Artificial Intelligence Applied to the Web

Chapter 3 Part 3 - Modern Web Search Approaches

Diego Cornejo, Felipe Hernández and Juan Velásquez

University of Chile
Departament of Industrial Engineering

Spring 2025

Outline

Modern Web Search Approaches

Motivation

The Current State of Search

- The search box has become the default user interface for interacting with data in most modern applications.
- We not only search explicitly, but also consume content streams customized to our tastes and interest.
- The concept of a “search engine” goes beyond websites like Google; it is present in nearly all our digital interactions.

Motivation

The Impact of LLMs and New Expectations

- The arrival of technologies like ChatGPT, Claude, and Gemini has skyrocketed expectations for the intelligence level of search technologies.
- Users no longer expect just a list of “ten blue links”, but a much more sophisticated search experience.
- These Language models are being actively integrated into major search engines, influencing their evolution.

Motivation

User Expectations for Current Search Technology

Users today expect search technology to be:

- **Domain-aware**: It should understand entities, terminology, and categories specific to the user case, not just generic text statistics.
- **Contextual and personalized**: It should take into account user context (location, previous searches, profile) and domain context (inventory, business rules) to better interpret intent.
- **Conversational**: It should be able to interact in natural language and guide users through a multi-step discovery process.

Motivation

User Expectations for Current Search Technology

- **Multi-modal**: It should be able to resolve queries issued by text, voice, or images, and search across these different content types.
- **Intelligent**: It should deliver predictive type-ahead, understand what users mean (spelling correction, intent classification), and constantly get smarter.
- **Assistive**: It should move beyond delivering links to provide direct answers, summaries, explanations, and available actions.

Motivation

The Critical Role of the Search engine in Modern AI

- Even the best conversational AI models can "hallucinate" (make up bad answers) if they are not tethered to a reliable information source, such as a search engine index.
- **Retrieval Augmented Generation** (RAG) emerges as the stardard technique to ground AI systems.
- In RAG, a search engine or vector database is used as a knowledge source to provide LLMs with accurate and up-to-date information as context.

Motivation

The Goal of AI-Powered Search

- The goal is to use automated machine learning techniques to deliver on all the desired search capabilities (domain-aware, contextual, conversational, etc.).
- While many organizations spend years manually tuning their systems, AI-Powered Search aims to automate most of that process.

What is AI-Powered Search?

Fundamental Definitions

- **Artificial Intelligence (AI)**: In the context of software development, AI generally describes any computer program that can perform a task that previously required human intelligence.
- **Search**: Refers to any technology that enables users to query for and find information. It involves two critical steps:
  - **Matching**: Finding documents that match a query.
  - **Ranking**: Ordering those documents by relevance to the query.

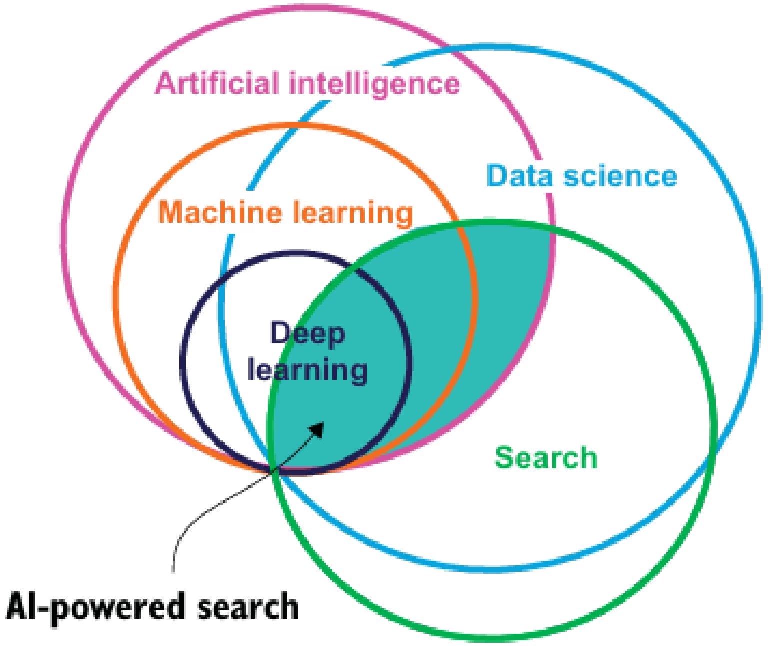
What is AI-Powered Search?

A Conceptual Diagram

- **AI-Powered Search** is the intersection of the fields of search (information retrieval) and Artificial Intelligence.
- The relationship between key disciplines is as follows:
  - **Machine Learning (ML)**: A subset of AI that use data to train models to perform tasks.
  - **Deep Learning (DL)**: A further subset of ML that focuses on training artificial neural networks.
  - **Data Science**: A discipline that heavily overlaps with both AI and search but is not a complete subset of either.

