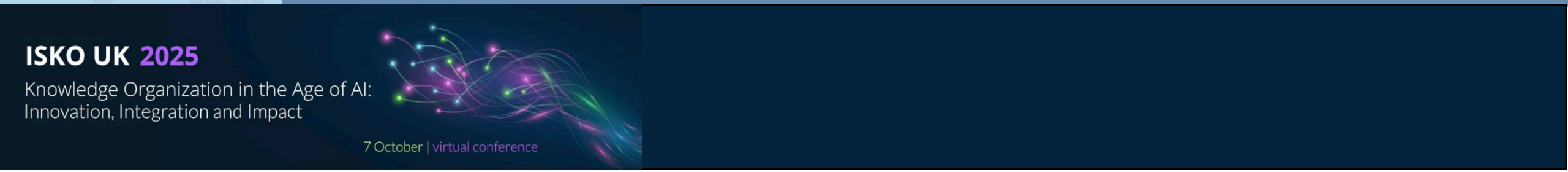
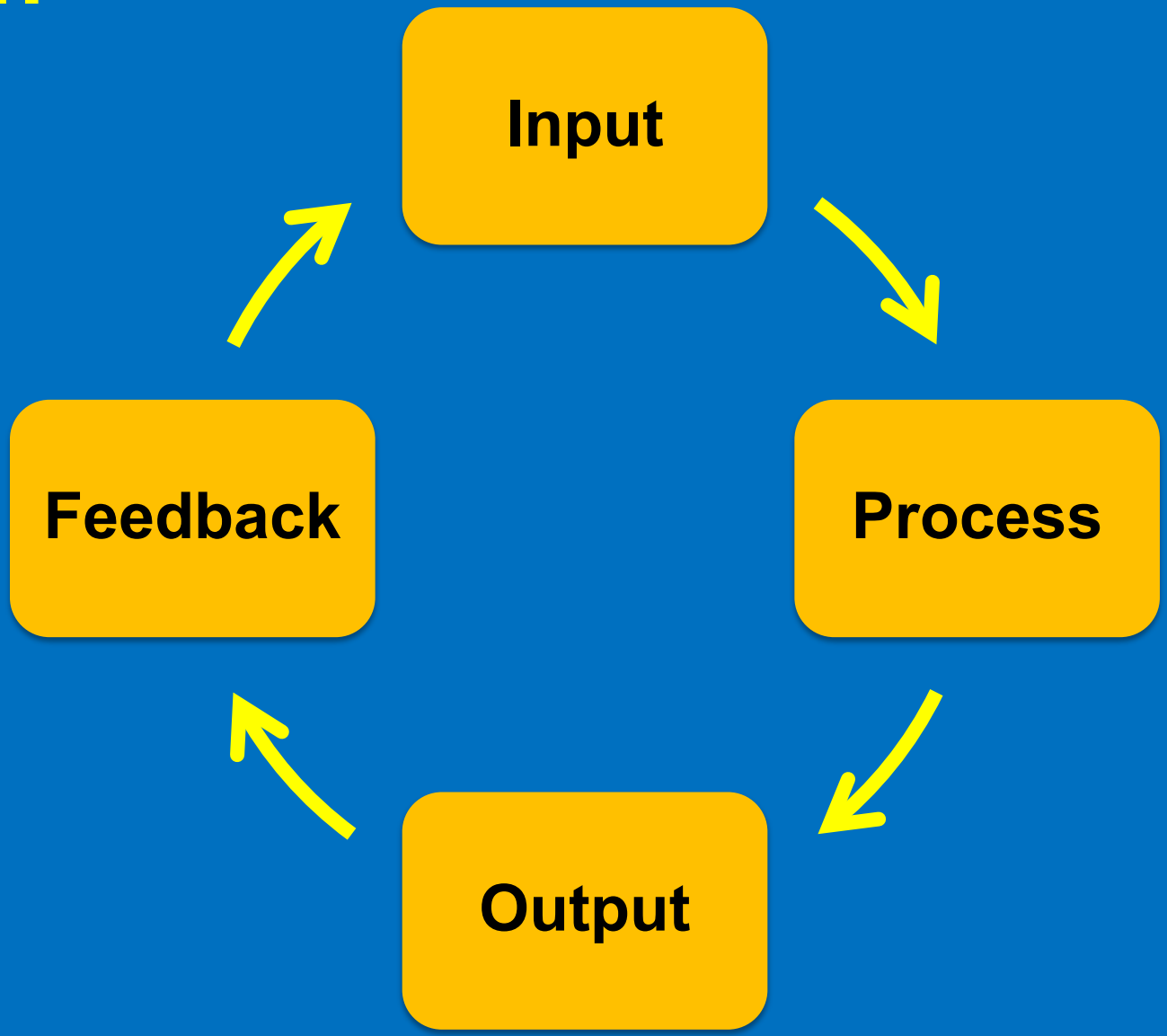


The Case for General Purpose Categorizers as Part of the AI Ecosystem

Joseph Busch

Copyright 2025 Taxonomy Strategies LLC. All rights reserved.

Datafusion



Content processing pipeline

Categorize content: What is the content about?

