Microsoft Research

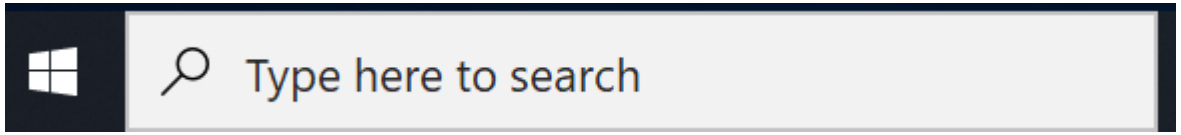
Summit 2021

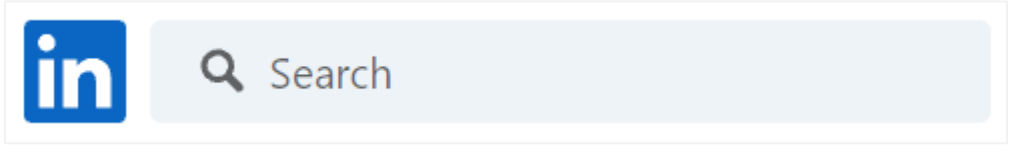
Universal Search & Recommendation

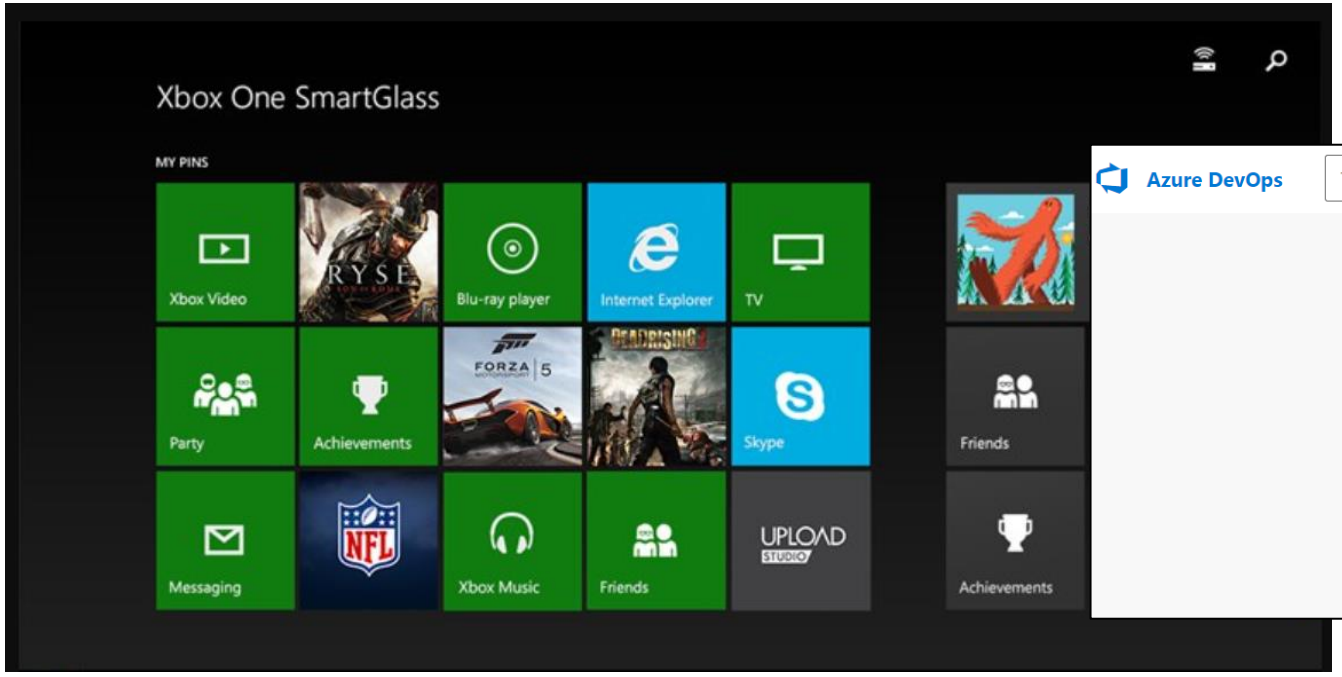
Paul Bennett

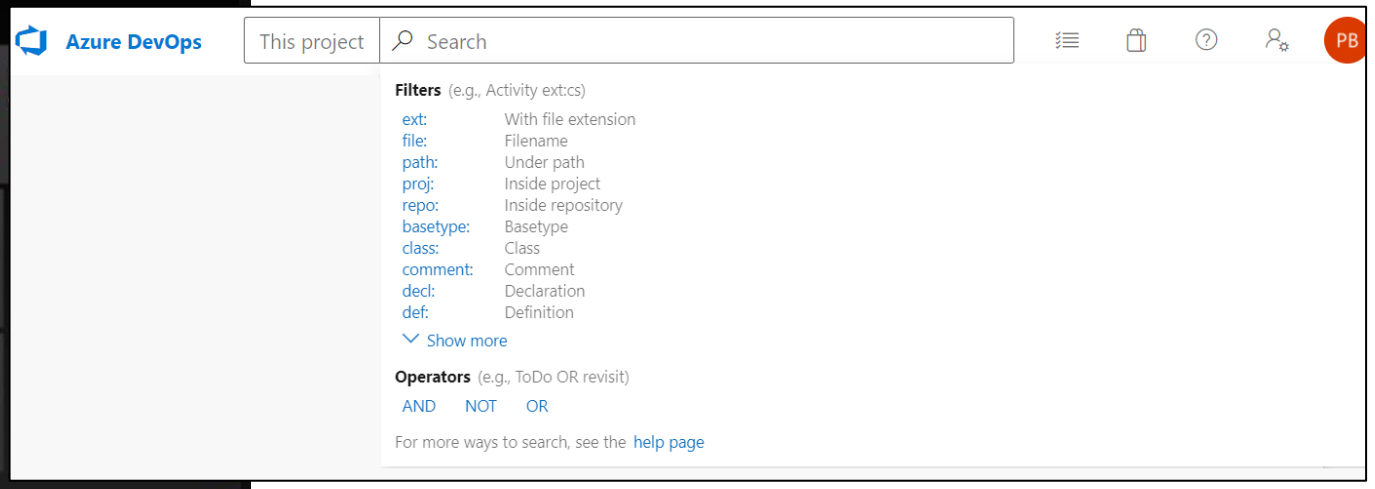
When I say “search” what is it you think of searching?