What is AI-Powered Search?

A Conceptual Diagram

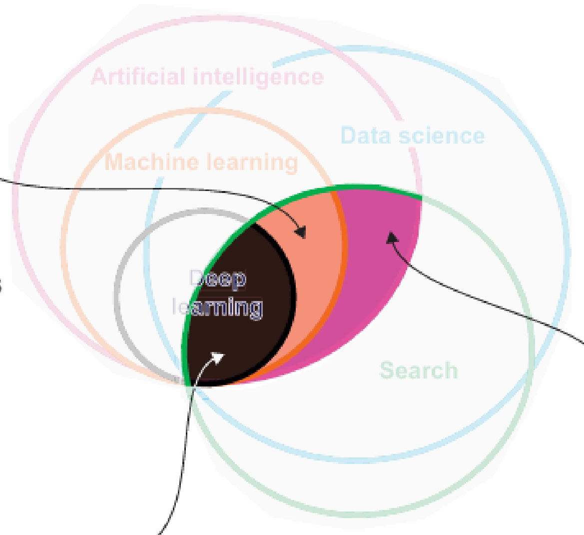


What is AI-Powered Search?

Specific AI-Powered Search Techniques

Machine learning (ML):

- **Signals boosting models**
- **Learning to rank**
- **Semantic search**
- **Collaborative filtering**
- **Personalized search**
- **Content clustering**
- **NLP/entity resolution**
- **Semantic knowledge graphs**
- **Document classification**
- ***Deep learning***
- **Etc.**



Artificial intelligence (AI):

- **Question/answer systems**
- **Virtual assistants**
- **Chatbots**
- **Rules-based relevancy**
- ***Machine learning***
- **Etc.**

Deep learning (DL):

- **Foundation models/LLMs**
- **Neural search/vector search**
- **Word embeddings**
- **Multimodal search (image, video, etc.)**
- **Generative search and summarization**
- **Etc.**

What is AI-Powered Search?

Specific AI-Powered Search Techniques

- **AI-only category**:
  - Includes AI techniques that are often build using machine learning but do not fundamentally require it.
  - Examples of these techniques include question-answering systems, virtual assistants, chatbots, and rules-based relevancy.
  - It is possible to build chatbots based entirely on rules to understand user intents, and question-answering systems can function solely on rules and ontologies.
  - However, the lines between categories are often blurred because machine learning is frequently used to **learn these types of rules and ontologies**.

What is AI-Powered Search?

Specific AI-Powered Search Techniques

- **Machine Learning category**:
  - This subcategory begins when algorithms start using data to train models.
  - It uses **behavioral signals** from users (clicks, likes, purchases, etc.) to build models that learn to better rank documents.
  - ML is also used to learn **knowledge graphs** to better understand the domain and interpret user queries.
  - This enables **semantic search** through knowledge graphs and other natural language processing techniques.

What is AI-Powered Search?

Specific AI-Powered Search Techniques

- **Deep Learning category**:
  - Involves the use of neural networks to build models that can understand user queries and documents, as well as rank and summarize search results.
  - Text is used to train LLMs to understand the meaning of words and phrases, generate answers, and create summaries.
  - LLMs are a type of model that interprets text content, often trained on massive amounts of text from the internet. Can also be trained on images, audio, or video to enable **multimodal search** (e.g., text-to-image).