Identify people, organizations, places, events, things: Identify and disambiguate entities (Kotonu is the French name of the capital of Benin; Apple the technology company vs. Apple the record company)

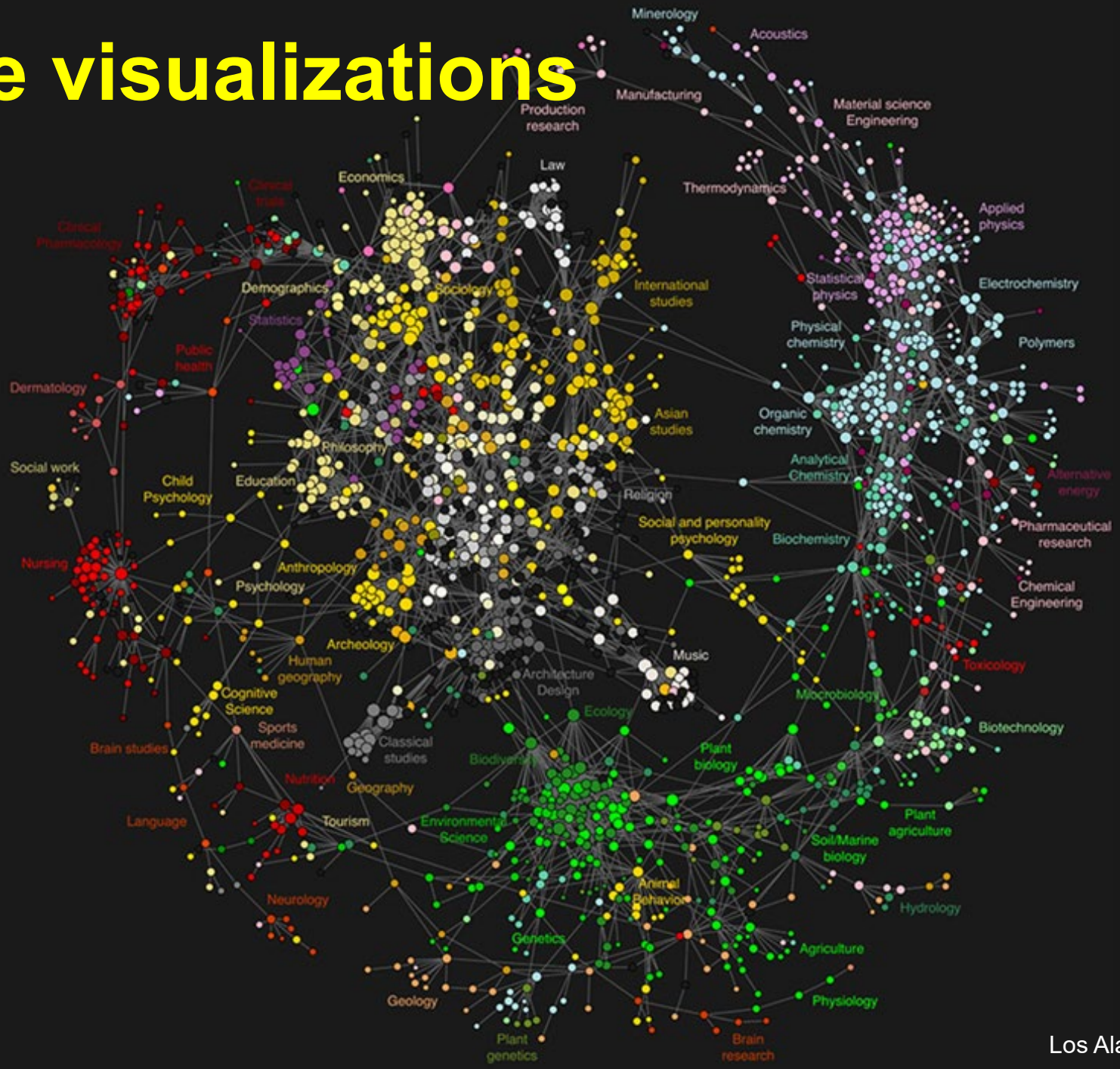
Identify grammatical relationships in sentence: Find subjects and objects of all verbs in the text

Extract actions: Identify related series of verbs (enter restaurant, sit down, read menu, order food, eat food, pay bill, leave)

Normalize disparate names for entities: Change “Kotonu” to long/lat - 6.3703° N/2.3912° E, “2.4 million” to 2,400,000, and “2006” to 1/1/2006-12/31/2006

