DATA FUSION 2025

Мечтают ли андроиды о большом и чистом эксплицитном знании?

Докладчик: к.т.н., доц. Алексей Незнанов

2025-04-16

О докладчике

- Алексей Андреевич Незнанов
  - К.т.н., доцент
    - Член *IEEE*, РАИИ, НБМЗ
  - *Senior Data Scientist* в ООО ТКШ
  - Доцент ФКН НИУ ВШЭ (http://cs.hse.ru/ai/)
  - Консультант «маленького гида по большим данным» на Постнауке (http://postnauka.ru/author/a_neznanov)
  - Автор учебника, учебных пособий, 11 авторских курсов и более 65 научных публикаций

Основные аббревиатуры

- ИИ – **И**скусственный **И**нтеллект
- МО – **М**ашинное **О**бучение

- БД – **Б**аза **Д**анных
  - СУБД – **С**истема **У**правления **БД**
- БЗ – **Б**аза **З**наний
  - СУБЗ – **С**истема **У**правления **БЗ**
- БЯМ – **Б**ольшая **Я**зыковая **М**одель

- ИП – **И**нтерфейс с **П**ользователем
- ПО – **П**рограммное **О**беспечение

Что определяло 2023 год?

1. Быстрое изменение ландшафта **специализированных средств высокопроизводительных вычислений**
2. Хайп вокруг больших языковых моделей (***LLM***)
   - Вовлечение большого числа пользователей инструментов МО из разнообразных прикладных областей
3. Дальнейший отрыв **лидеров рынка** как от «простых смертных», так и от «академии» в широком смысле
4. Учёт требований жизненного цикла моделей машинного обучения в *ETL*-процессах
   - Специализация хранилищ данных для МО (векторные СУБД и др.)
5. Всплеск интереса к качеству **логического вывода** и **гибридизации** средств статистического анализа с онтологическими средствами
6. Переоценка значимости **качества данных**, с учётом задач разметки данных для обучения
   - *ISO/IEC* 8183 (*Data life cycle framework*), *ISO/IEC DIS* 5259 (*Data quality for analytics and machine learning*), …
7. Накопление опыта выявления **дрейфа данных**

8. Давно ожидаемое добавление в корпоративные решения задач, связанных с этикой и сдвигом точки зрения искусственного интеллекта
9. Осознание зрелости некоторых классических решений и выстраивание системы референсных реализаций
10. Попытки применения инструментов искусственного интеллекта в проектировании и разработке ПО
11. Появление удобных компонентов интерфейса с пользователем для доступа к ИИ

Что определяло 2024 год?

1. На хайпе **бенчмаркинг**! Все всё сравнивают и не могут сравнить… Области Парето растут, критерии множатся
2. Ландшафт специализированных средств **высокопроизводительных вычислений** меняется всё быстрее
   - *Google Ironwood* (http://blog.google/products/google-cloud/ironwood-tpu-age-of-inference/), *Quantum Machine Learning* (*QML*), …
3. Хайп рассуждений больших языковых моделей, обернувшийся окончательной гибридизацией средств статистического анализа с онтологическими средствами, «рассуждение» рассуждению рознь…
4. ИИ-Агенты и ребрендинг *RPA* – первые экосистемы нового поколения, как следствие – корпоративный *prompt-engineering* в многоагентных системах
5. Попытка «простых смертных» и «академии» приблизиться к лидерам в генеративном ИИ
   - *DeepSeek* (http://github.com/deepseek-ai) и борьба за «открытость» с учётом пункта 2
6. *ETL*-процессы окончательно и обязательно учитывают ***MLOps***, включая разметку данных и фича-инжиниринг
7. Очередная **переоценка** значимости **качества данных**, с учётом проблем использования синтетических данных и свойств искусственности/синтетичности генеративного ИИ
   - *Data-Efficient LLMs*
8. Переворот (разворот на 180!) в исследованием **этики**, **сдвигов** точки зрения и **токсичности** ИИ
   - Выявление связи цензуры и качества – все переругались…
9. Накопление опыта применения ИИ в **проектировании и разработке ПО**
   - “*vibe coding*”, *cursor.ai* как имя нарицательное, *NVIDIA OpenCodeReasoning* (http://huggingface.co/datasets/nvidia/OpenCodeReasoning)
10. Стандартизация удобных компонентов интерфейса с пользователем и *huggingface*-экосистема
    - *Hugging Face Spaces* (http://huggingface.co/spaces)

Практика RAG и эксплицитное знание

- ✓ Смысл
- ✓ Устойчивость и качество традиционных решений
- ✓ Классификация
- ✓ Что смогли БЯМ в этом году в смысле «рассуждений»
- ✓ Семантический поиск
- ✓ Онтологии и семантическое обогащение
- ✓ Инструменты

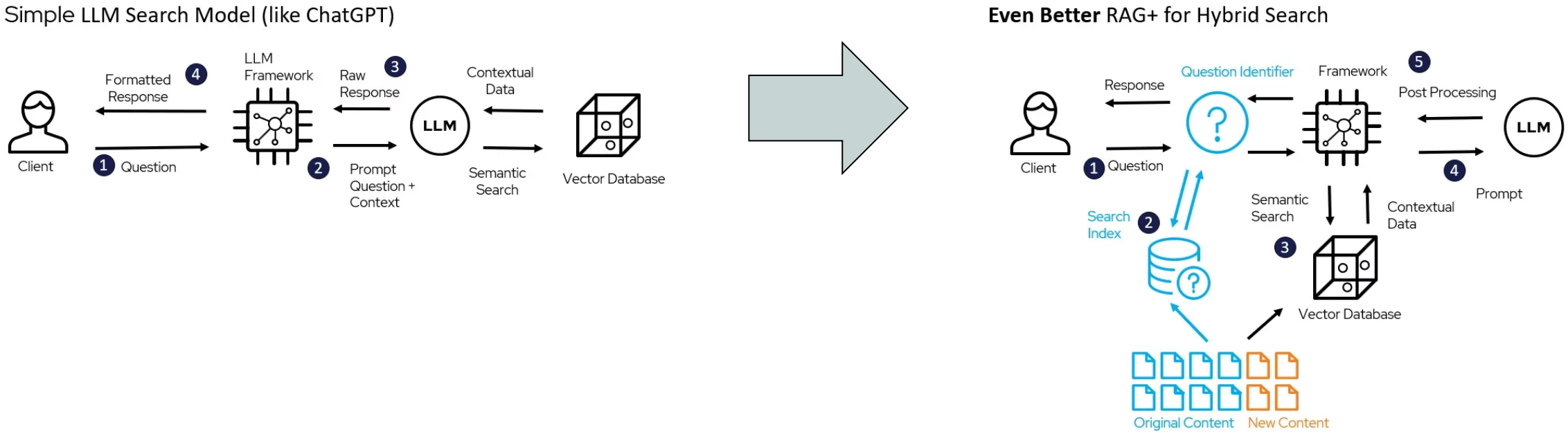
RAG как концепция и её обобщение

- *RAG* – это не только «лоскут ветоши» (см. *google*-перевод), но и три-четыре методологии генерации, дополненной извлечением данных и знаний из внешних источников этих самых данных и знаний для улучшения работы БЯМ

- Акценты:
  1. **Не хочется менять базовую БЯМ**
  2. Хочется минималистичную надстройку над существующими хранилищами
  3. Хорошо бы повысить качество «поиска» (релевантность)
  4. Ну и учесть динамику