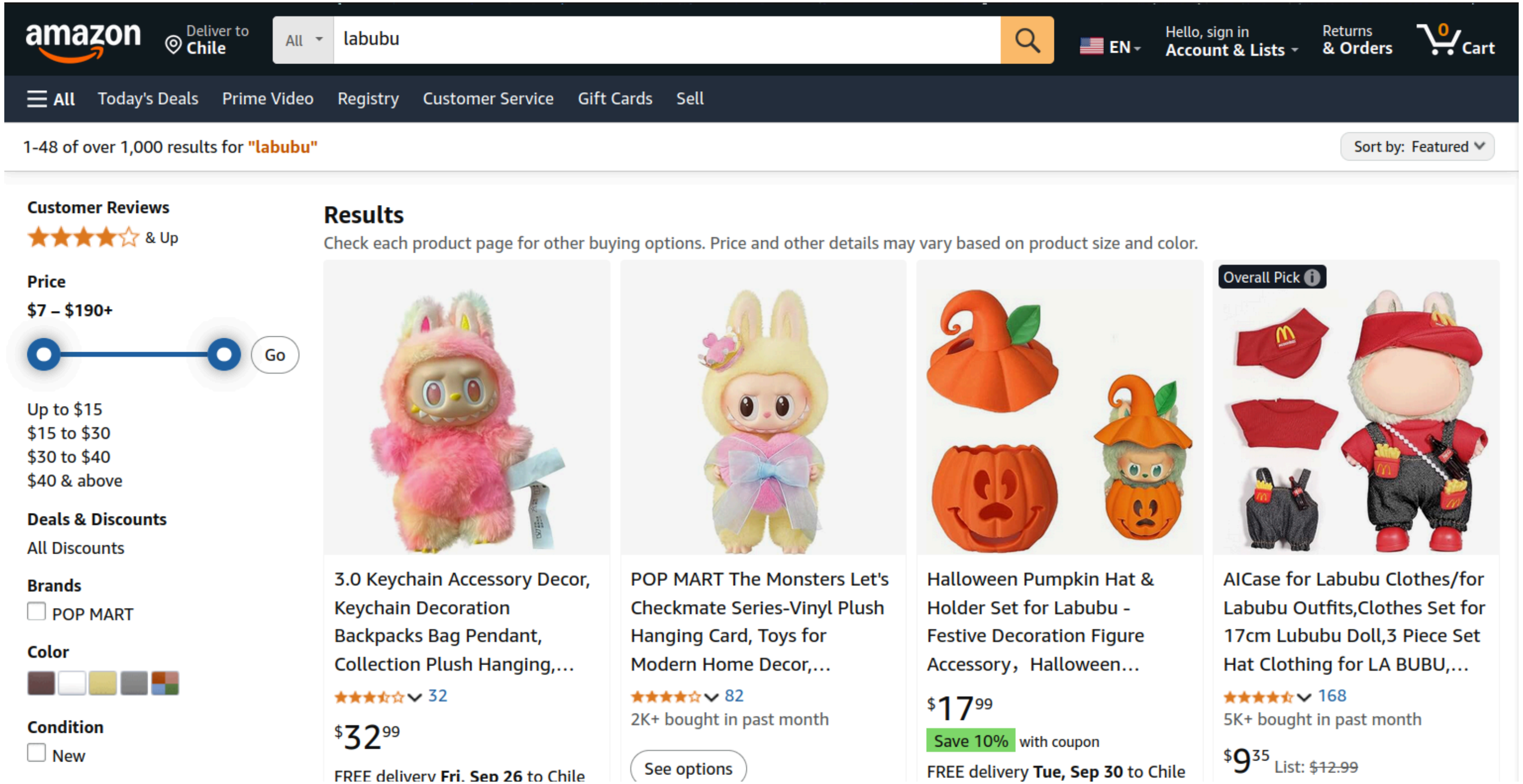
Understanding User Intent

Search Engine vs Recommendation Engines

- **Search Engine**:
  - Typically thought of as a technology for explicitly entering queries and receiving a response.
    - Usually exposed via a text box direct discovery of content.
- **Recommendation Engine**:
  - Typically does not accept direct user input.
  - Delivers content based on what the engine learns about users, calculating best matches for their interest and behaviors.
  - Commonly uses three approaches: content-based, behavior-based, and multimodal recommenders.