Search is Everywhere in Microsoft

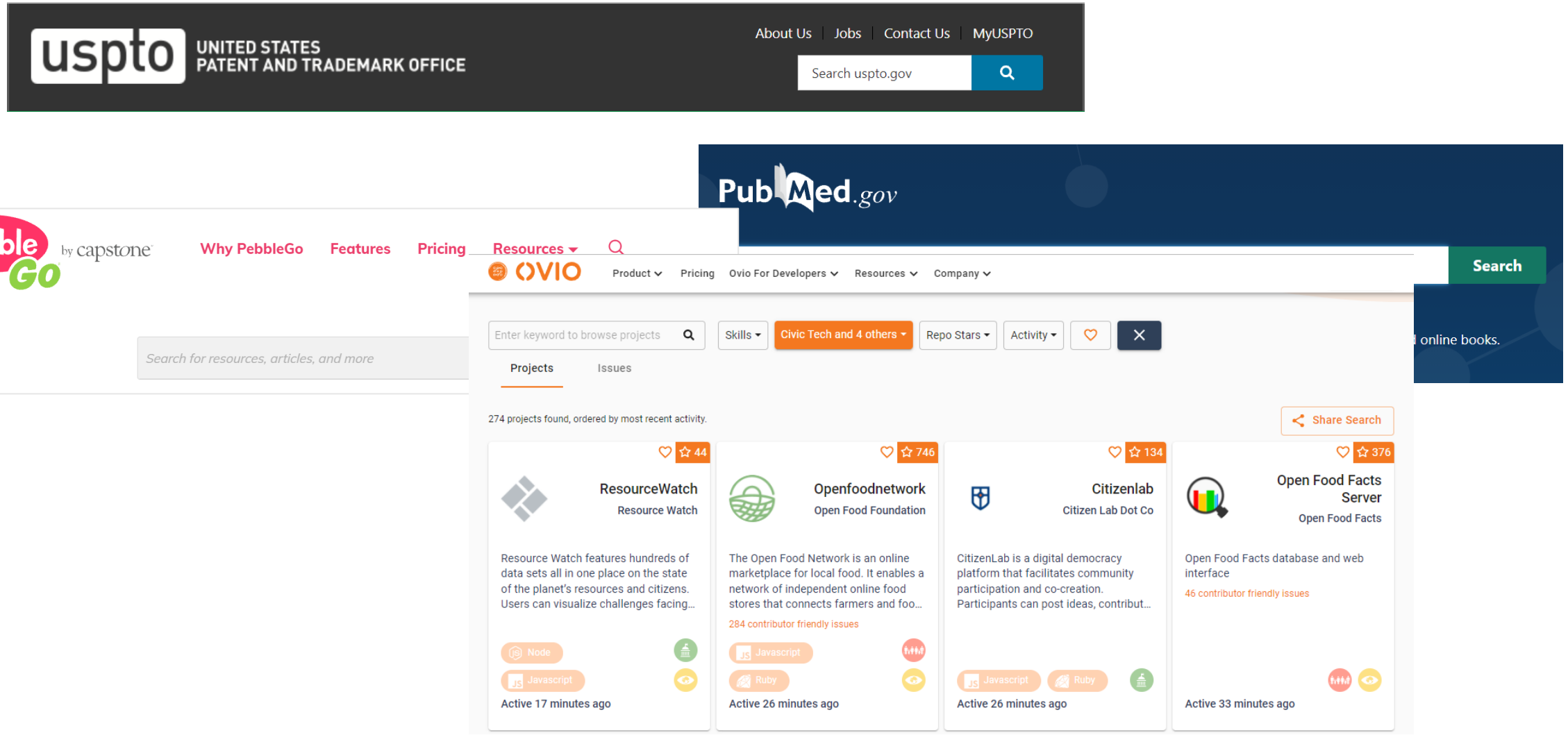








And in the World



Leaving My Fingerprints: Motivations and Challenges of Contributing to OSS for Social Good. Huang et al. ICSE 2021.

Why search is not a “solved problem”

Search engines perform best where search logs are available to learn from implicit feedback.

Many key scenarios lack large search logs.

We need to be able to truly understand content, queries, and people.

Universal Search and Recommendation


Office


Azure


Health


Dynamics


LinkedIn


Xbox


Bing


GitHub

High-performing search should not be limited to the open web.

Provide push-button customized search and recommendation

…. for any person,

any vertical,

any data.

Trends: Toward Universal Search and Recommendation

Dense retrieval based on representation learning as the foundation

- Efficiency and accuracy advances enable dense retrieval to outperform inverted index.

Universal search and recommendation to scenarios without search logs

- Robust weakly & self-supervised learning generalize to the tail.

Universal search and recommendation for all data

- Representation learning enables dense representations to plug into dense retrieval stack.

Universal hyper-personalization

- Learning from feedback from all scenarios provides best search and recommendation experience for each user everywhere.

Current Search Landscape

Web Search
and Limited Scenarios

Bag-of-Words	**1970s-1990s**
Classic IR Features	**1990s-2000s**
Learning to Rank	**2000s-2010s**
Neural IR	**2010s-2020s**
Dense Retrieval	**2020s-Future**

Current Search Landscape

Web Search and Limited Scenarios

Bag-of-Words	1970s-1990s
Classic IR Features	1990s-2000s
Learning to Rank	2000s-2010s
Neural IR	2010s-2020s
Dense Retrieval	2020s-Future

Search Most Everywhere Else

Bag-of-Words	1970s-1990s

"BM25" [1970s/1980s]