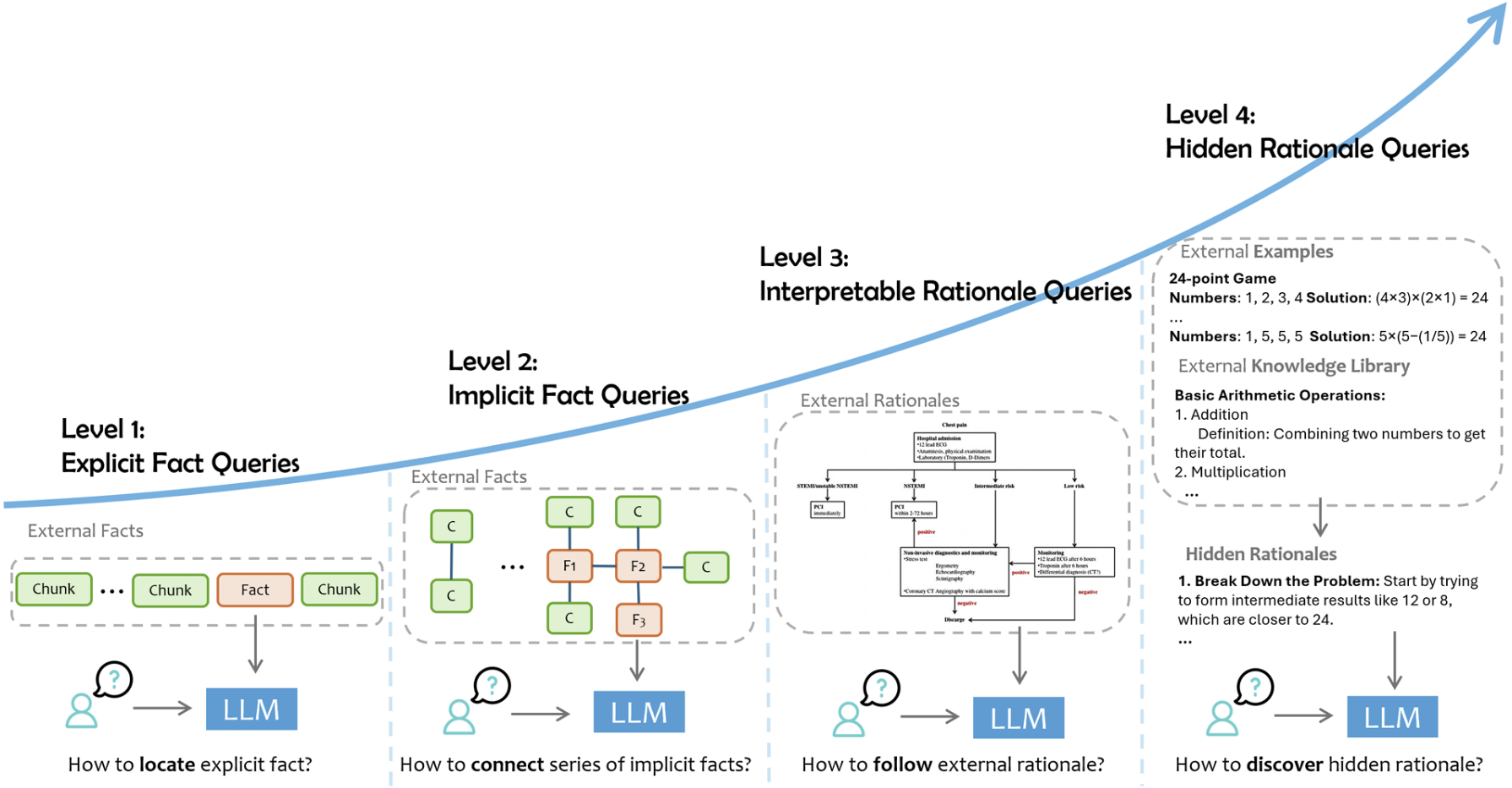
RAG и семантический поиск

- *Retrieval-Augmented Generation* (***RAG***) – подмешивание ко входу БЯМ проверенных данных в текстовой или мультимодальной форме
  - Для нормальной работы требуется дообучение БЯМ, БД и семантический поиск
    - *What is Retrieval Augmented Generation* (*RAG*)? (http://pureinsights.com/blog/2023/what-is-retrieval-augmented-generation-rag/)



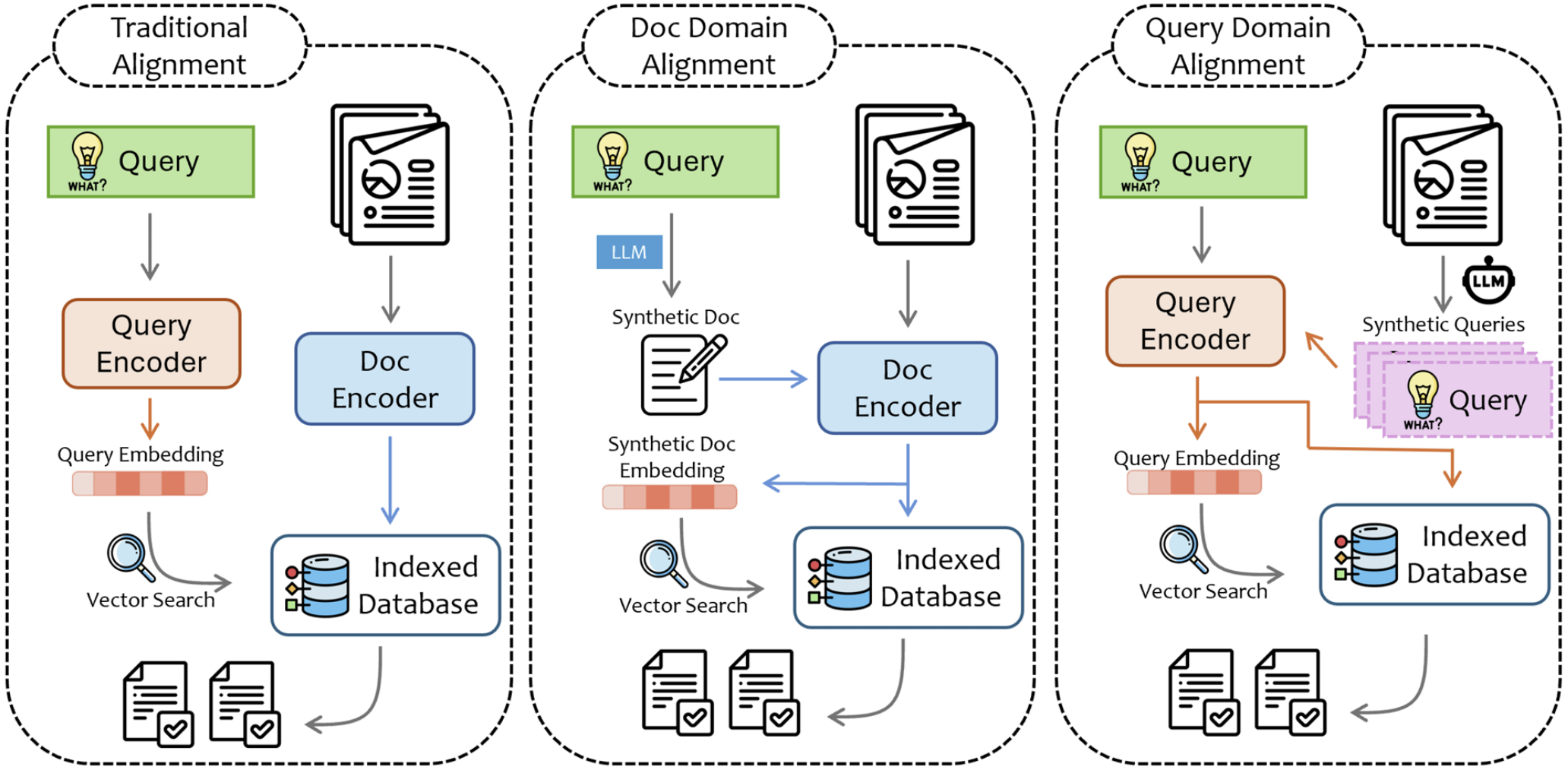
Обзор *RAG* и чуть шире (1)