Understanding User Intent

Search Engine

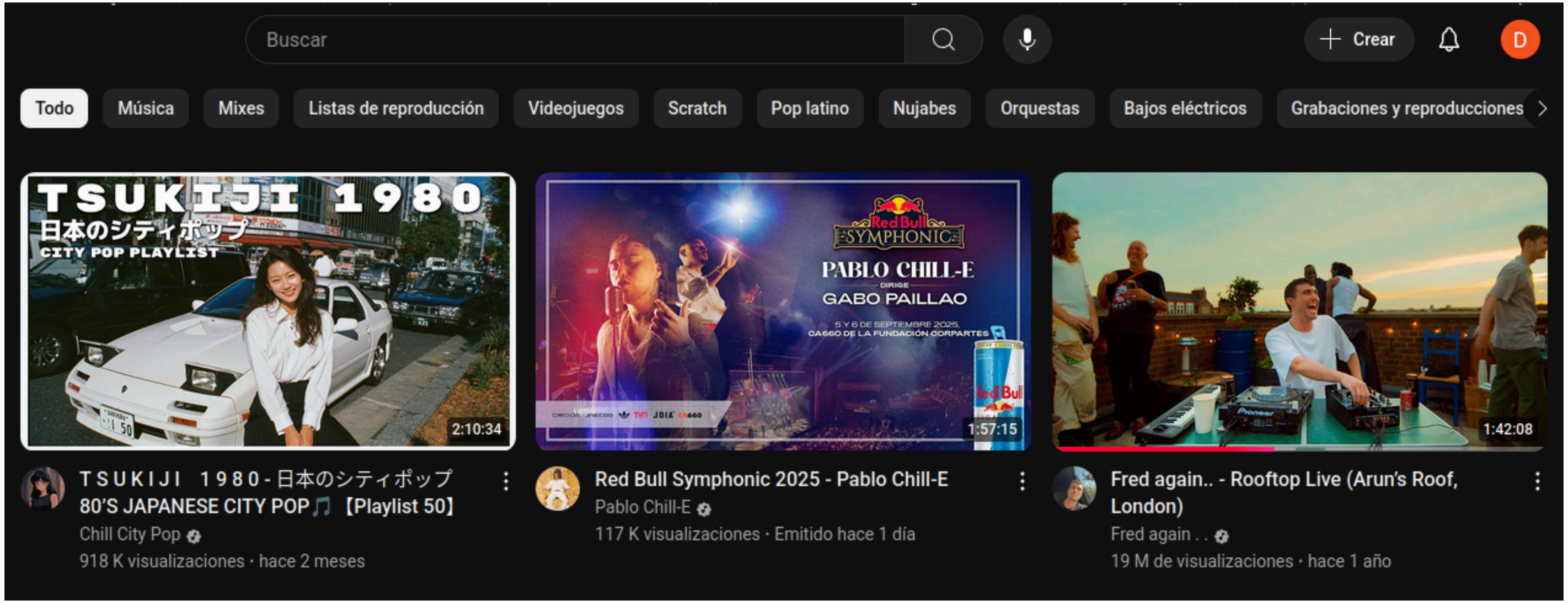


Understanding User Intent

Recommendation Engine

Understanding User Intent

Recommendation Engine



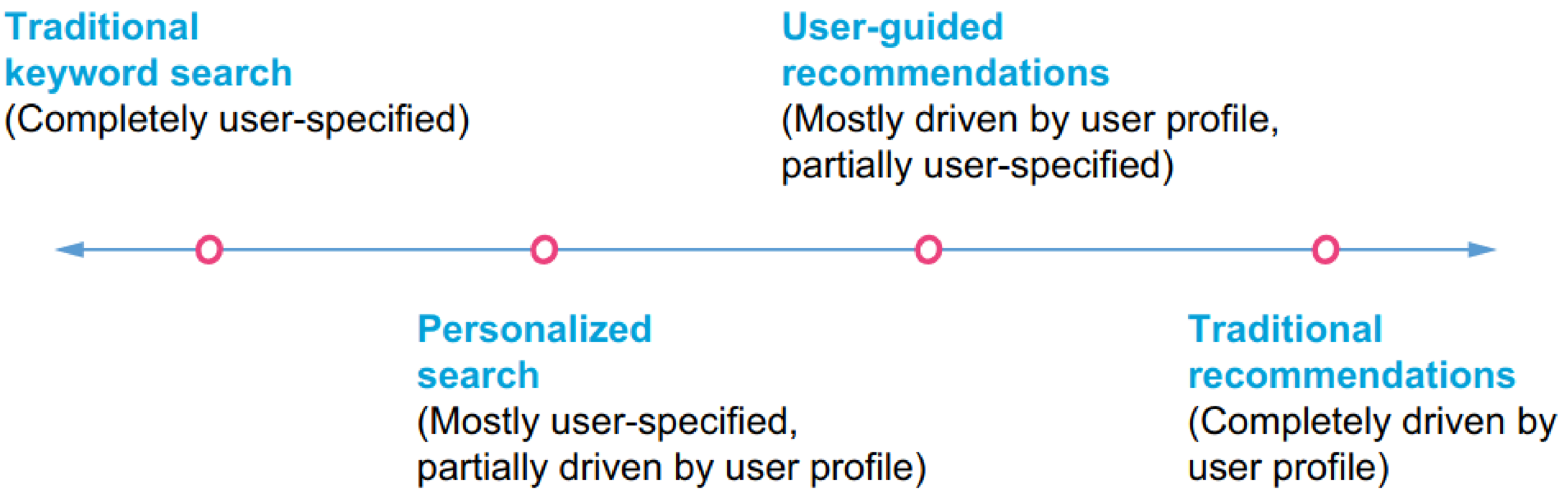
Understanding User Intent

The Personalization Spectrum

- Search and recommendation engines are not separate systems but two sides of the same coin.
- The goal in both cases is to understand a user's information need and deliver relevant results.
- They exist on a spectrum of personalization:
  - **Traditional Keyword Search**: Completely user-specified.
  - **Personalized Search**: Mostly user-specified, partially driven by user profile.
  - **User-Guided Recommendations**: Mostly driven by user profile, partially user-specified.
  - **Traditional Recommendations**: Completely driven by user profile.