Current Search Landscape

Web Search and Limited Scenarios

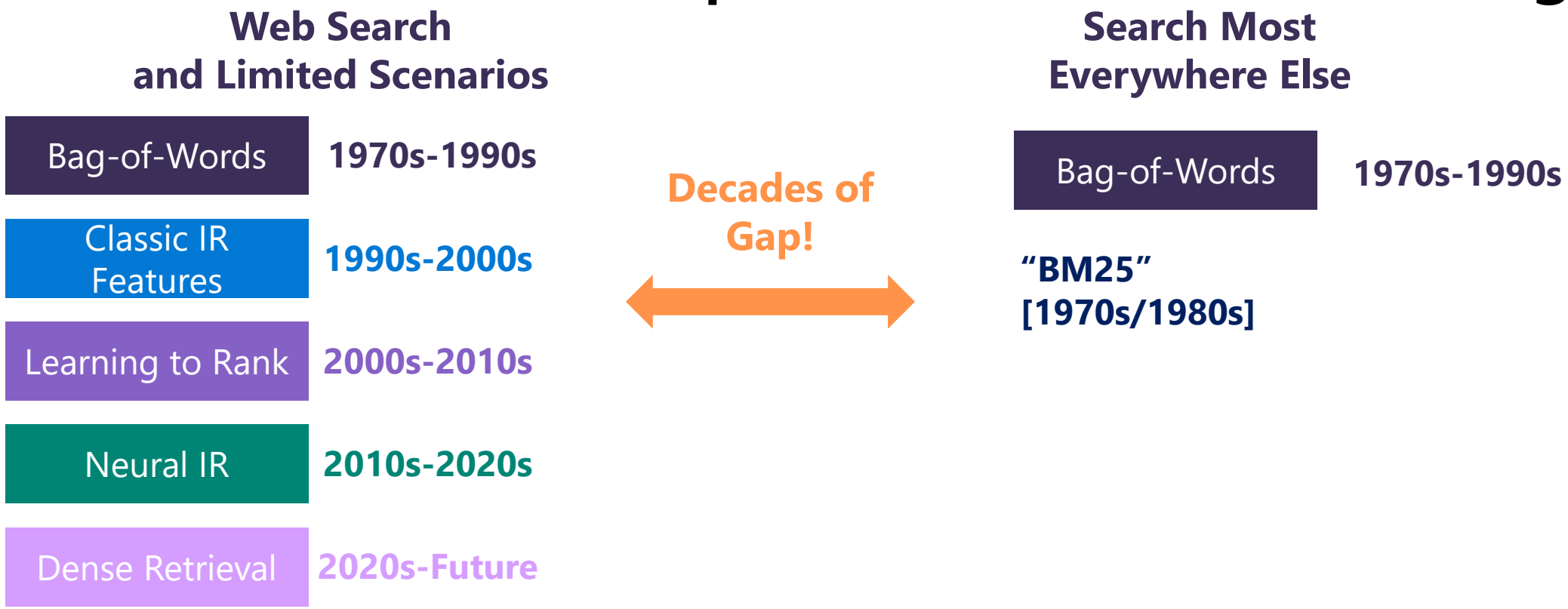
Bag-of-Words	1970s-1990s
Classic IR Features	1990s-2000s
Learning to Rank	2000s-2010s
Neural IR	2010s-2020s
Dense Retrieval	2020s-Future

Decades of Gap!

Search Most Everywhere Else

Bag-of-Words	1970s-1990s
"BM25" [1970s/1980s]	

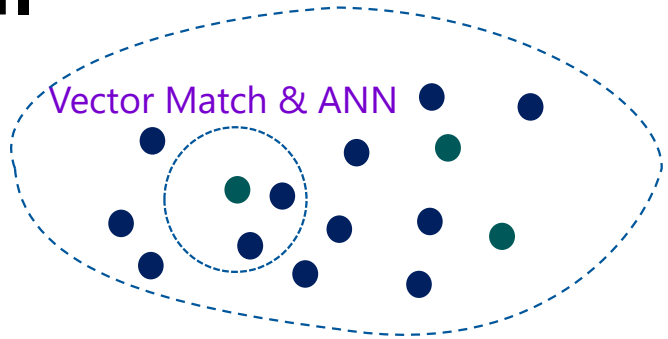
Current Search Landscape and Recent Breakthroughs



Recent breakthroughs in dense retrieval provide a promising path to *close the gap in a scalable way*:

1. **Efficient serving** with ANN retrieval
2. **Short texts match** with Dense retrieval: title/query matching in ads and common documents.
3. **Full document body vectorization** with Dense retrieval.

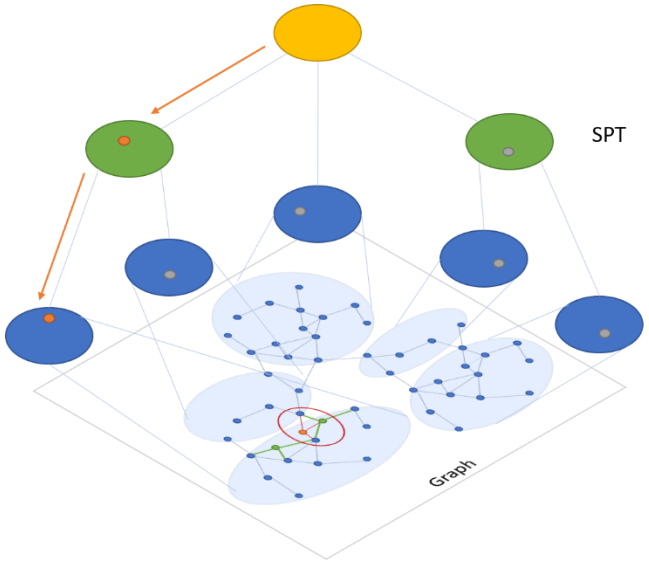
Breakthrough #1: Efficient ANN Retrieval



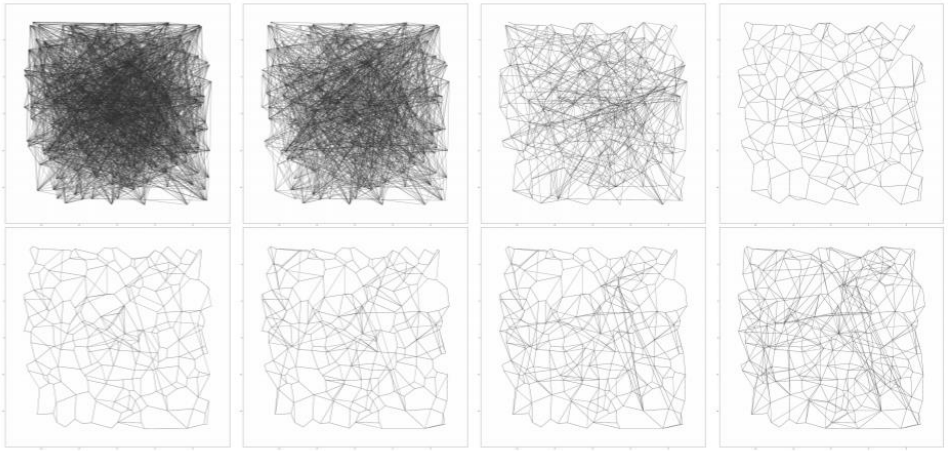
Large scale, efficient serving of dense retrieval with ANN index

- Approximation of KNN with sub linear efficiency
- Often by enforcing sparsity in the searching graph
- Bing has industry leading ANN infrastructure with SPTAG [1] and DiskANN [2]
- Serving multiple scenarios in Bing and expanding

SPTAG



DiskANN



[1] SPTAG: A library for fast approximate nearest neighbor search. Chen et al. 2019.
[2] DiskANN: Fast accurate billion-point nearest neighbor search on a single node. Subramanya et al. NeurIPS 2019.