- *Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely*. 2024 (http://arxiv.org/abs/2409.14924)
  - *Four levels of queries*:



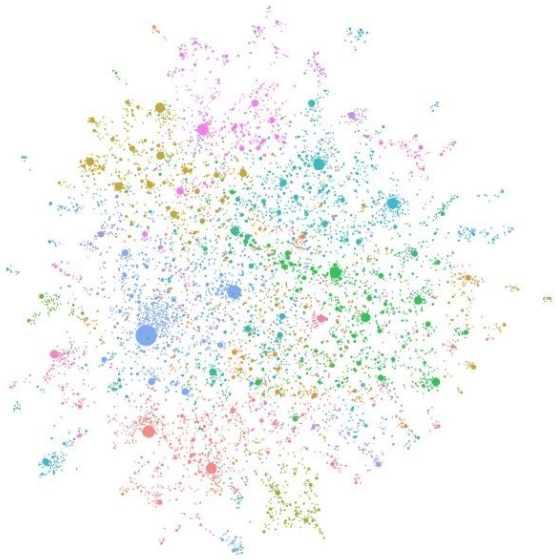
Обзор *RAG* и чуть шире (2)

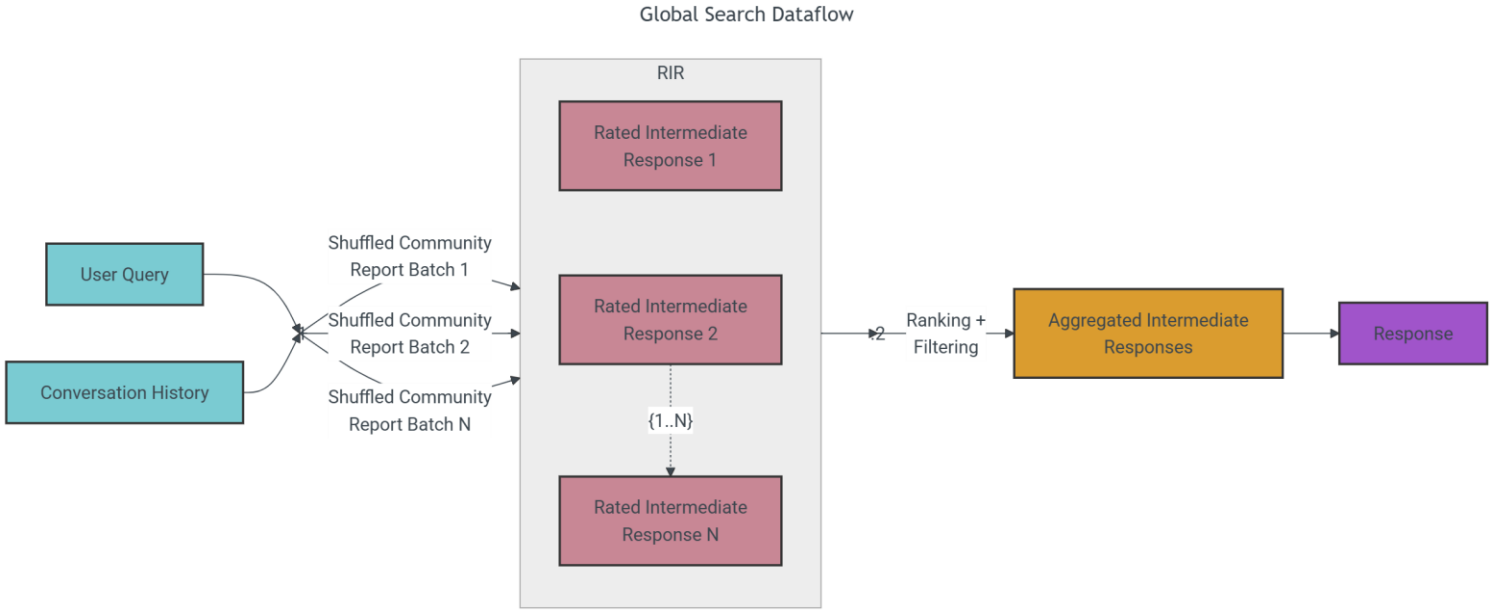
- Three Types of Query-Document Alignment:



Microsoft GraphRAG

- *The Future of AI: GraphRAG – A better way to query interlinked documents* (http://techcommunity.microsoft.com/blog/aiplatformblog/the-future-of-ai-graphrag-%E2%80%93-a-better-way-to-query-interlinked-documents/4287182)
  - *GraphRAG docs* (http://microsoft.github.io/graphrag/)
  - *GraphRAG repository* (http://github.com/microsoft/graphrag?ysclid=m6bwexmusj863175340)
    - *GraphRAG is a technique that combines the power of knowledge graphs and large language models (LLMs) to improve the accuracy and relevance of responses to user queries*





Несемантический поиск – *Lucene*

- *Apache **Lucene*** (http://lucene.apache.org)
  - Стандарт де-факто!

- *Lucene **query syntax***
  - *Apache Lucene – Query Parser Syntax* (http://lucene.apache.org/core/2_9_4/queryparsersyntax.html)
  - *Lucene query syntax in Azure AI Search* (http://learn.microsoft.com/en-us/azure/search/query-lucene-syntax)

Lucene Query Syntax Cheat Sheet

Keyword Matching:	
`title: foo`	Search for the word “foo” in the title field.
`title: "foo bar"`	Search for the phrase “foo bar” in the title field.
`title:foo -title:bar`	Search for word “foo” and not “bar” in the title field.

Wildcard Matching:	
`title:foo*`	Search for any word that starts with “foo” in the title field.
`title:foo*bar`	Search for any word that starts with “foo” and ends with bar in the title field.

Proximity Matching:	
`"foo bar"~4`	Search for “foo bar” within 4 words from each other.

Range Searches:	
`mod_date: [20020101 TO 20030101]`	To match documents whose field(s) values are between lower/upper bound.

Boosts:	
`(title:foo OR title:bar) ^1.5 (body:foo OR body:bar)`	The higher the boost factor, the more relevant the term will be.

Boolean Operators:	
`NOT type: mysql`	To search for all transactions except MySQL transactions.
`mysql.method: SELECT AND mysql.size: [10000 TO *]`	To search for all MySQL SELECT queries with large attachments.
`(mysql.method: INSERT OR mysql.method: UPDATE) AND responsetime:[30 TO *]`	Use the DISTINCT clause to get back only unique field values.

Fuzzy Searches:	
`roam~`	Lucene supports fuzzy searches based on the Levenshtein Distance.
`roam~0.8`	Starting with Lucene 1.9 an additional parameter can specify required similarity.

