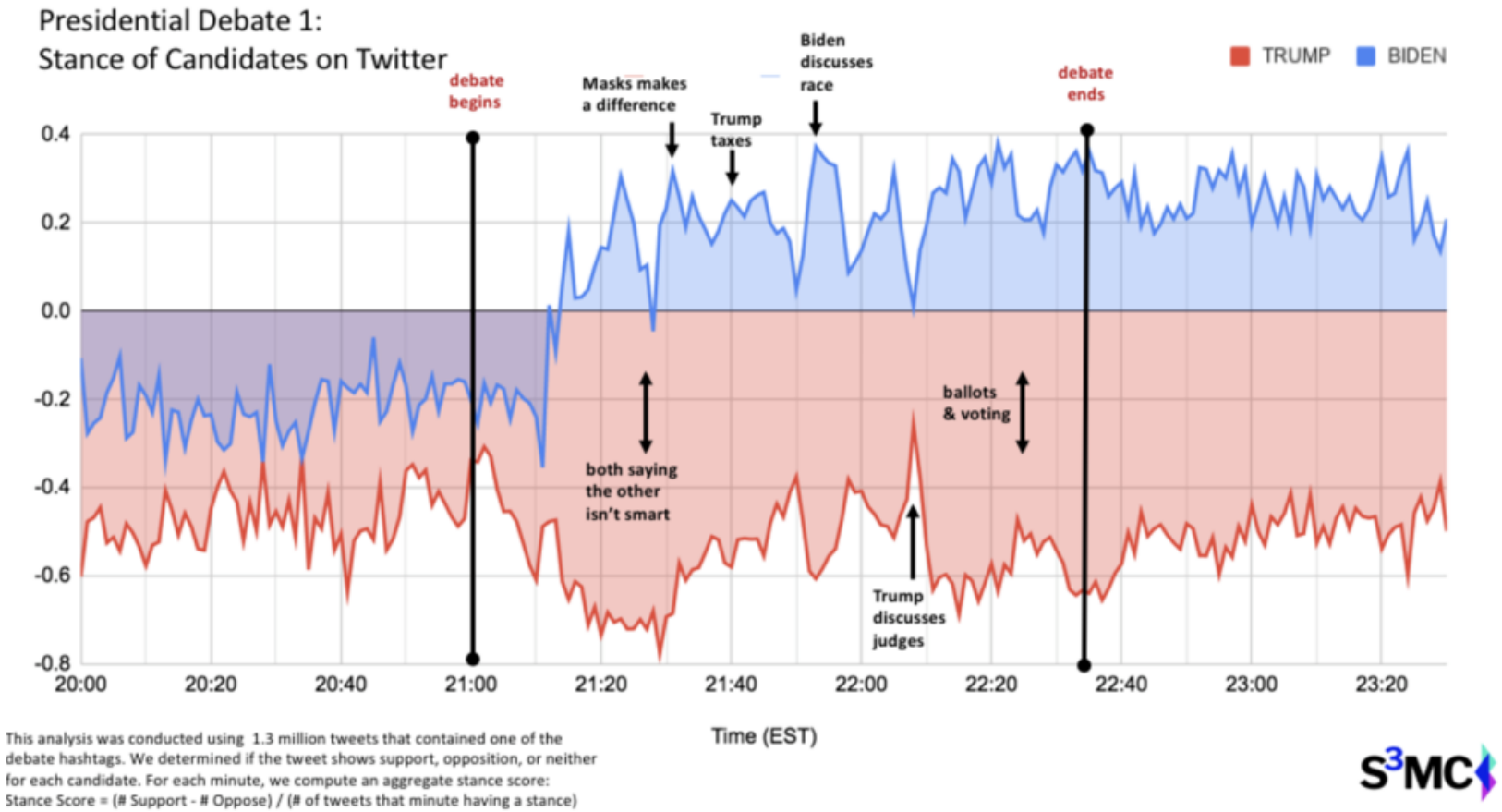
HARVARD UNIVERSITY §

http://www.ccs.neu.edu/home/amislove/twittermood

http://hedonometer.org



Stance Detection



This analysis was conducted using 1.3 million tweets that contained one of the debate hashtags. We determined if the tweet shows support, opposition, or neither for each candidate. For each minute, we compute an aggregate stance score:
Stance Score = (# Support - # Oppose) / (# of tweets that minute having a stance)