Breakthrough #2: Short Text Dense Retrieval

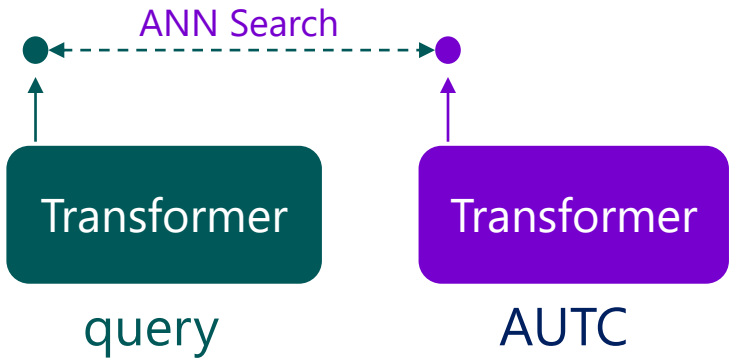
Effectively capturing user interests in short texts with dense retrieval in Bing Ads

- Ads encourages “softer matches”
- Bid keyword ensures a lower bound of relevance
- Short text fields easier for neural network-based dense retrieval
- DSSM [1], RNNs, and now Transformers
- Extreme Classification (XC) as a form of dense retrieval [2]

[1] Learning Deep Structured Semantic Models for Web Search using Clickthrough Data. Huang et al. CIKM 2013.
[2] SiameseXML: Siamese Networks meet Extreme Classifiers with 100M Labels. Dahiya et al. 2021. ICML 2021.

Breakthrough #2: Short Text Dense Retrieval

Effective relevance match with document's short text fields (AUTC) for the Web



AUTC: Anchor, URL, Title, and Click Stream

- Short texts fields
- More paraphrase-like matching with query
- Helpful mainly for common documents with AUTC

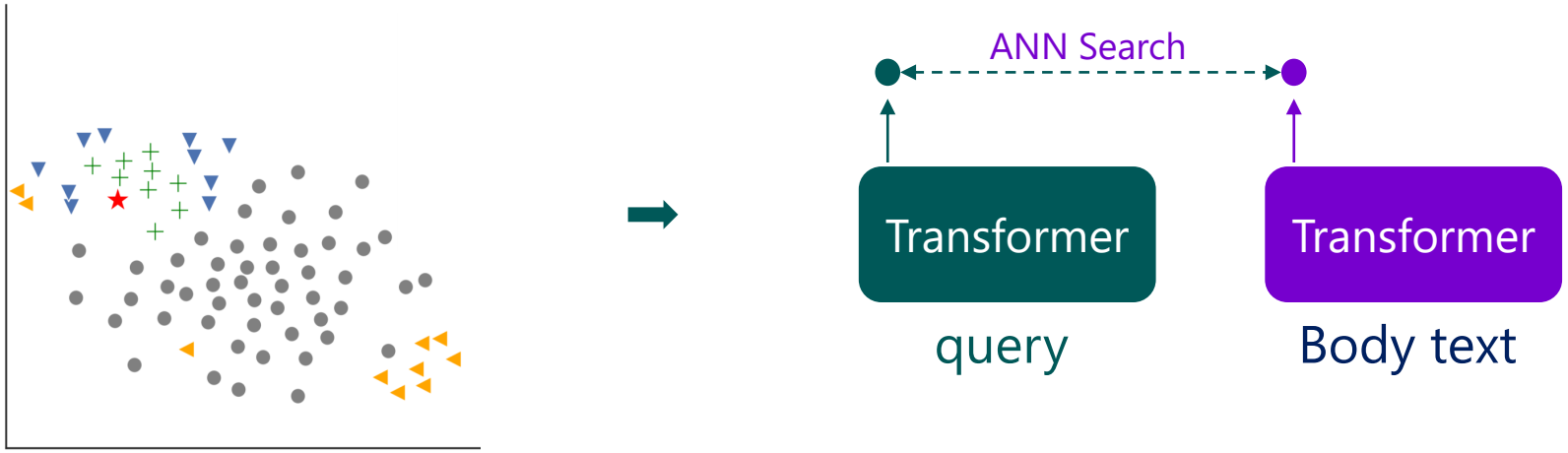
Mature dense retrieval stack at scale

- Accelerated hardware inference
- Efficient indexes for dense vectors
- .. Much more

Breakthrough #3: Full Body Vectorization Dense Retrieval

Effective Full Body Vectorization (FBV) Dense Retrieval with ANCE Training [1]

ANCE Training



ANCE enables dense retrieval model to capture relevance match with full body text of documents

- No hard dependency on AUTC: Generalize to entire web in Bing
  Generalize *beyond* the web
- 10+ recall gain
- Better information compression with 2x improvement in serving cost

[1] Approximate Nearest Neighbor Contrastive Learning for Dense Text Retrieval. Xiong et al. 2021. ICLR 2021.

The Research Path

2010 and Before

- Many attempts from community to leverage learned representations, e.g., LSI and LDA.

2014 - 2016

- Attempts to leverage word2vec and other embedding based models for search.
- Pessimistic view on embedding-based search in literature
- Many switched to reranking with term-level soft matches