Grouping	
`(jakarta OR apache) AND website`	To search for either “jakarta” or “apache” and “website” use the query.
`title:(+return +"pink panther")`	Search for a title that contains “return”

Log	
`normalMsgDate`	First Log Date
`normalDateMax`	Last Log Date
`count`	Log Count
`normalDate`	Log Date
`logSourceName`	Log Source
`entityName`	Log Source Entity
`logSourceHostName`	Log Source Host
`logMessage`	Log Message
`logSourceTypeName`	Log Source Type
`sequenceNumber`	Log Sequence Number

Traffic	
`kBytesIn`	Host (Impacted) KBytes Received.
`kBytesOut`	Host (Impacted) KBytes Sent.
`impactedHostTotalKBytes`	Host (Impacted) KBytes Total.
`itemsPacketsIn`	Host (Impacted) Packets Received.
`itemsPacketsOut`	Host (Impacted) Packets Sent.
`impactedHostTotalPackets`	Host (Impacted) Packets Total.
`kBytes`	KBytes Inbound.
`outboundKBytes`	KBytes Outbound.

Applications	
`action`	Action.
`amount`	Amount.
`command`	Command.
`parentProcessId`	Parent Process ID.
`parentProcessPath`	Parent Process Path.
`process`	Process Name.
`processId`	Process ID
`serviceName`	Known Application

Created by logit.io

«Семантический поиск» – *Elastic search*

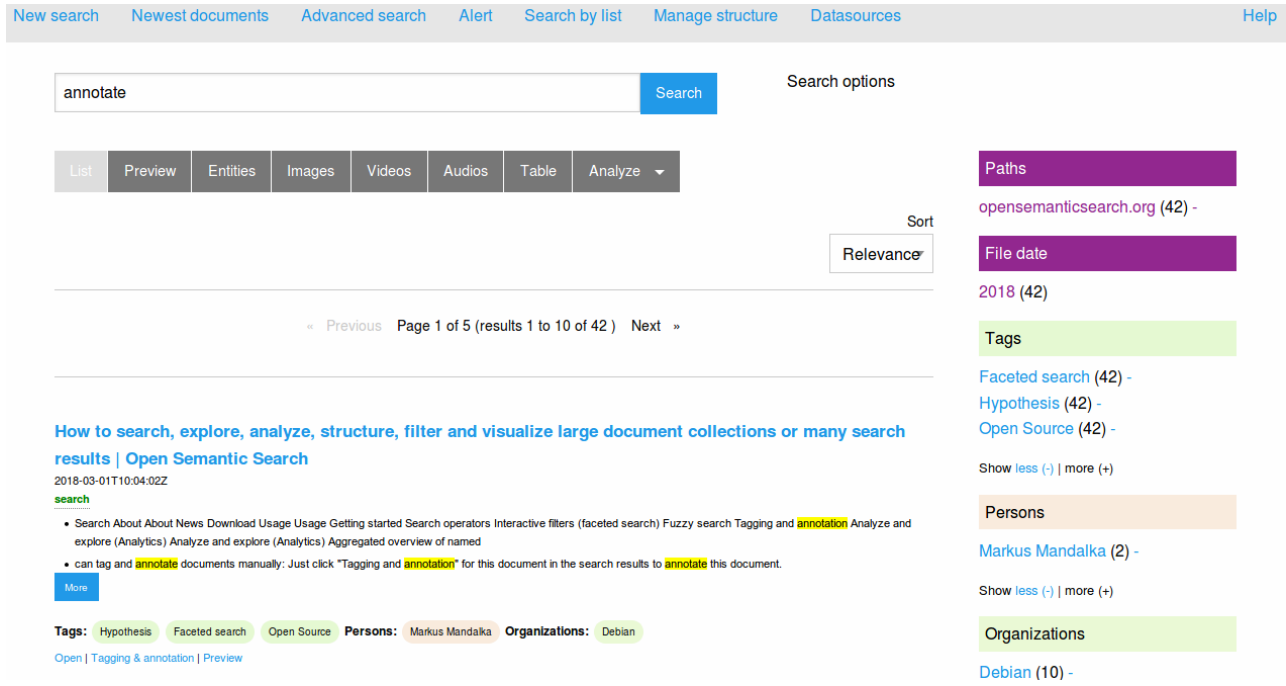
- ***Elasticsearch*** (http://www.elastic.co/elasticsearch)
  - The most widely deployed vector database?
  - The heart of the Elastic Stack

«Семантический поиск» – *Opensearch*

- ***Opensearch*** (http://opensearch.org)
  - *Open-source, enterprise-grade search and observability suite that brings order to unstructured data at scale*
  - *Mostly compatible with Elasticsearch OSS*
  - *Query domain-specific language* (http://docs.opensearch.org/docs/latest/query-dsl/)
  - *Top Tips for Querying OpenSearch* (http://logit.io/blog/post/top-tips-for-querying-opensearch/)

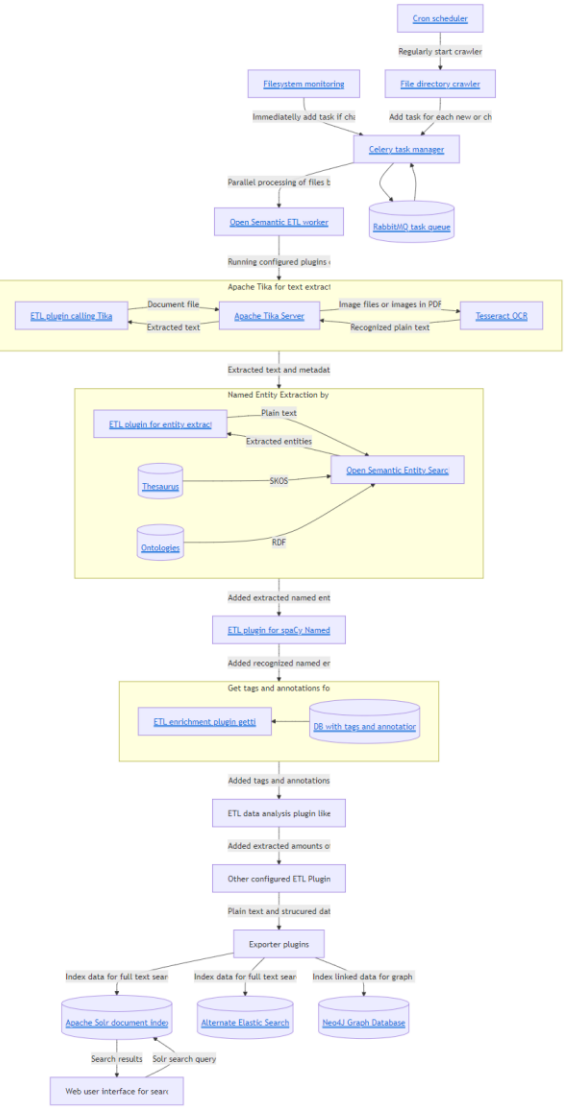
«Семантический поиск» – *Open Semantic Search*

- ***Open Semantic Search*** *(OSS)* (http://opensemanticsearch.org)
  - *Open-Source Search Engine with Apache Lucene / Solr*
  - *Flowchart of document processing and data flow* (http://github.com/opensemanticsearch/open-semantic-search/blob/master/docs/doc/modules/README.md)