Understanding User Intent

The Personalization Spectrum



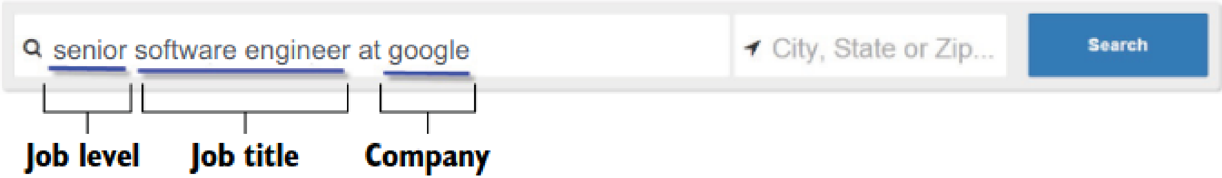
Understanding User Intent

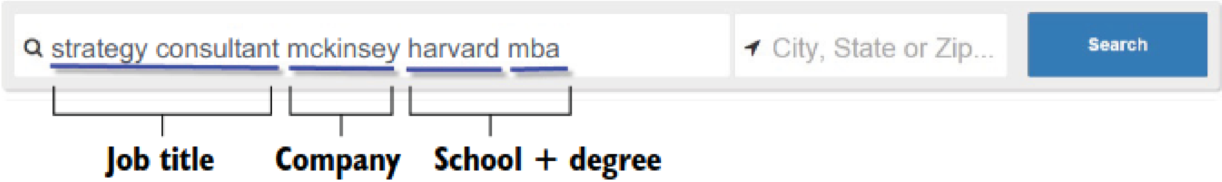
The Third Dimension: Domain Understanding

- It is not enough to just match keywords and recommend based on user interactions; the engine must also deeply understand the specific domain.
- This includes:
  - Learning all important domain-specific phrases, synonyms, and related terms.
  - Identifying entities in documents and queries.
  - Generating a knowledge graph that relates those entities.
  - Disambiguating the many nuanced meanings of domain-specific terminology.
- The ultimate goal is to **search on “things”, not “strings”.**