2018

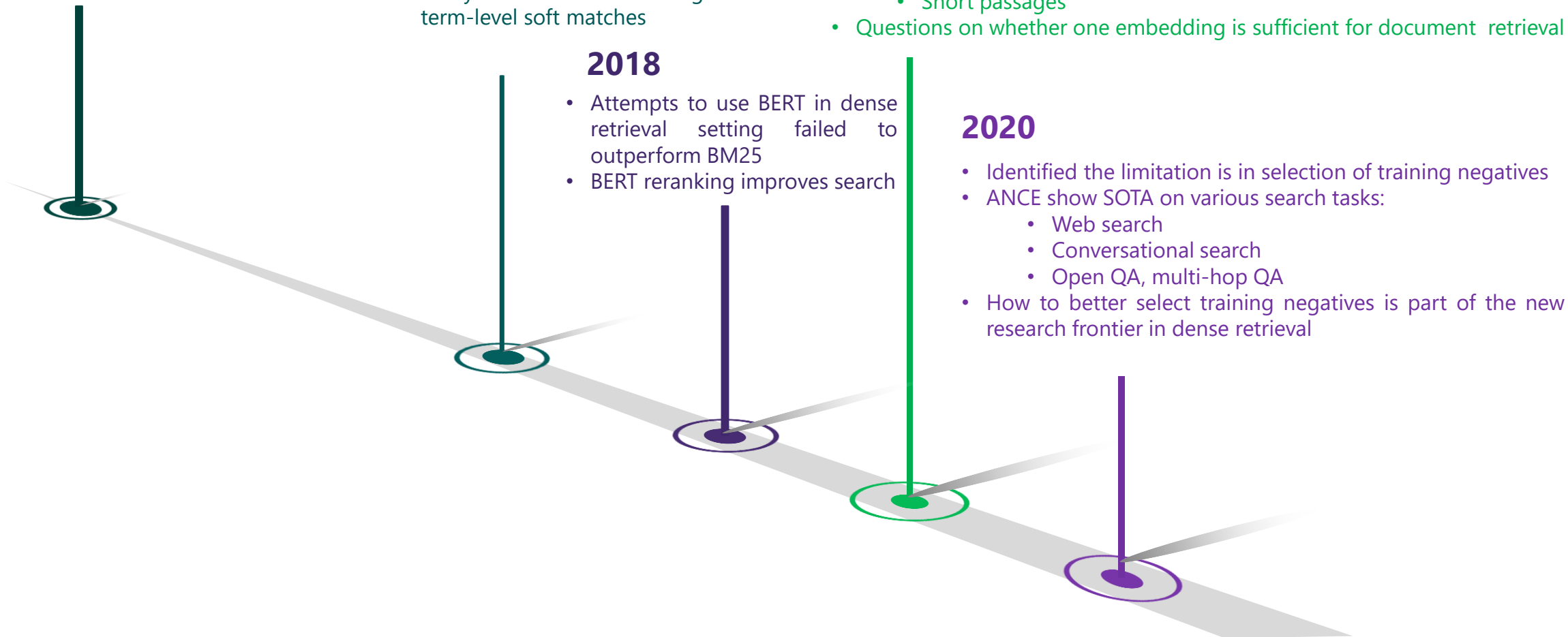
- Attempts to use BERT in dense retrieval setting failed to outperform BM25
- BERT reranking improves search

2019

- Dense retrieval with BERT performs on par with BM25 on Open-QA
  - More “paraphrase” questions
  - Short passages
- Questions on whether one embedding is sufficient for document retrieval

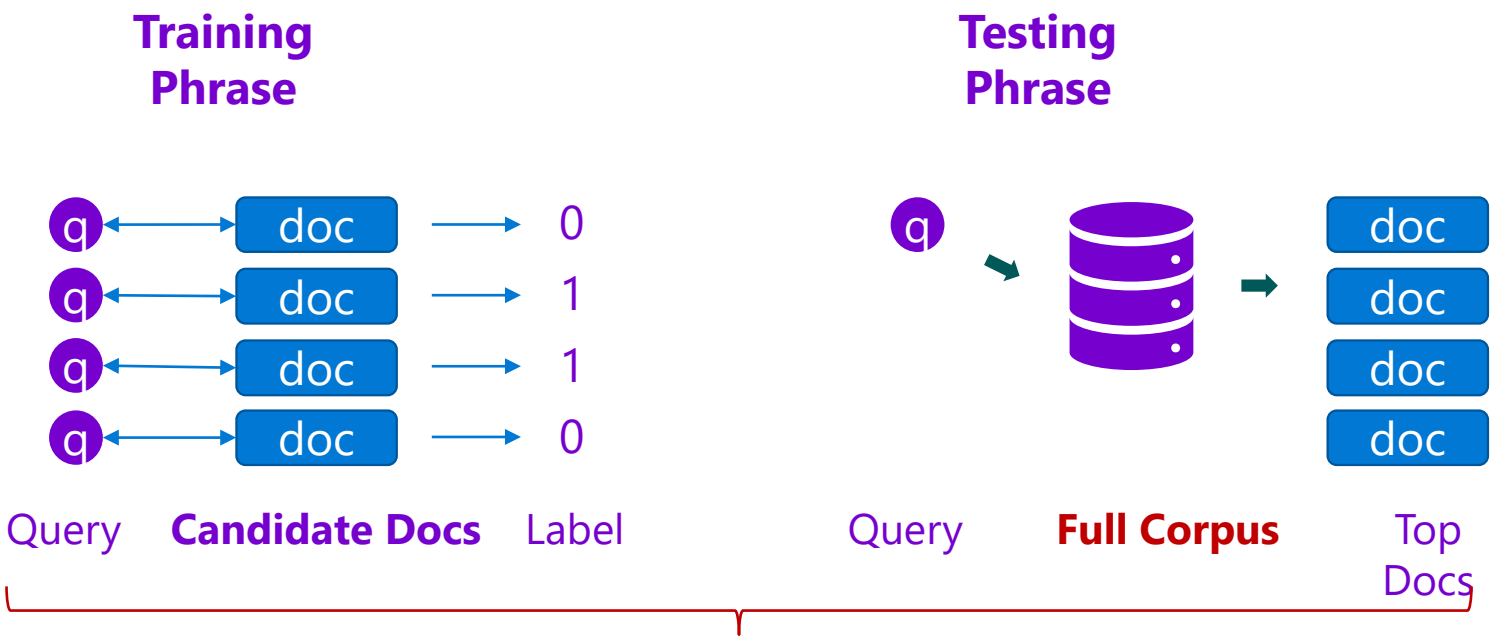
2020

- Identified the limitation is in selection of training negatives
- ANCE show SOTA on various search tasks:
  - Web search
  - Conversational search
  - Open QA, multi-hop QA
- How to better select training negatives is part of the new research frontier in dense retrieval



Why is a dense representation for text retrieval hard, intuitively?