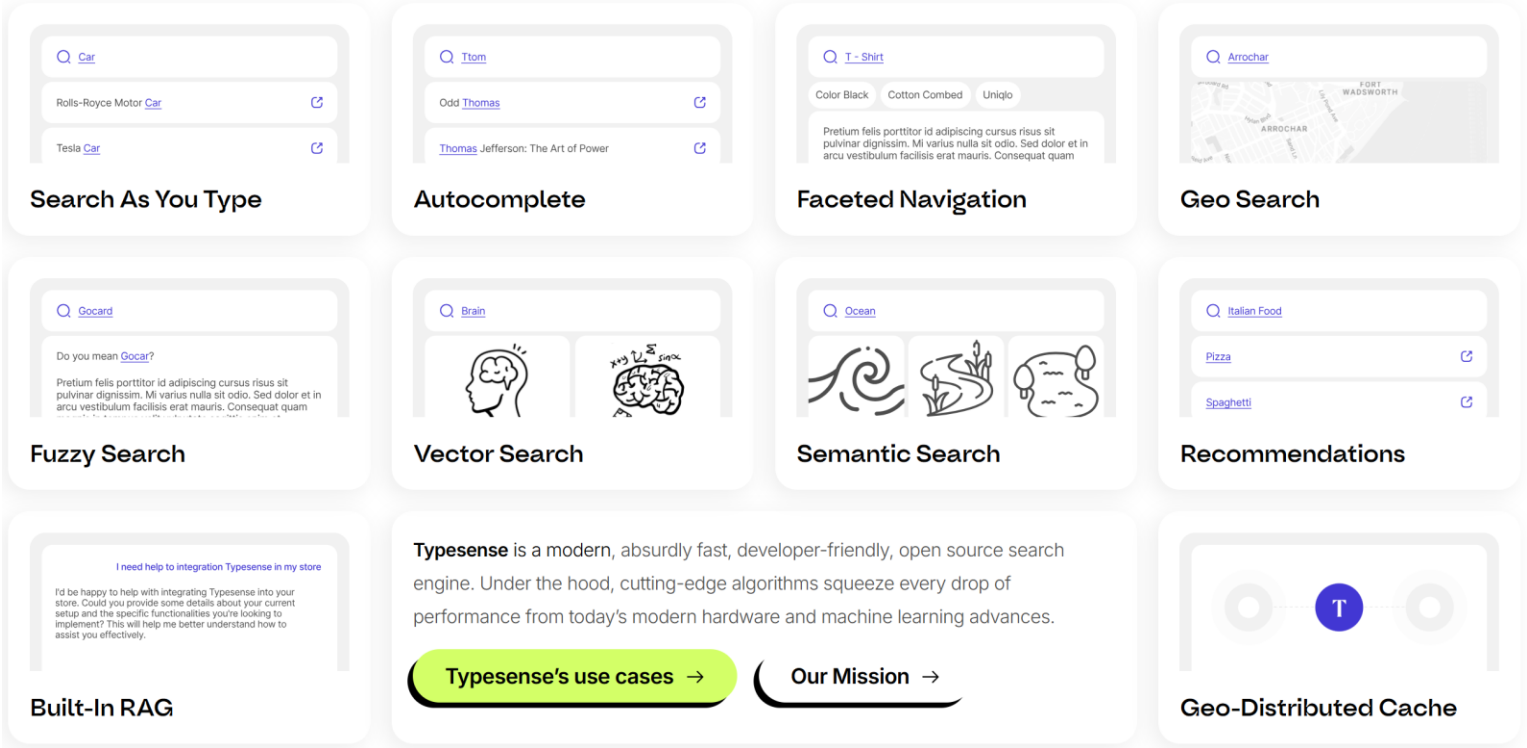
OSS: архитектура и компоненты

- User Interface: Client and user interface
  - **Search query forms**: Search query form for full text search
  - **Explorer and navigator**: Search with full text search and navigate (exploratory search) the index or search results with **interactive filters (facets)**
    - **Viewers**: Parts of the UI to show different views
    - Code: /solr-php-ui/templates/
  - **Annotators**: Web Apps for tagging documents or CMS with forms and fields to manage meta data like tags or annotations
  - **Search Apps**: Applications and user interfaces for search like search with lists tool or named entities manager
- **Index and search server (Solr or Elastic Search)**
  - Datamodel/Schema: src/solr.deb/var/solr/data/opensemanticsearch/conf/managed-schema
  - Storage: /var/solr/data
  - Log: /var/solr/logs/
- Open Semantic ETL: Framework for data integration, data analysis, data enrichment and ETL
  - Connectors, importers, ingestors or crawlers: Import data from a data sources
  - **Parsers:** *Apache Tika* to extract text and metadata from different file formats and document formats
  - **Entity extraction and entity linking**: Open Semantic Entity Search API
  - Data enrichment plugins and enhancer: Enhancing content with additional data like meta data (i.e. tagging or annotations) or analytics (i.e. OCR)
  - **ETL Exporter** or Loader for Solr or Elastic Search: Indexing the data to search index
- **Trigger**: Your CMS or your file system will notify the web service (API) when there is new data or when content changed
- Web services (REST-API): Available via standard network protocol HTTP and waiting until you or another service demands actions like crawling a directory or a webpage and starting this actions
- Queue manager (Celery on RabbitMQ): Managing task queue and starting of text extraction, analysis, data enrichment and indexing jobs by the right balance of parallel workers
- Scheduler: Managing starting of scheduled indexing jobs



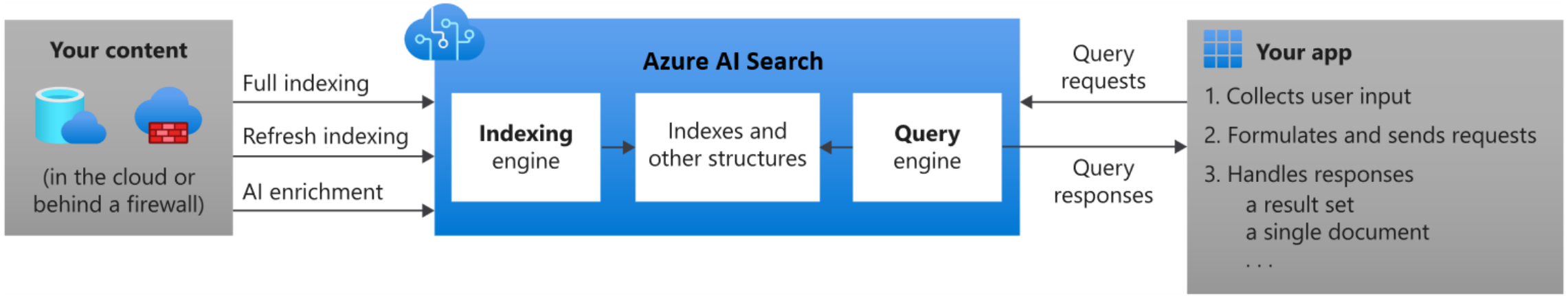
«Семантический поиск» – *Typesense*

- ***Typesense*** (http://typesense.org)
  - *Modern, privacy-friendly, open source search engine built from the ground up using cutting-edge search algorithms*