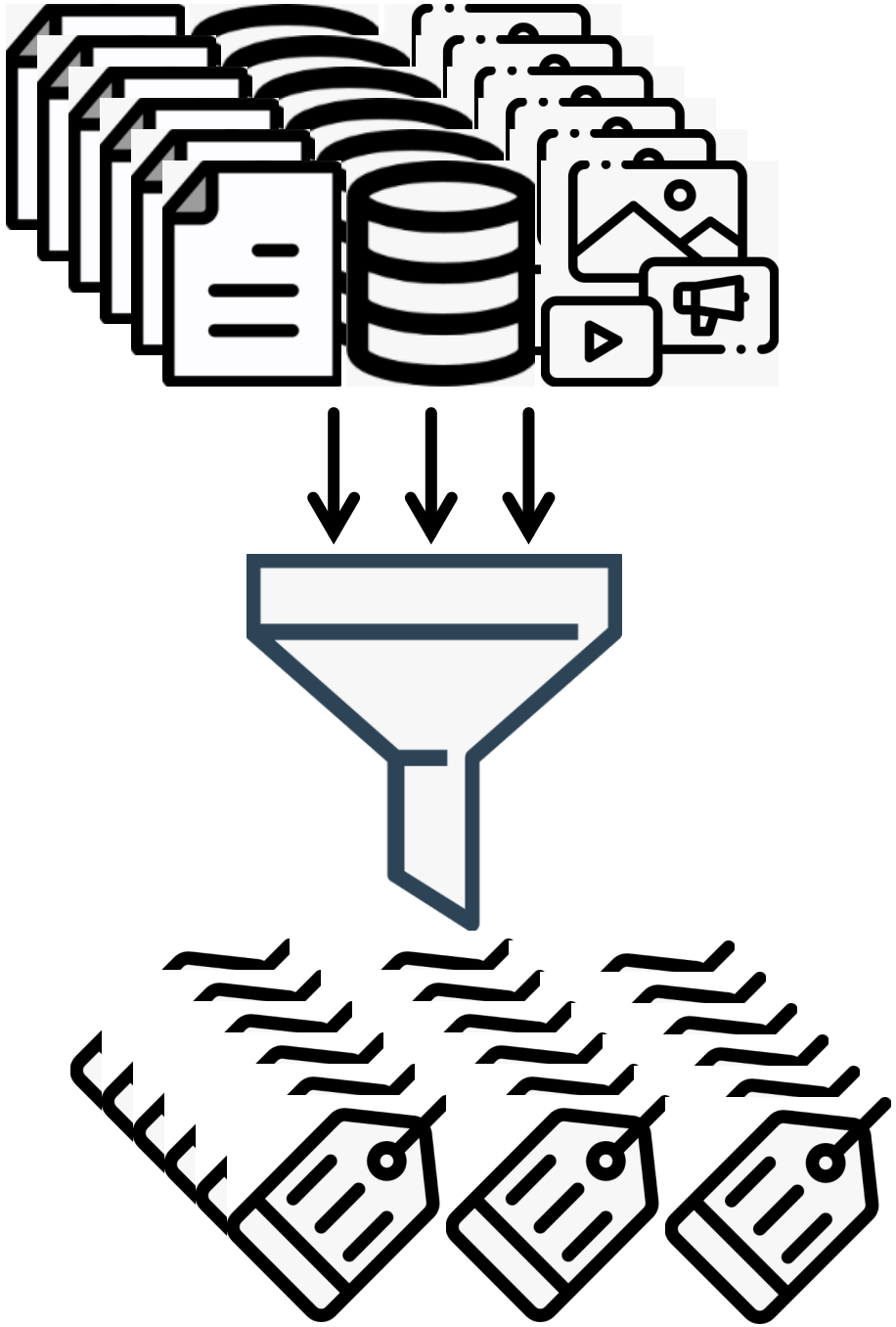
After Mark Butler, CTO, ONTOAi

Generate visualizations



Los Alamos Map of Knowledge

Tagging at scale



Web search engines, 1995-

Google

WWW invented by Tim Berners-Lee

Alta Vista launched

First Dublin Core Workshop

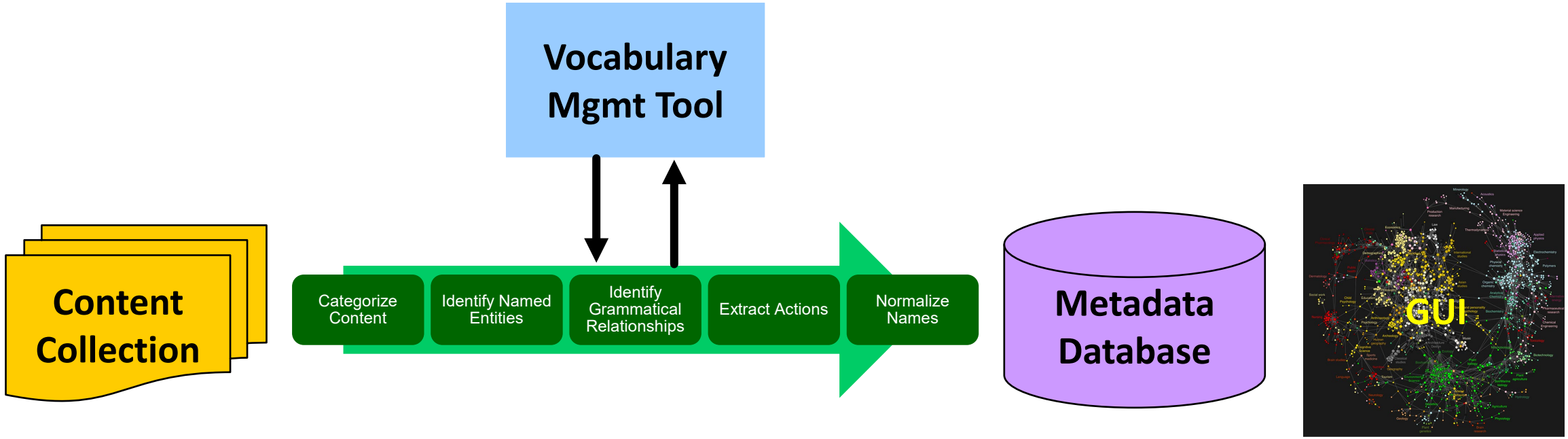
First NKOS Workshop

Google overtakes Alta Vista

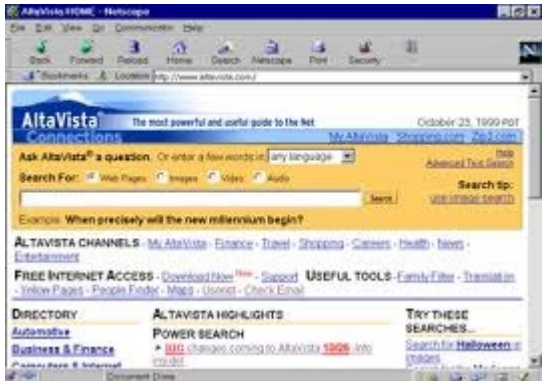
1990 1992 1994 1996 1998 2000 2010

What is metacode

- Unique identifiers (URIs) assigned to concepts in knowledge maps
- Vocabulary management tool for creating and managing metacodes
- Content tagging pipeline for assigning metacodes to content items
- Metadata database for storing metacodes associated with content objects
- GUI for viewing and browsing metacodes in knowledge maps