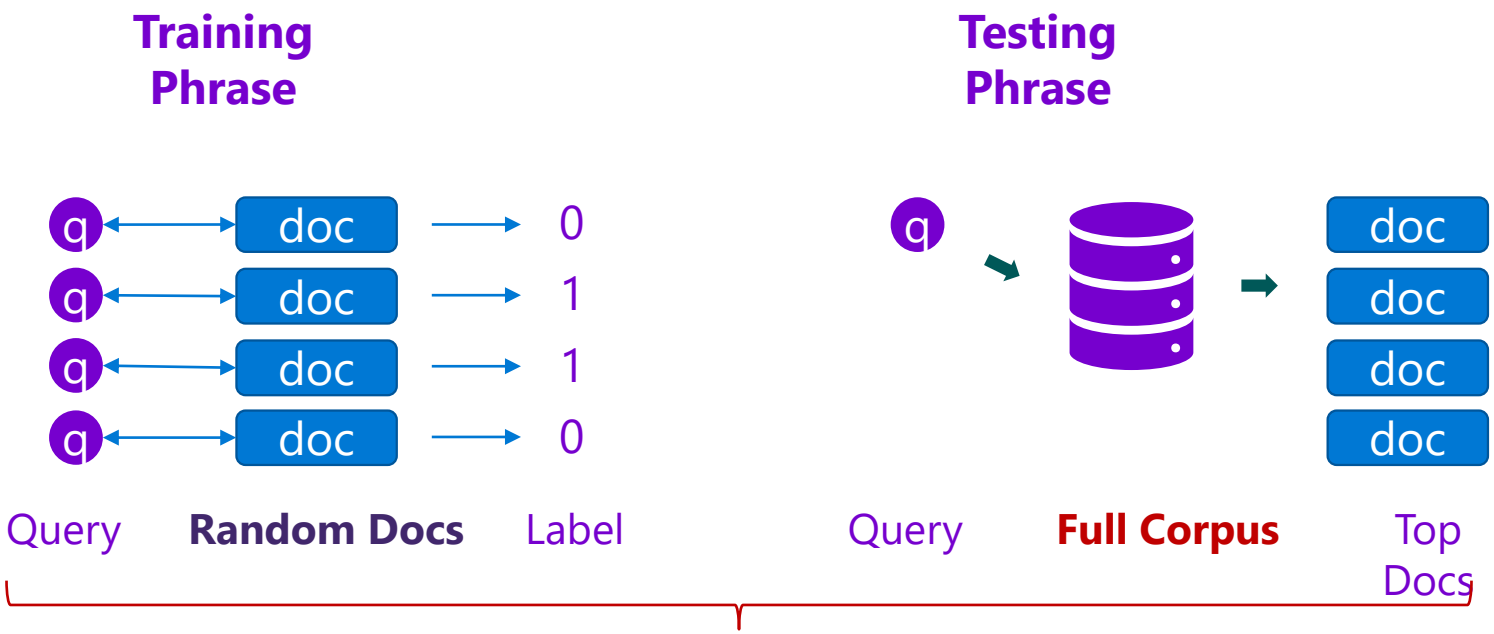
Dense Retrieval with standard learning to rank:



- Candidate docs are from existing sparse retrieval pipeline
  - Very different from those to retrieve in testing!
- Different training and testing distributions!

Why is a dense representation for text retrieval hard, intuitively?

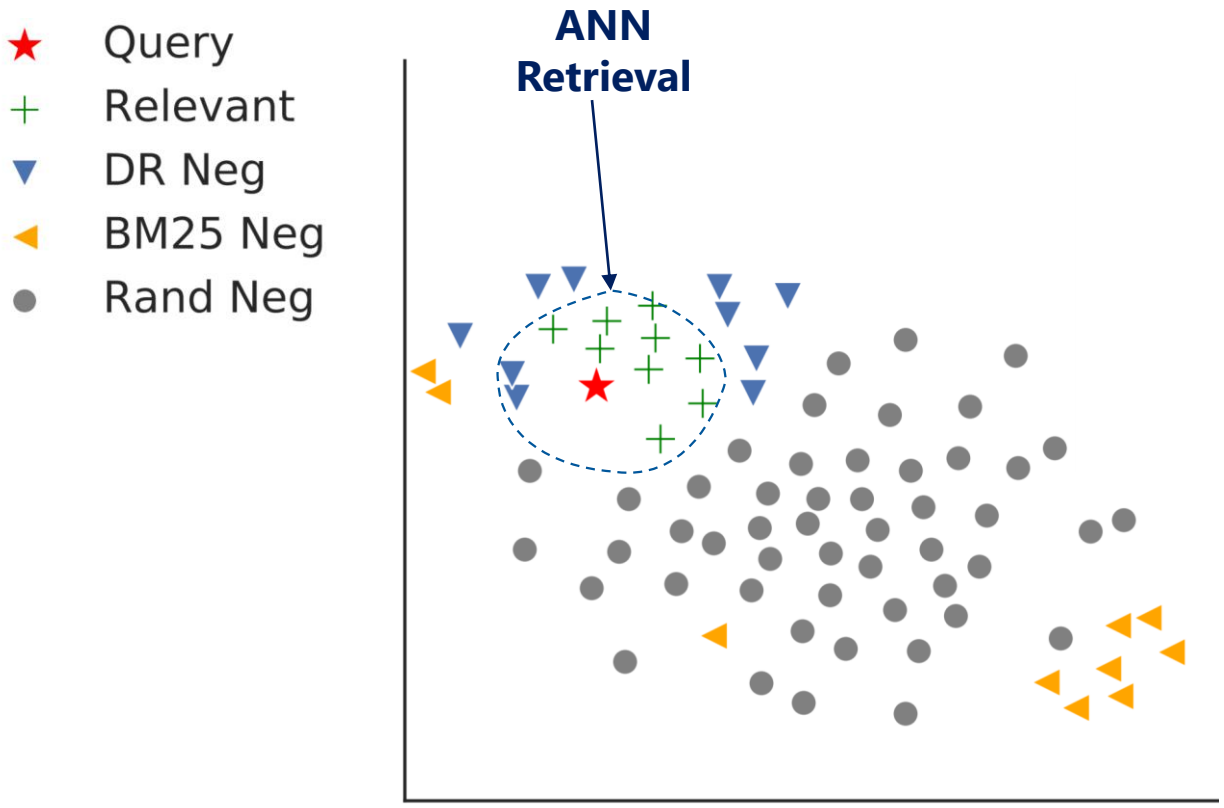
Dense Retrieval with random negatives:



- Random documents are often too easy
- Retrieval is more about distinguish relevant from hard negatives