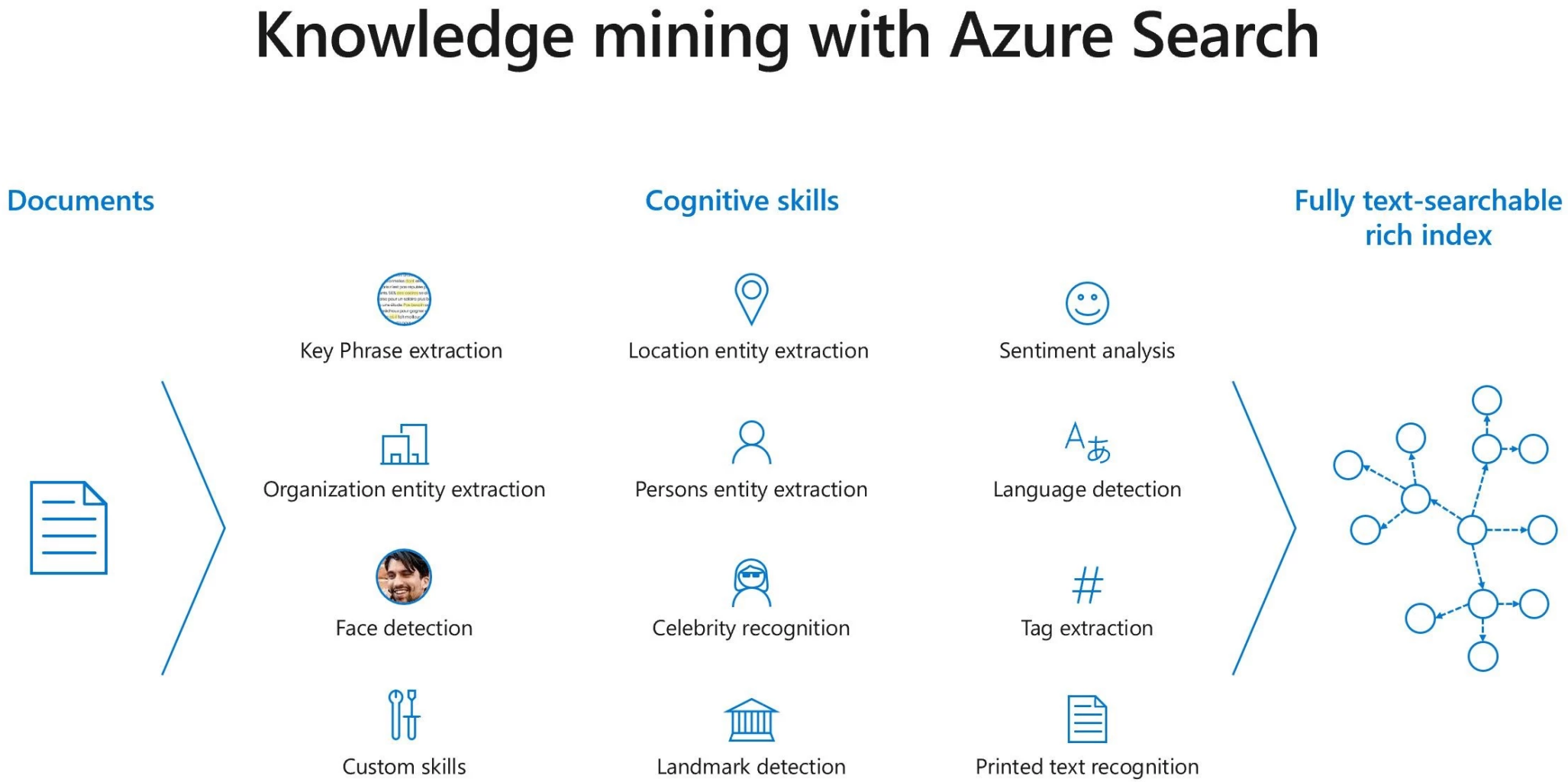


Платформа *Azure AI (Cognitive) Search*

- *Microsoft Azure **Cognitive Search** → Azure **AI Search***
  (http://azure.microsoft.com/en-us/products/ai-services/ai-search/)
  - *“Deliver better AI experiences with a knowledge retrieval system built for advanced retrieval augmented generation (RAG) and modern search”*
  - *Uses Azure AI Foundry and AI Data enrichment components*

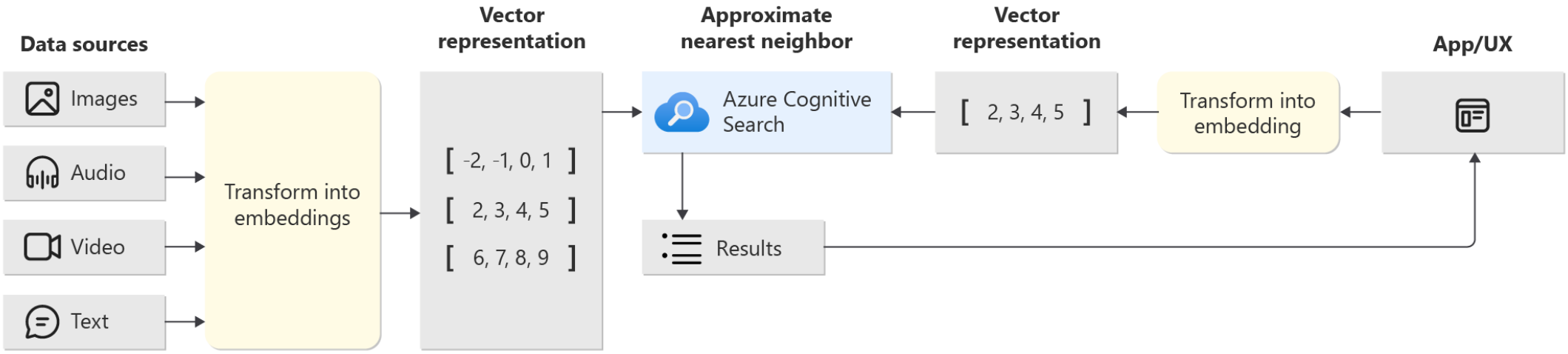


Исходная идея *Microsoft AI Search*



«Вектора» в *Azure AI Search*

- *Vectors in Azure AI Search* (http://learn.microsoft.com/en-ca/Azure/search/vector-search-overview)
  - *Supports hybrid multilanguage search*
  - *Stores the data that you query over in pure vector store – use it any time you need long-term memory or a knowledge base, or grounding data for Retrieval Augmented Generation (RAG) architecture*