Datafusion/Metacode/Interwoven, 1998-2002



INTERWOVEN

Google

WWW invented by Tim Berners-Lee

Alta Vista launched
First Dublin Core Workshop

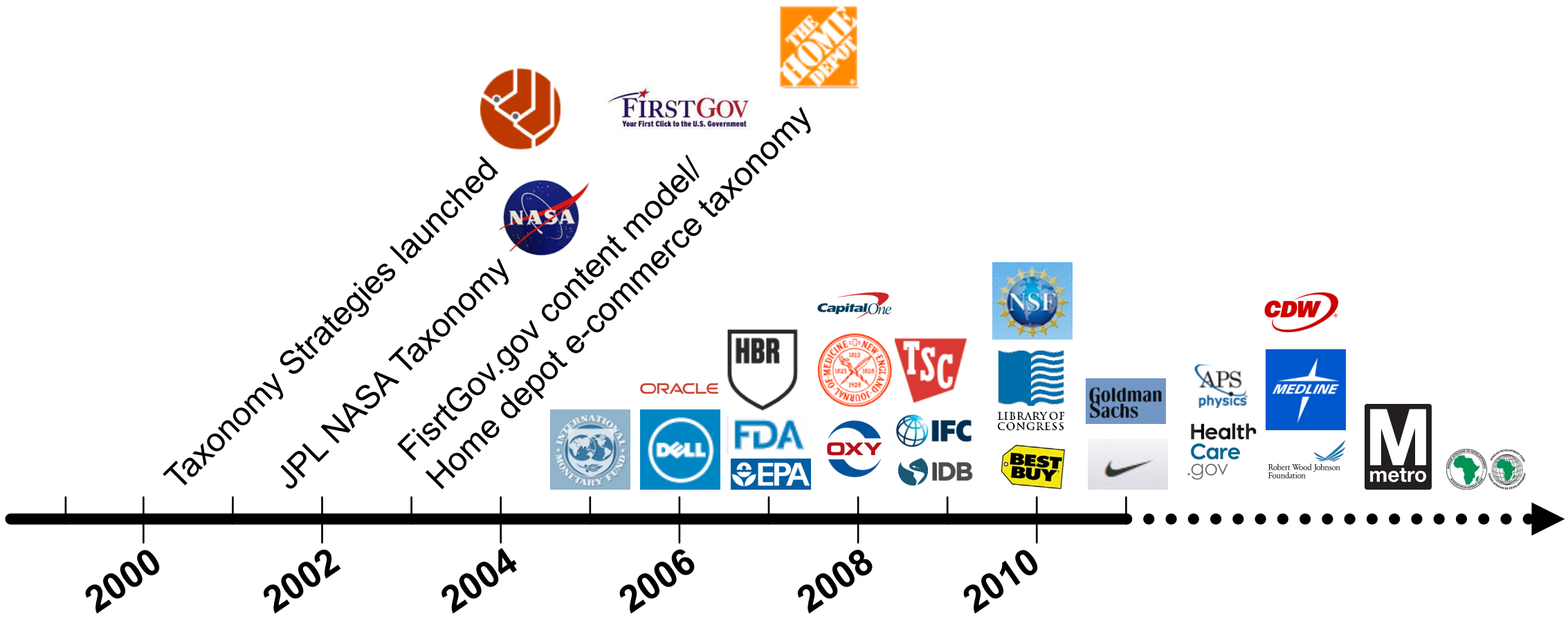
First NKOS Workshop

Google overtakes Alta Vista

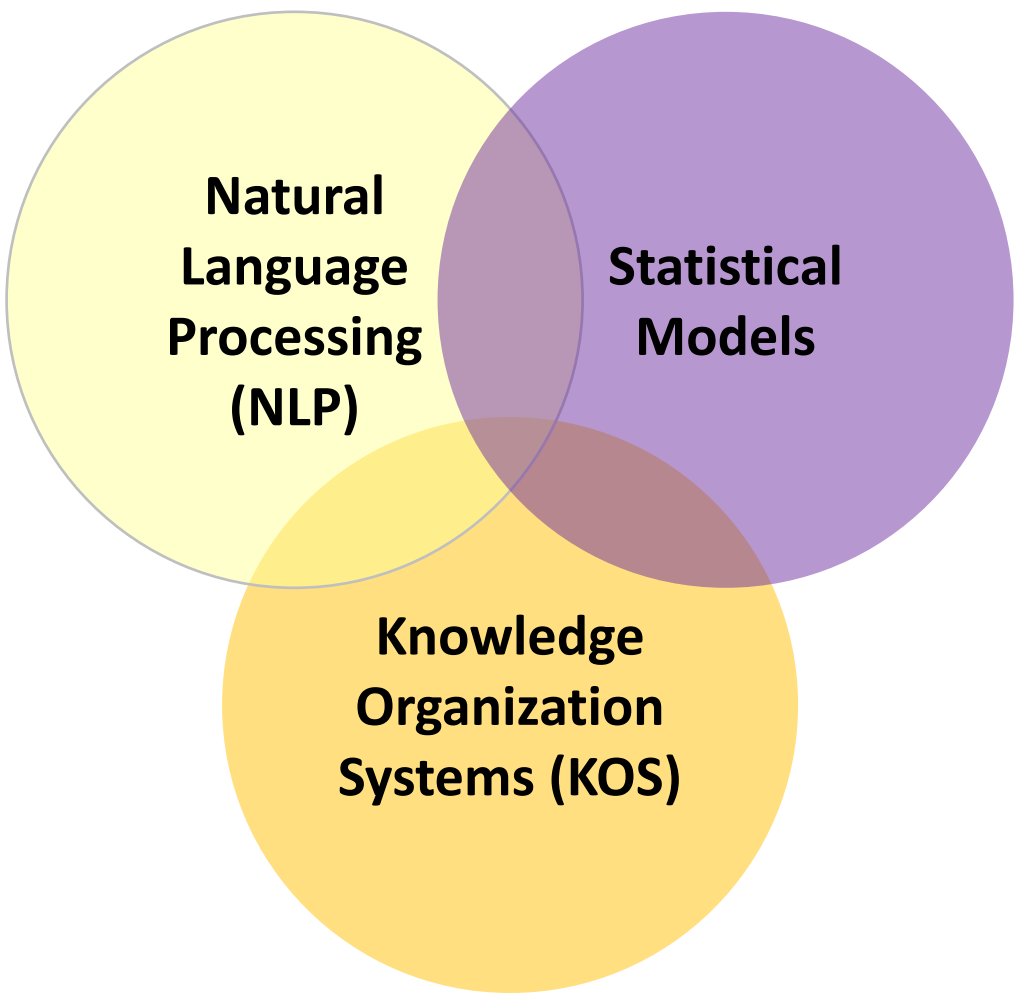
Interwoven acquires Metacode for $150M

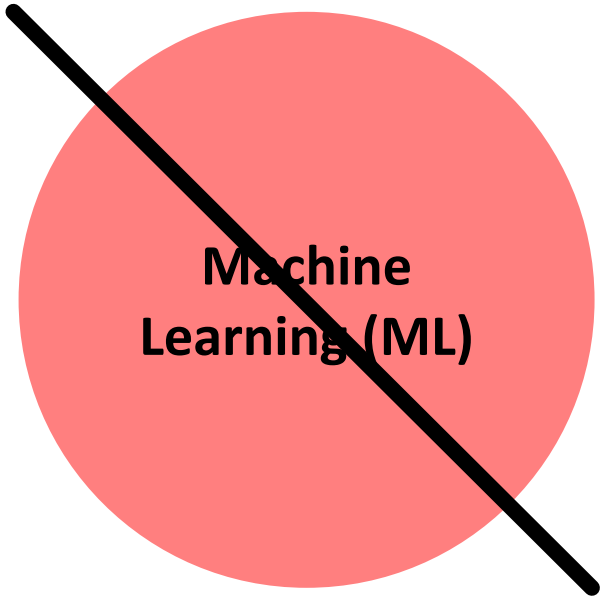
1990 1992 1994 1996 1998 2000 2010

Taxonomy Strategies, 2002-



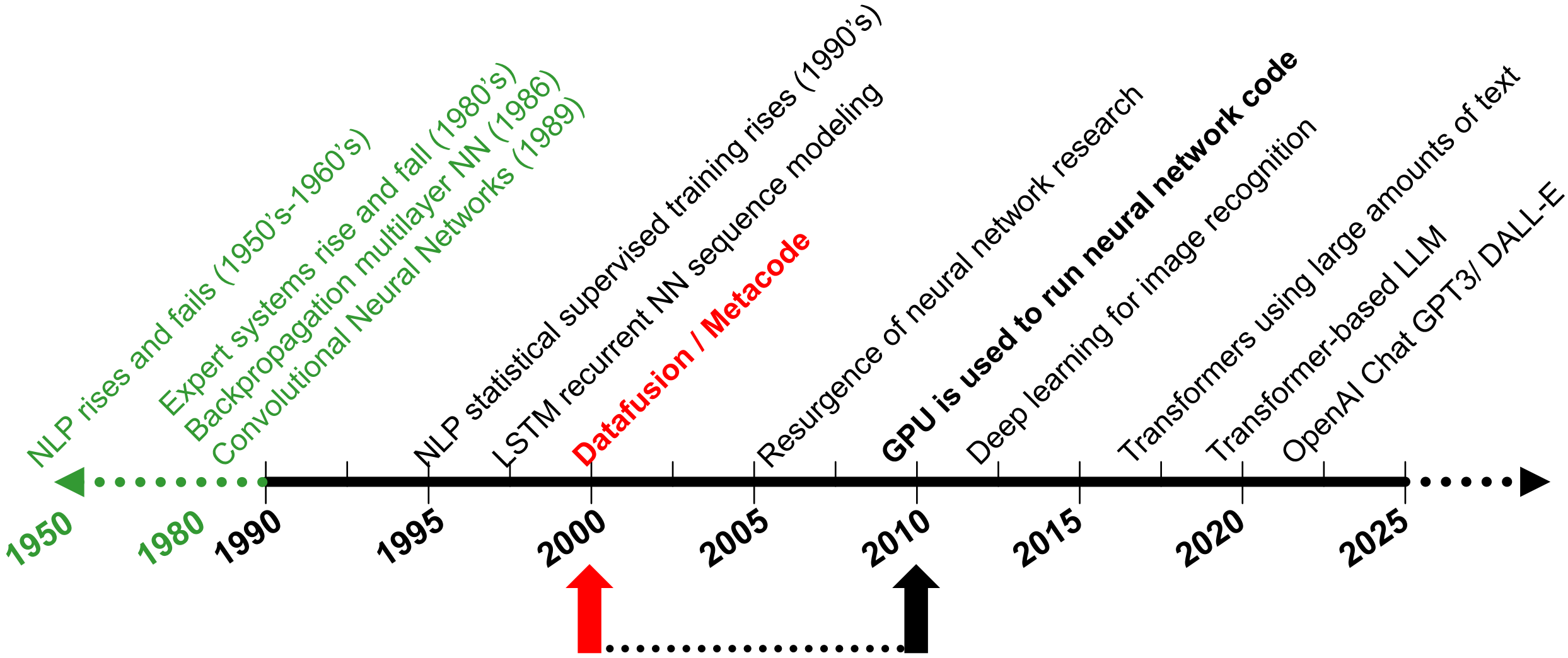
Datafusion components





AI Timeline

1956: John McCarthy coins the term “artificial intelligence” at Dartmouth meeting



Web search

- Keyword matching and ranking of web pages
- Natural language, Boolean, and semantic search
  - Query

LLM