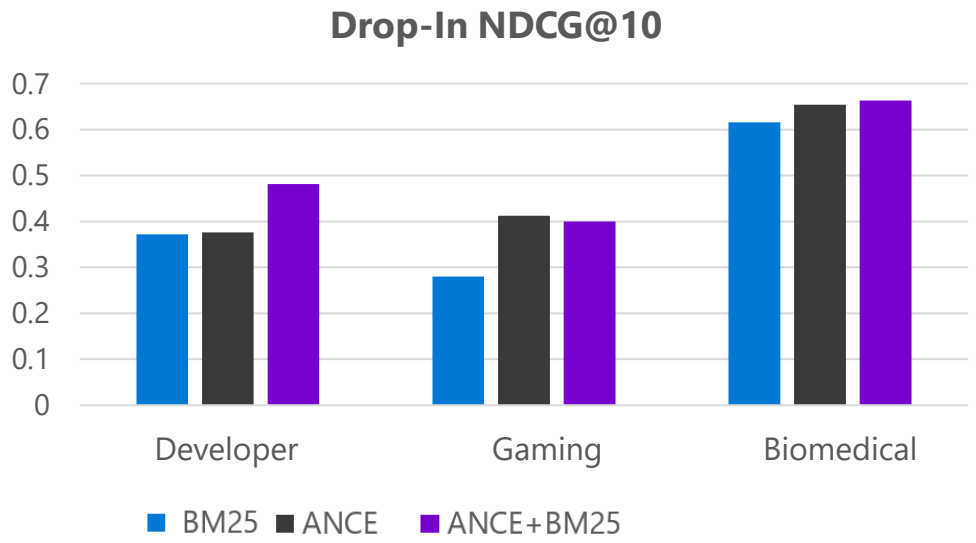
ANCE: Global ANN Negatives for Dense Retrieval Training

Dense Retrieval Inference



Generalization to Unseen

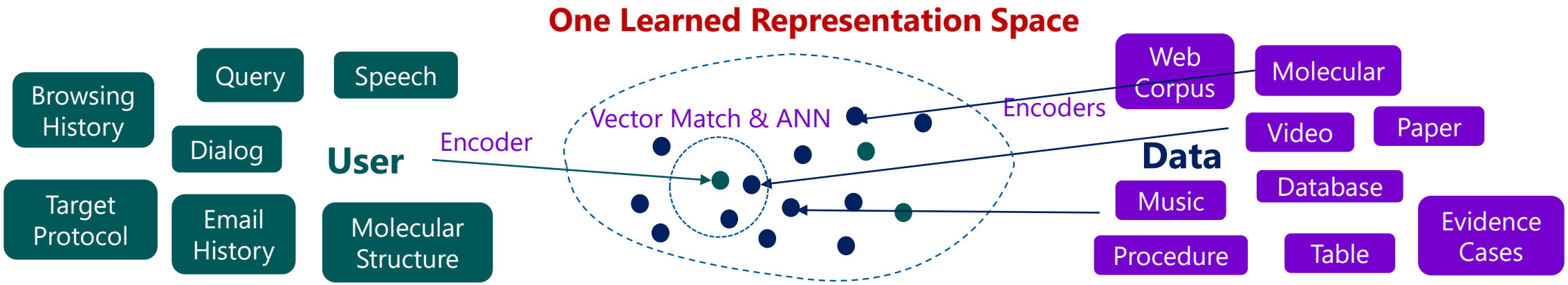
Directly apply trained ANCE models as a fixed API, no change at all, plug-and-play



BM25 + ANCE lead to stable gains on developer, gaming, and biomedical workloads.

Toward Universal Search and Recommendation

Accurate: Best search and recommendations experience with semantic understanding and user modeling
Generalizable: One solution for all search and recommendation scenarios
Scalable: Index and serving the entire world



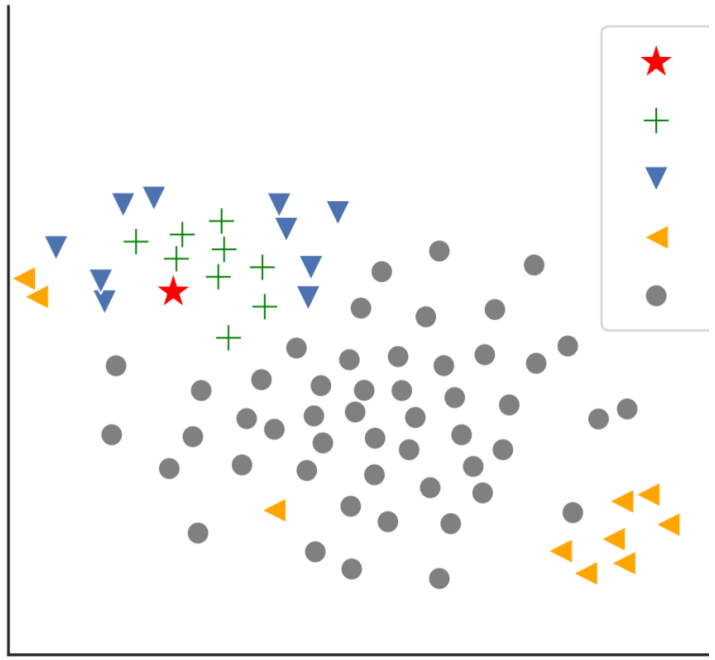
Representation Learning in Search

Goals:

- Accurately captures the semantics of content
- Strong generalization ability to all domains

Technical directions:

- **Efficient Relevance-Oriented Pretraining** that provides better starting point for universal search and recommendation.
- **Optimization techniques** that better captures user feedback signals in the representation space
- **Theoretical Understanding** of the properties of the Representation space in generalization
- **Document Representation Models** that focus on capturing semantics in long documents.
- **Multi-modal representations** that can jointly model semantics of image-text-video.



Better representation learning that improves the overall performance and generalization ability of universal search and recommendation

Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training, Li et al., AAAI 2020.
M3P: Learning Universal Representations via Multitask Multilingual Multimodal Pre-training, Ni et al., CVPR 2021.