Онтологически-контролируемый поиск

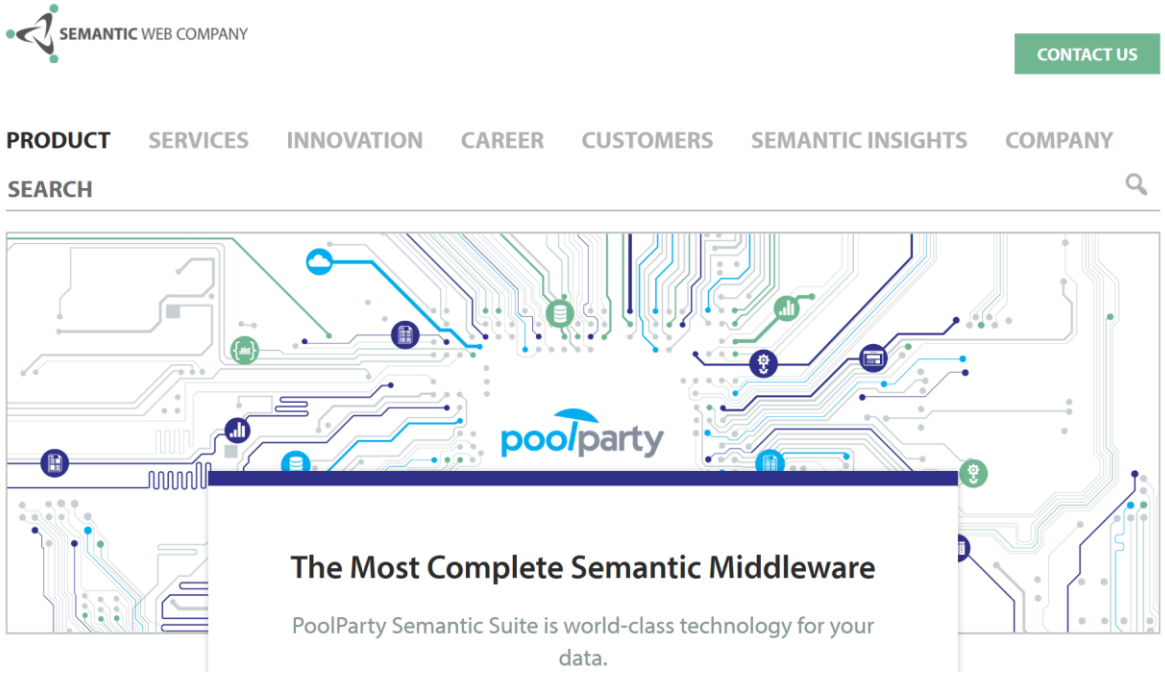
- Давняя идея с многочисленными подводными камнями
  - Главная проблема стоимости как создания, так и внедрения
- Много вариантов предложено в 2010-е, а сейчас переделывается с применением БЯМ
  - Иллюстративное студенческое:
    - *Ontology Based Meaningful Search Using Semantic Web and Natural Language Processing Techniques*. 2013 (http://ictactjournals.in/paper/IJSC_Vol4Iss1_P6_662_666.pdf)
  - Из свежего:
    - *FENIX: A Semantic Search Engine Based on an Ontology and a Model Trained with Machine Learning to Support Research*. 2021 (http://www.aircconline.com/csit/papers/vol11/csit110709.pdf)

«Семантический поиск» – *uNaice*

- ***uNaice*** *Semantic Search* (http://unaice.com/semantic-search-engine/)
  - *The product search function that understands what your customers are looking for*
  - *Explicit ontology*

Корпоративные *KMS – PoolParty*

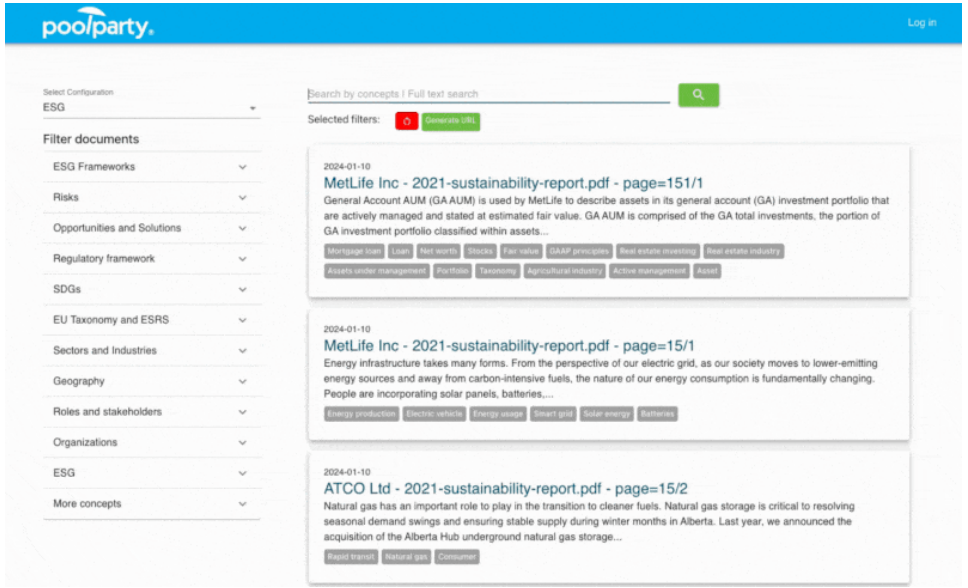
- *Semantic Web Company* ***PoolParty Semantic Suite*** (http://semantic-web.com/poolparty-semantic-suite/)
  - *How PoolParty and ISO 25964 fit together* (http://semantic-web.com/**2016**/03/04/how-poolparty-and-iso-25964-fit-together/)



PoolParty Semantic Search

- За счёт чего:
  - *PoolParty Taxonomy Manager + Ontology Management*
    - *Combine the power of graph databases and SPARQL engines*
  - *PoolParty Extractor*
    - *Transform and harmonize documents/data into RDF*
  - *PoolParty Recommendor*
  - *PoolParty GraphSearch API*

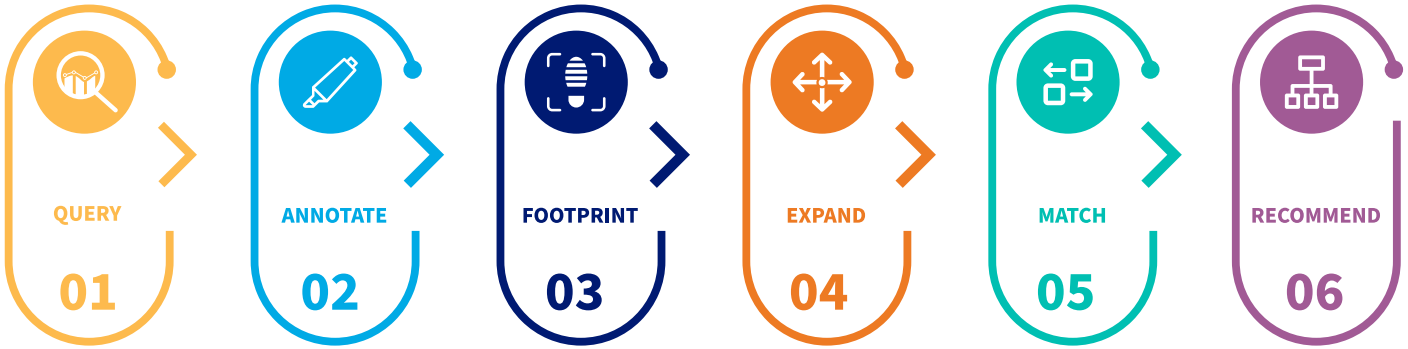
- Реально семантические фацеты!
  - *Filter and narrow down search queries with facets that are based on a taxonomy*

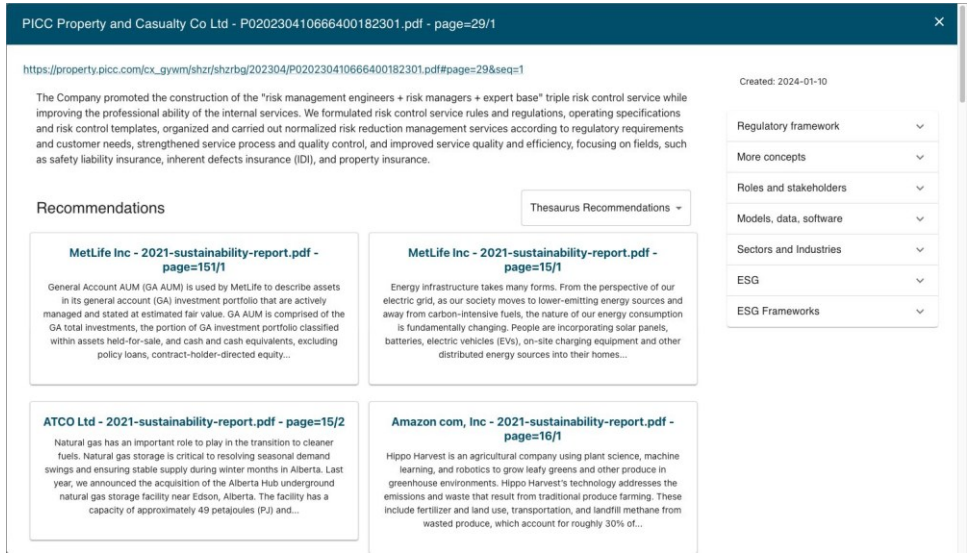


PoolParty Semantic Recommender

poolparty
A Graphwise Company

- *Recommender Systems* (http://www.poolparty.biz/recommender-systems)
  - *Traceable recommendation workflow of the PoolParty recommendation system*