- **Trained to analyze text and estimate the most probable word following a given sequence**
  - 13 trillion tokens
    - 1 token = 4 characters (about three quarters of a word)
  - 9.74 trillion words
  - 80 million books
- **Natural language user interface**
  - Prompt

Zero shot prompts

Who	What	Why	How
As a …	**I want to …**	**So that …**	**With these outputs …**
As a knowledge engineer,	I want to improve generative AI results using knowledge organization schemes	So that I can improve search results.	Give me some practical use cases.
As a task manager at a multilateral development organization,	I want a generic file plan for sovereign operating projects	So that I can file project documents and dispose or archive them when appropriate.	Format the file plan as a table. Include retention period and typical access rights.

Some prompting techniques

Technique	Description	Use Case
Chain of Thought	Use analogies, a table, contrasting examples, etc.	Simulated human reasoning
Self-Verification	Score multiple solutions	Quality assessment
Self-Consistency	Choose most consistent reasoning	Reduced errors and improved reliability
Decomposition	Break down complex problems	Modular problem-solving
Least-to-Most	Solve increasingly complex subproblems	Educational applications
Plan and Solve	Generate a structured plan	Project planning
Skeleton of Thought	Generate an outline	Writing, planning
Tree of Thought	Search through possible paths	Decision trees, exploration
Recursive Thought	Recursively solve sub-problems	Hierarchical problems