Understanding User Intent

The Third Dimension: Domain Understanding





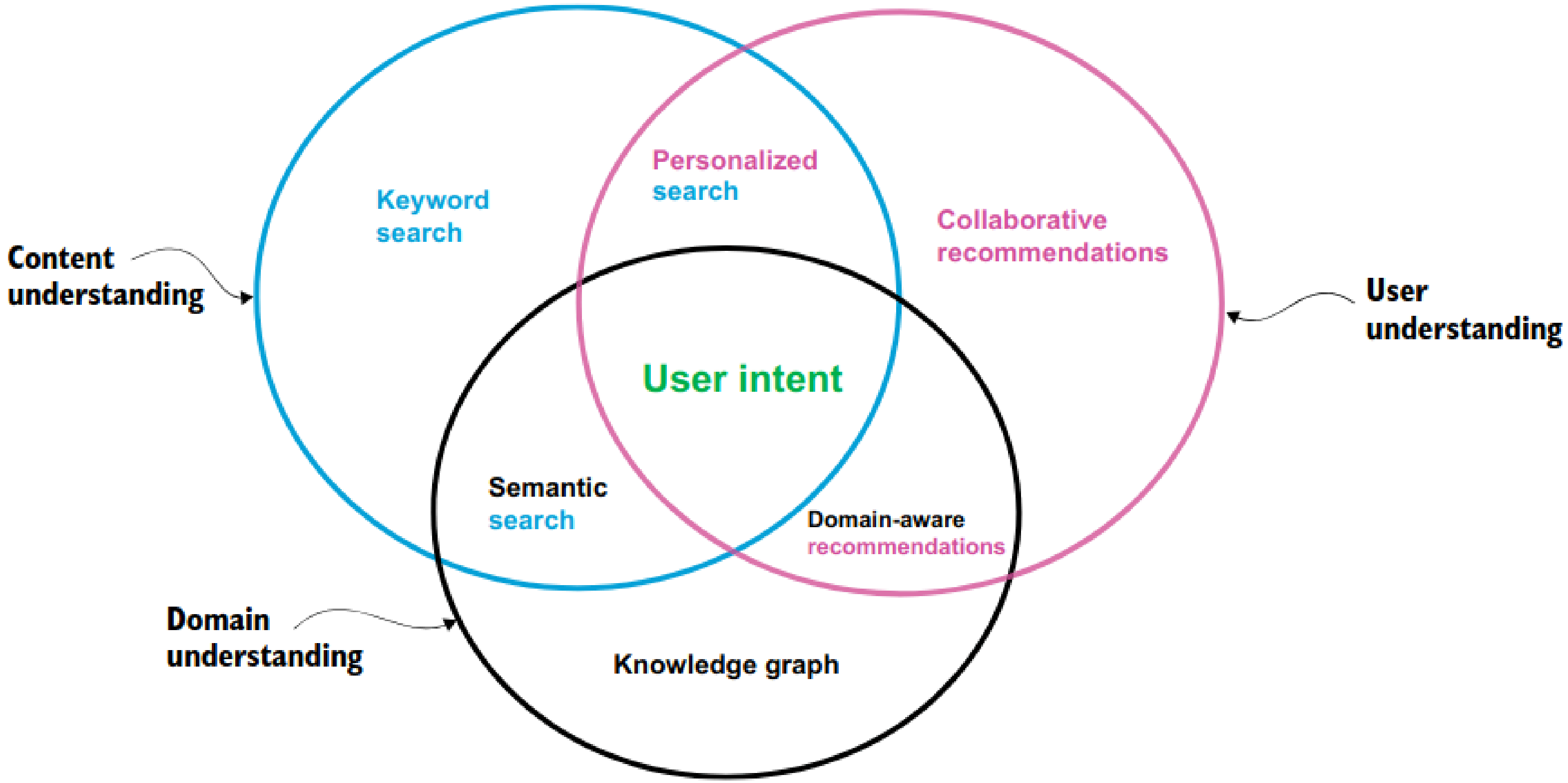
Understanding User Intent

The Three Pilars of User Intent

- To truly understand user intent, an AI-powered search system need to combine three key pilars:
  - **Content Understanding**: The ability to find the right content based on keywords, language patterns, and attributes. This corresponds to traditional keyword search.
  - **User Understanding**: The ability to understand each user's specific preferences to return more personalized results. This corresponds to collaborative recommendations.
  - **Domain Understanding**: The ability to interpret words, phrases, concepts, entities, and relationships within your domain-specific context. This is powered by a knowledge graph.

Understanding User Intent

Achieving True User Intent



Understanding User Intent

Achieving True User Intent

- The intersection of these pillars create more advanced search capabilities:
  - Content + User Understanding: **Personalized Search**.
  - Content + Domain Understanding: **Semantic Search**.
  - User + Domain Understanding: **Domain-Aware Recommendations**.
- The "holy grail" for AI-powered search is to harness the intersection of all three categories. This requires:
  - An expert understanding of the domain.
  - An expert understanding of the users and their preferences.
  - And expert ability to match and rank arbitrary queries against any content.