S3MC

Budak, Kawintiranon, Singh, Soroka, 2020

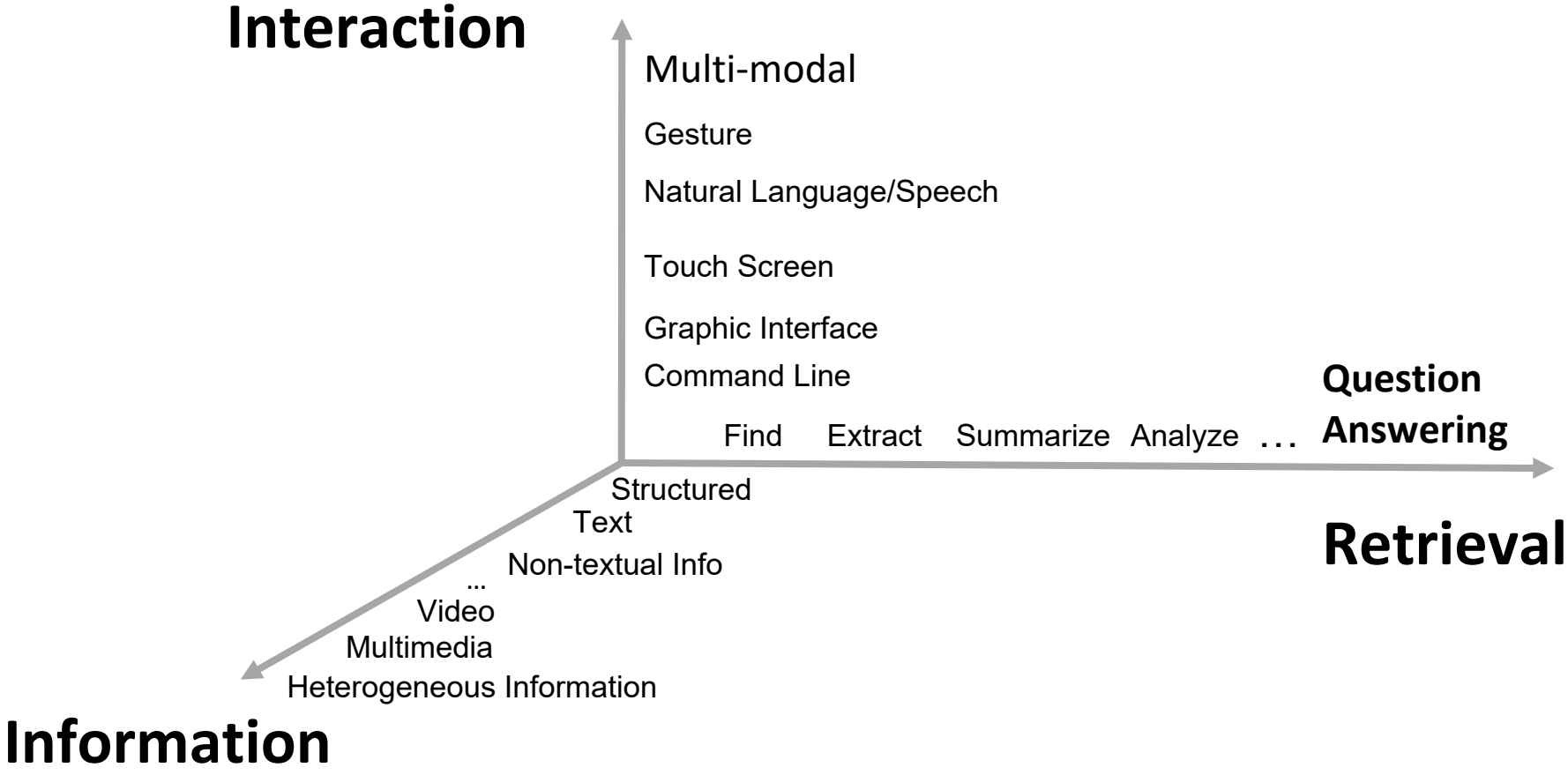
Conversation and IR

Conversation and IR are interrelated

Use conversation to refine/perform search tasks

Use search to aid in dialog

Broad Interpretation of Interactive Information Retrieval

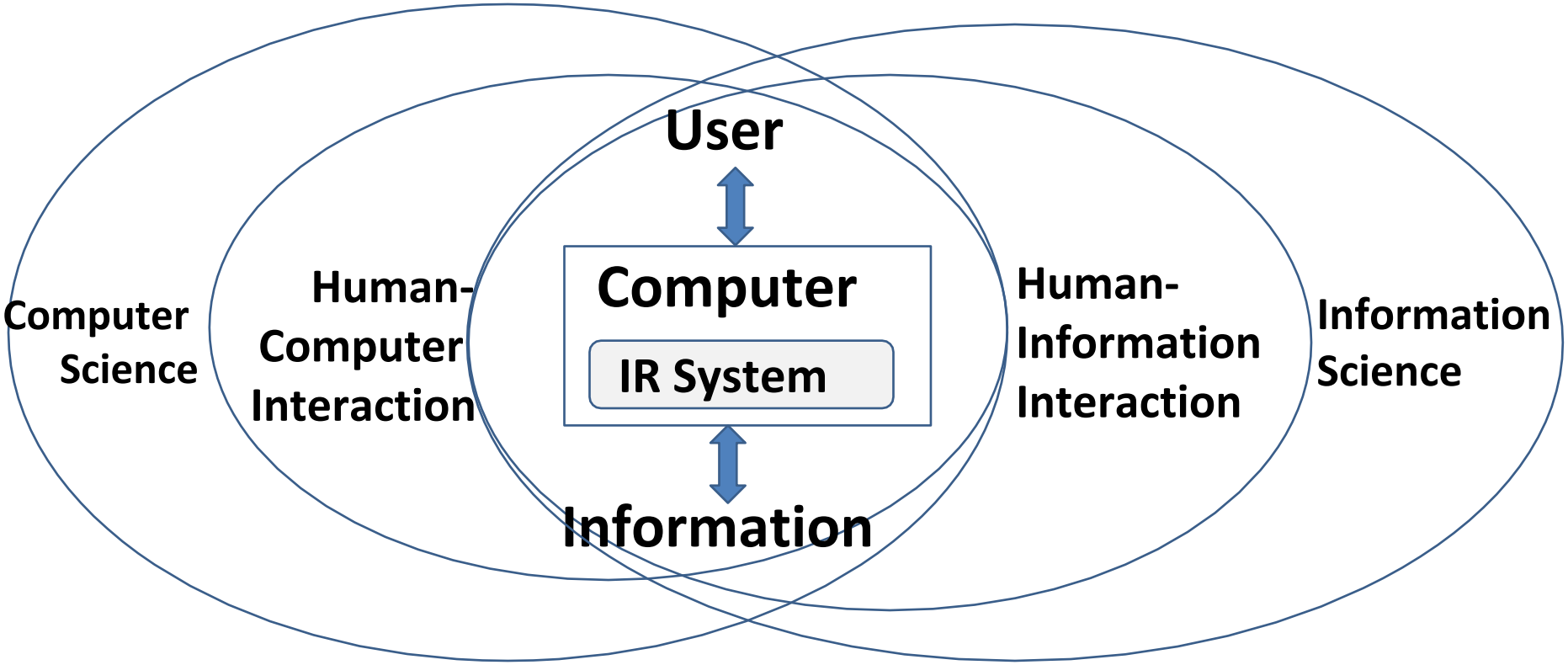


Multiple Perspectives of IIR

- Cognitive IR framework (broadest): interactions can be between a person and a system as well as between people:

  "... the interactive communication processes that occur during the retrieval of information by involving all the major participants in IR, i.e. the user, the intermediary, and the IR system." Ingwersen, 1992
- HCI view: interactions can only be between a person and a system, but the system can go beyond supporting only retrieval to support task and interface can be complex
- Search engine application view: interactions are restricted to a search engine interface (iterative query reformulation, browsing, clicking, ….)

IIR as subarea of Computer Science (HCI) and Information Science (HII)



Interactive IR = Cooperative Game-Playing

- Retrieval process = cooperative game-playing
- Players: Player 1= search engine; Player 2= user
- Rules of game:
  - Player take turns to make “moves”
  - First move = “user entering the query” (in search) or “system recommending information” (in recommendation)
  - User makes the last move (usually)
  - For each move of the user, the system makes a response move (shows an interaction interface), and vice versa
- Objective: help the user complete the (information seeking) task with minimum effort & minimum operating cost for search engine

Unification of search and recommendation

Major benefits of IR as game playing

- **General**
  - A formal framework to integrate research in user studies, evaluation, retrieval models, and efficient implementation of IR systems
  - A general tool for conceptualizing important research topics in Interactive IR
- **Specific**
  - Naturally optimize performance on an entire session instead of that on a single query (optimizing the chance of winning the entire game)
  - Optimize the collaboration of machines and users (maximizing collective intelligence)
    - May see this as "human in the loop AI"
  - Emphasize the two-way communications between a user and a system (e.g., active feedback)