Source: Advanced Prompting Techniques Guide

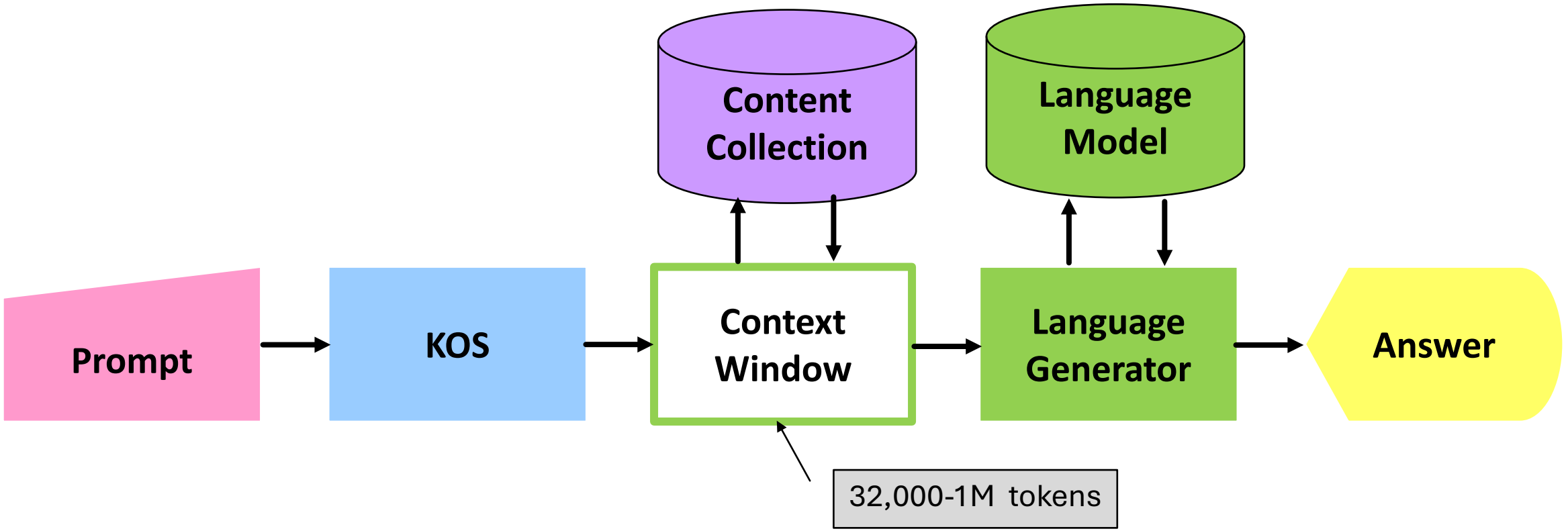
Everyone is fiddling with AI...
... but the results are unreliable.

What do you like about the AAT?

A: It helps me think.

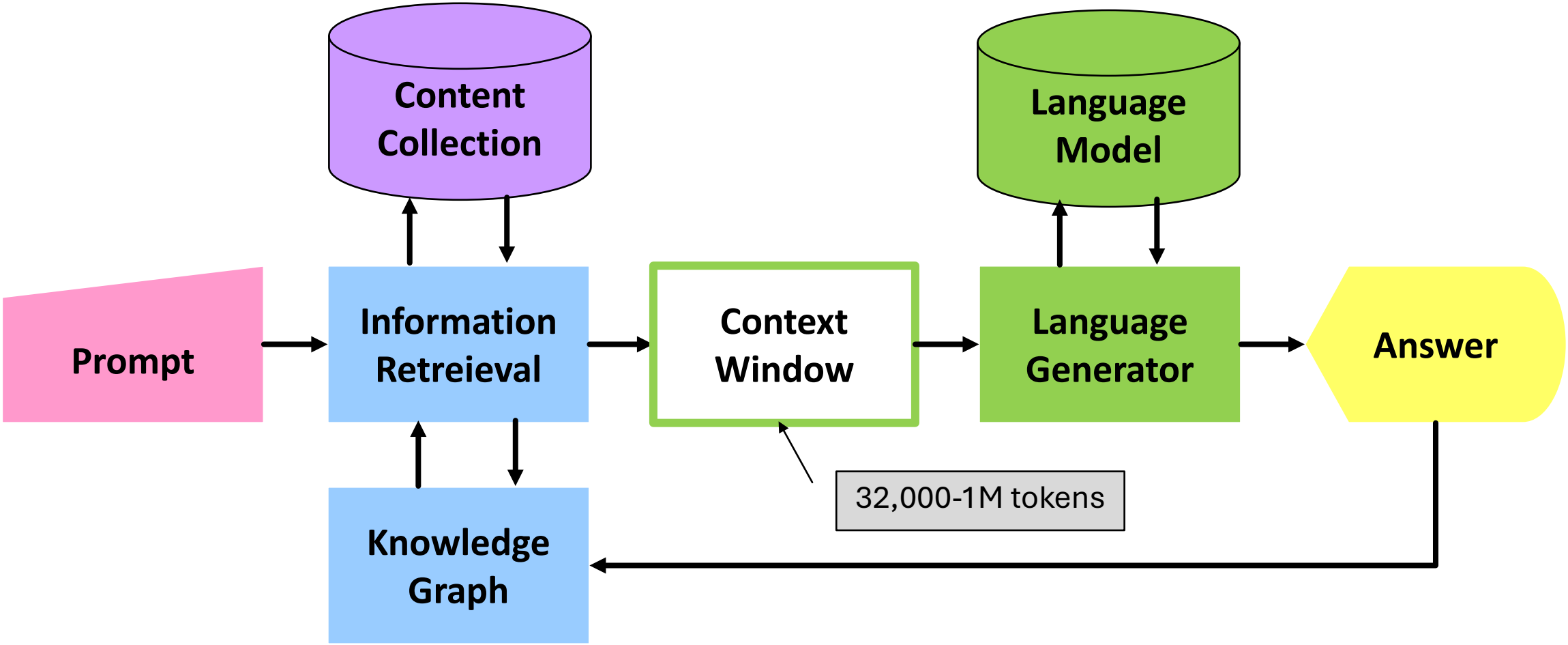
Integrating KOS in prompts using the context window