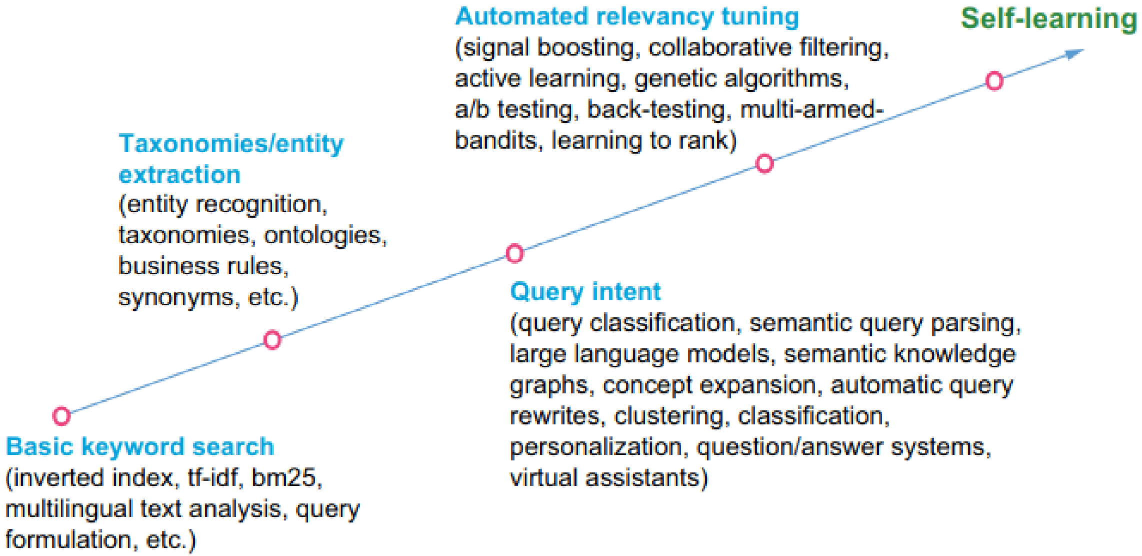
How AI-Powered Search Works

The Search Intelligence Progression

- Search intelligence typically matures along a predictable path.
- The stages of this progression are:
  - **Basic keyword search**: The typical starting point, using foundational algorithms like inverted index.
  - **Taxonomies/entity extraction**: Manually injecting domain understanding with synonyms, taxonomies, ontologies, and business rules.
  - **Query intent**: Focusing on correctly interpreting user queries through classification, semanting parsing, and knowledge graphs.
  - **Automated relevancy tuning**: Automating the tuning process through learning from user signals, A/B testing, and building machine-learning models.
- The end goal is a completely automated, **self-learning** engine.

How AI-Powered Search Works

The Search Intelligence Progression



How AI-Powered Search Works

Reflected Intelligence through feedback loops

- Feedback loop are critical to building an AI-powered search solution.
- Without feedback, it's like an education consisting only of reading textbooks with no teachers, exams, or classmates, which would lead to a flawed understanding.
- Traditional search engines often operates this way, acting on their initial configurations the same way every time for repeated user queries.
- However, search engines are the perfect type of system for interactive learning when feedback loops are introduced.

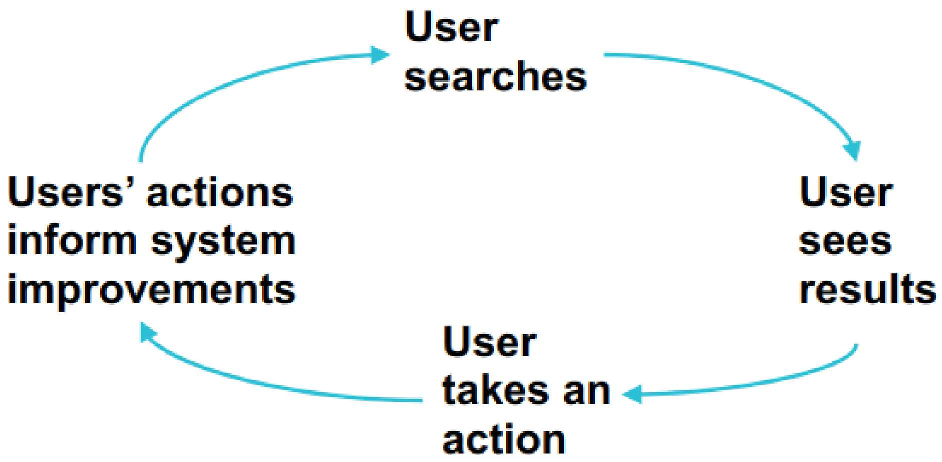
How AI-Powered Search Works

The Flow of a Search Feedback Loop

- Step 1: A user issues a query. This query executes a search.
- Step 2: The system return results. These can be specific answers, a list of answers, or links to pages.
  - Step 3: The user takes one or more actions.
    - These actions usually start with clicks on documents.
    - They can lead to other context-specific actions, such as:
      - Adding an item to a shopping cart and purchasing it.
      - Giving an item a thumbs up or thumbs down.
      - Liking or commenting on the result.

How AI-Powered Search Works

The Flow of a Search Feedback Loop



These actions can then be used to generate an improved relevance ranking model for future searches.

How AI-Powered Search Works

The Power of Signals

- User interactions like searches, clicks, likes, add-to-carts, and purchases are collectively referred to as **signals**.
- Signals provide a constant stream of feedback that can be used by machine learning algorithms to power user, content, and domain understanding.

How AI-Powered Search Works

The Power of Signals

- Key refelected intelligence algorithms powered by signal include:
  - **Popularized relevance**: Signals-boosting algorithms use aggregated signals to boost the rankings of the most important documents for popular queries.
  - **Personalized relevance**: Collaborative filtering algorithms use signals to generate recommendations and user profiles to personalize search results.
  - **Generalized relevance**: Learning to rank algorithms train ranking classifiers based on relevance judgments generated from user signals to create models that can apply to all queries.