IR-based Conversational Agents

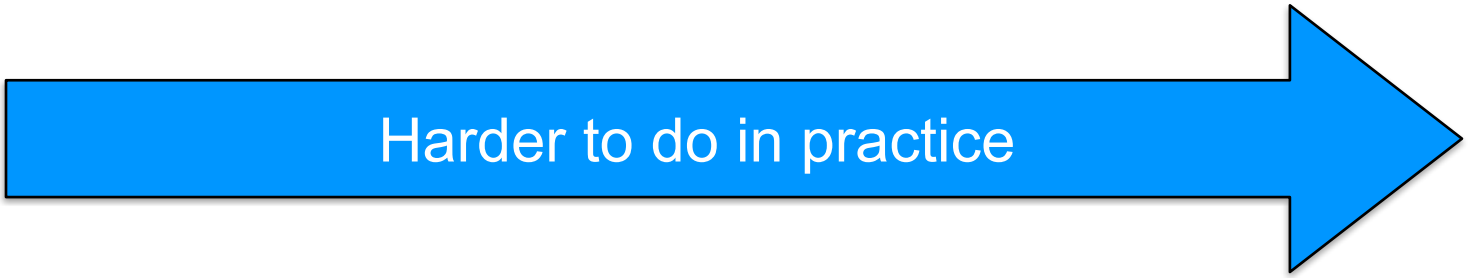
- Customer service settings have rich stored dialogs between representatives and customers
- Re-imaging dialog as IR:
  - Query = user utterance
  - Documents: past customers' utterances
  - Retrieves: representative responses to past customer
- Avoids generating texts from neural methods—a human wrote each reply originally!
  - Prioritized in customer-facing settings

Using IR chatbots in Practice

- Personalized assistance
- Follow-up questions

- Providing Information
- Triaging new inquiries
- Answer common questions

- Interactivity / Negotiation
- Question Revisions



Can Machine Learning And NLP Help Predict Suicidal Risk?

Yoni Levine Follow
Sep 11, 2017 · 7 min read

https://medium.com/@yoni.levine/from-depression-to-suicide-how-the-way-we-speak-can-predict-the-way-we-feel-ed359e54c81

From Depression To Suicide, How The Words We Use Can Predict The Way We Feel.

Unfortunately, the number of people suffering from depression is on the rise. According to this study by the NIH, nearly 7 percent of US adults, and over 10 percent of those between the ages of 18–25, have had a major depressive episode in the last year. The CDC published another surprising and frightening statistic: less than one-third of Americans taking one antidepressant medication, and less than one-half of those taking multiple antidepressants have seen a mental health professional in the past year.

Google AI Blog

The latest news from Google AI

Deep Learning for Electronic Health Records

Tuesday, May 8, 2018

Posted by Alvin Rajkomar MD, Research Scientist and Eyal Oren PhD, Product Manager, Google AI

When patients get admitted to a hospital, they have many questions about what will happen next. When will I be able to go home? Will I get better? Will I have to come back to the hospital? Having precise answers to those questions helps doctors and nurses make care better, safer, and faster — if a patient's health is deteriorating, doctors could be sent proactively to act before things get worse.

Predicting what will happen next is a natural application of machine learning. We wondered if the same types of machine learning that predict traffic during your commute or the next word in a translation from English to Spanish could be used for clinical predictions. For predictions to be useful in practice they should be, at least:

1. **Scalable**: Predictions should be straightforward to create for any important outcome and for different hospital systems. Since healthcare data is very complicated and requires much data wrangling, this requirement is not straightforward to satisfy.
2. **Accurate**: Predictions should alert clinicians to problems but not distract them with false alarms. With the widespread adoption of electronic health records, we set out to use that data to create more accurate prediction models.

Together with colleagues at UC San Francisco, Stanford Medicine, and The University of Chicago Medicine, we published "Scalable and Accurate Deep Learning with Electronic Health Records" in Nature Partner Journals: Digital Medicine, which contributes to these two aims.

Digital Medicine

https://ai.googleblog.com/2018/05/deep-learning-for-electronic-health.html

The Promise and Peril of AI Legal Services to Equalize Justice

By Ashwin Telang - Edited by Edwin Farley, Teodora Groza, Pablo A. Lozano, and Pantho Sayed

March 14, 2023

What you should know

- RAGs: why are they needed?
- How does Naïve Rag work? How do advanced and modular RAG systems differ from the Naïve RAG approach?
- How to evaluate RAG systems.
- IR has connections to many fields—when you have an information need, anything can be IR
- Strong interconnection between IR and NLP for presenting information
  - Summarization, Information Extraction, Sentiment Analysis, Opinion Mining, …