- Incorporate synonyms using a domain thesaurus
- Broaden or narrow the concepts using a hierarchical taxonomy
- Provide specific contexts using facets



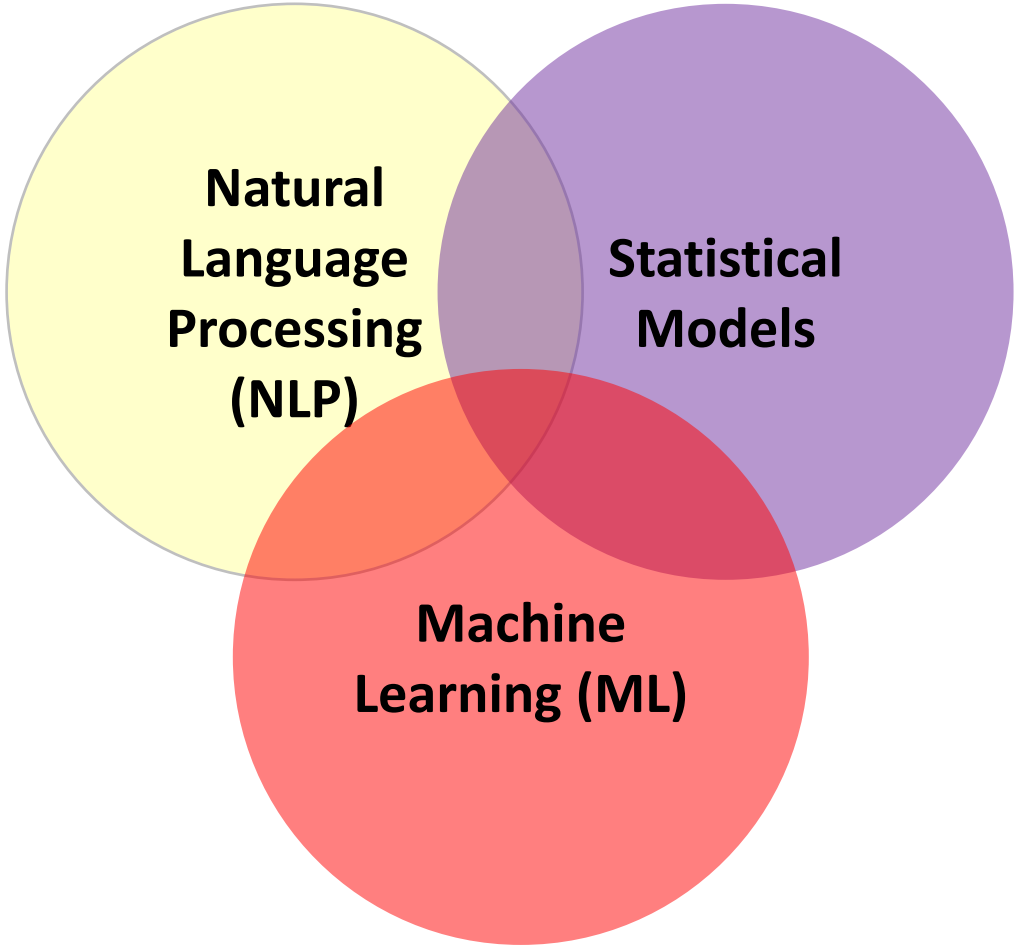
Retrieval Augmented Generation (RAG)

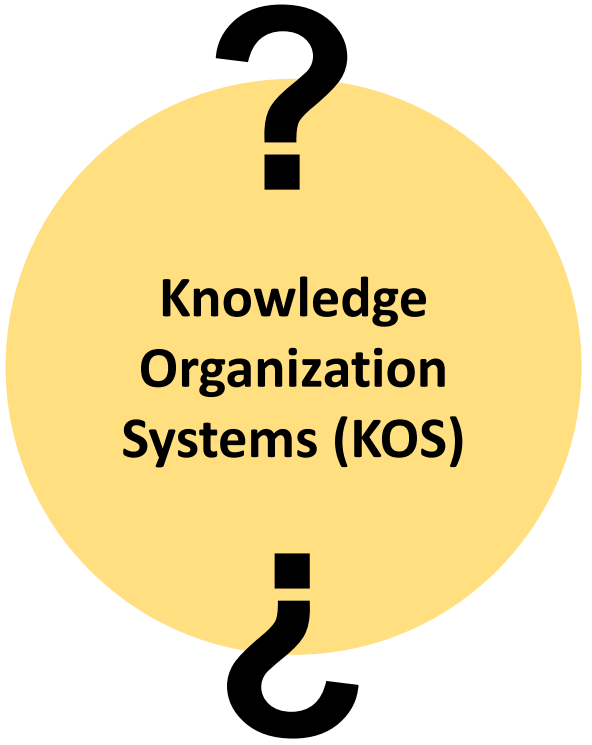
- Identify entities in the output, linking entities to knowledge graphs, and connecting models to knowledge bases
- Index source content to improve document recall and precision
- Annotate output to improve explainability and trustworthiness



The good thing about standards is there are so many to choose from.

AI components





Questions

Joseph Busch
jbusch@taxonomy strategies.com

Explore Related Presentations & Resources

https://taxonomystrategies.com/struggle-language-models-human-powered-instrument/