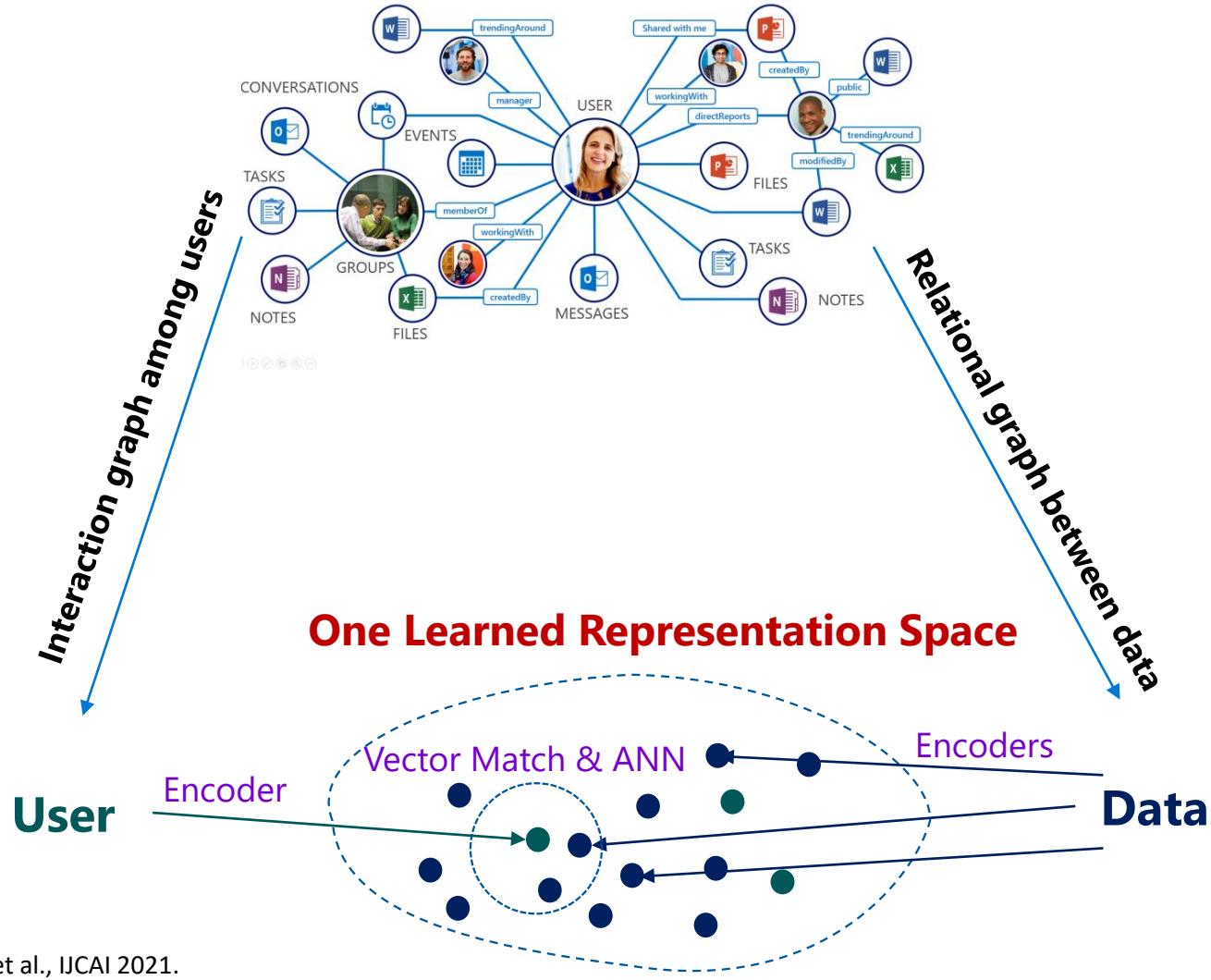
Graph/Structure Learning

Goals:

- Improve universal embedding estimation by leveraging relationships among (1) users, (2) data elements, and (3) interactions between people and data.

Technical directions:

- **Content+graph** learning to leverage implicit relational signals in universal embeddings
- **Composite embedding** methods to dynamically compute implicit representations that reflect temporal context
- **Pretrained graph models** to generalize across different settings.
- Public-**private learning** techniques to preserve privacy in recommendation settings.



User-as-Graph: User Modeling with Heterogeneous Graph Pooling for News Recommendation. Wu et al., IJCAI 2021.

Uni-FedRec: A Unified Privacy-Preserving News Recommendation Framework for Model Training and Online Serving. Qi et al., EMNLP Findings 2021.

Attentive Knowledge-aware Graph Convolutional Networks with Collaborative Guidance for Recommendation. Chen et al.,

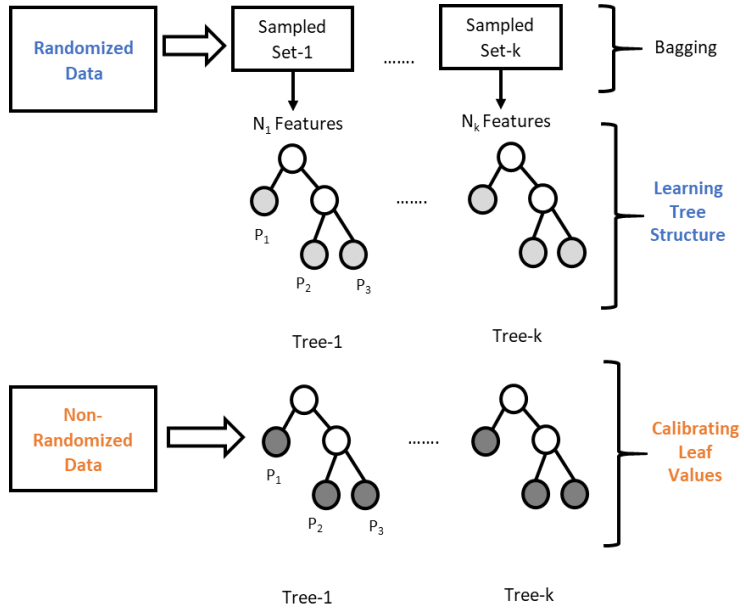
Causality and Robustness

Goals:

- Improve **robustness and generalizability** in dynamic environments (e.g., changing traffic, corpus), across domains and into the tail.
- Target **rewards beyond clicks**, such as long-term rewards, user outcomes, empowerment.

Technical directions:

- **Algorithmic advances**
  - Inverse reinforcement learning
  - Causal representation learning
  - Causal extreme learning
- **Incorporation of domain knowledge**
  - Simulations + causal modeling for decision making
  - Leveraging knowledge of experiments and environments
- **Causal ML under real-world constraints**
  - Causal modeling and online ML pipelines
  - Federated causal ML



Relying on causal rather than correlational relationships provides more stable and generalizable foundation for machine learning

Causal Transfer Random Forest: Combining Logged Data and Randomized Experiments for Robust Prediction, Zeng et al., WSDM 2021.
Domain Generalization using Causal Matching, Mahajan et al., ICML 2021.

Universal Search and Recommendation


Office


Azure


Health


Dynamics


LinkedIn


Xbox


Bing


GitHub

Provide push-button customized search and recommendation

.... for any person,

any vertical,

any data.