How AI-Powered Search Works

The Power of Signals

Signal collection & processing

pizza

User searches

User sees results

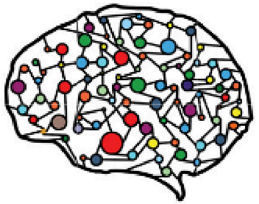
User takes an action

Users' actions inform system improvements

User	Query	Results
Alonzo	**pizza**	doc10, doc22, doc12, …
Elena	**soup**	doc84, doc2, doc17, …
Ming	**pizza**	doc10, doc22, doc12, …
…	…	…

User	Action	Document
Alonzo	**click**	doc22
Elena	**click**	doc17
Ming	**click**	doc12
Alonzo	**purchase**	doc22
Ming	**click**	doc22
Ming	**purchase**	doc22
Elena	**click**	doc2
…	…	…

Learned relevance models



Signal processing & machine learning

How AI-Powered Search Works

Content and Domain Intelligence

- While signals provide usage data, the **content** is also a rich source of information for feedback loops.
- The content of the documents forms a representative textual model of your domain.
- **Large Language Models** (LLMs) based on the **Transformer architecture** have revolutionized query and content interpretation.
  - They are deep neural networks trained on massive amounts of text.
  - They can recognize, summarize, predict, and generate new data.
  - They are used to generate **embeddings**, which are numerical vector representations of content's meaning, enabling a sophisticated ability to search on a query's meaning.

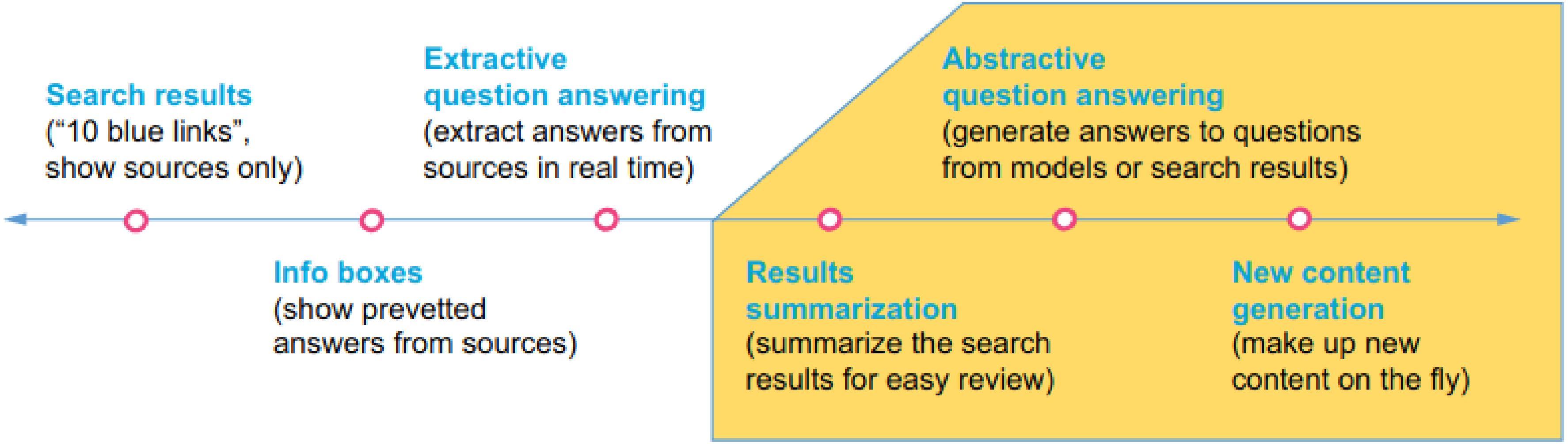
How AI-Powered Search Works

Generative AI & Retrieval Augmented Generation (RAG)

- Generative AI and AI-powered search are tightly intertwined.
- **Retrieval Augmented Generation (RAG)**: Search engines are used as a knowledge source for LLMs, allowing relevant context to be retrieved and passed to the LLM to ensure it has up-to-date and accurate data from which to answer. This is critical for preventing LLMs from "hallucinating".
- **Generative Search**: In turn, LLMs are critical components of search engines. They can be used to:
  - Interpret queries and generate embeddings.
  - Generate summaries of search results.
  - Generate answers to questions directly from search results.

How AI-Powered Search Works

Generative AI & Retrieval Augmented Generation (RAG)



How AI-Powered Search Works

Architecture for an AI-Powered Search Engine

- An end-to-end system for continuous learning requires several key building blocks:
  - A **core search engine** (indexing, matching, ranking).
  - **Index-time transformations** to enrich content as it's indexed (e.g., creating embeddings, extracting entities).
  - **Query pipelines** to interpret incoming queries (e.g., correcting misspellings, expanding with synonyms, rewriting the query).
  - A **job processing framework** to run batch jobs on content and signals to derive domain-specific intelligence.
  - A mechanism for **collecting the constant stream of user signals**.
  - Generated **models** that are used to constantly adjust future search results.

Artificial Intelligence Applied to the Web

Chapter 3 Part 3 - Modern Web Search Approaches

Diego Cornejo, Felipe Hernández and Juan Velásquez

University of Chile
Departament of Industrial Engineering

Spring 2025