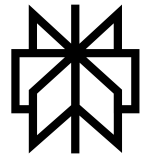




LLM-based Research Agents

- *OpenAI Deep research* (http://openai.com/index/introducing-deep-research/)

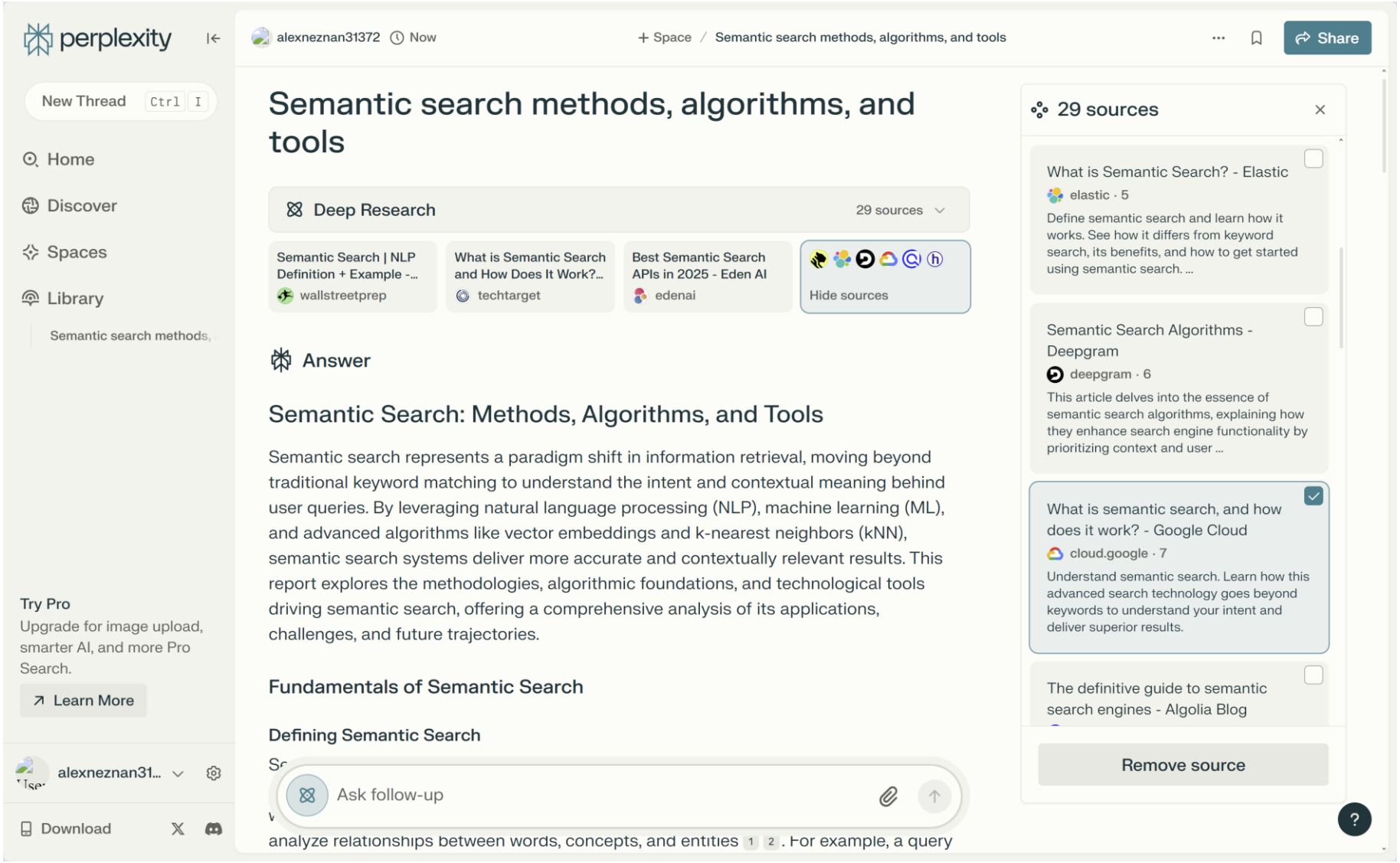
- *Perplexity* (http://www.perplexity.ai)



- …

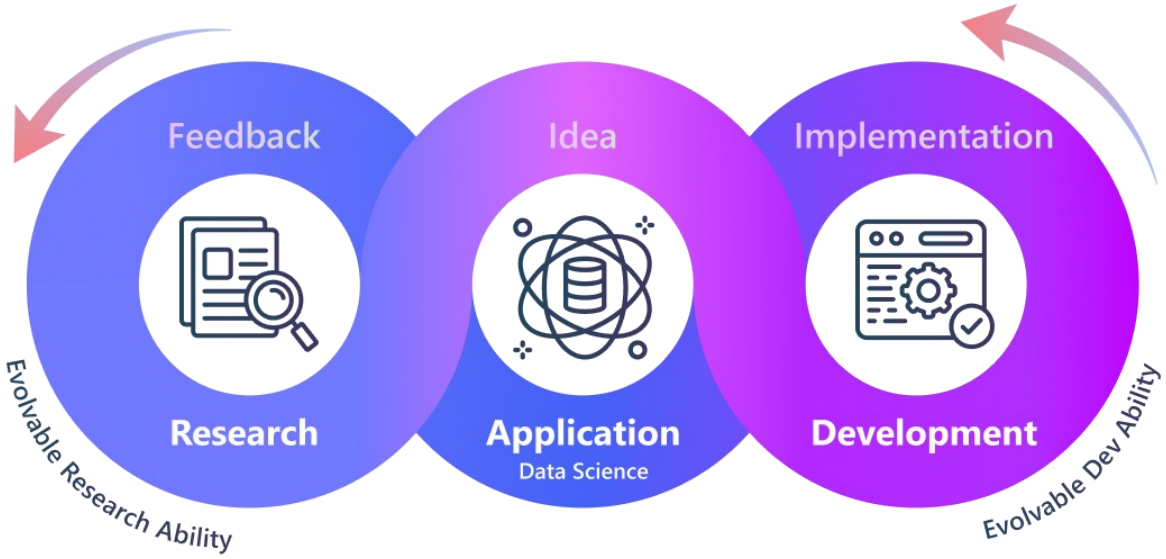
- Вменяемое управление источниками!

Пример *Deep Research* в *Perplexity*



Microsoft RDAgent

- *Microsoft **RDAgent*** (http://github.com/microsoft/RD-Agent)
  - *Data Science Agent aims to automate the most critical and valuable aspects of the industrial R&D process*
  - *The advent of RDAgent can be your Data Mining Agent!*
    - *Iteratively proposing data & models and implementing them by gaining knowledge from data*

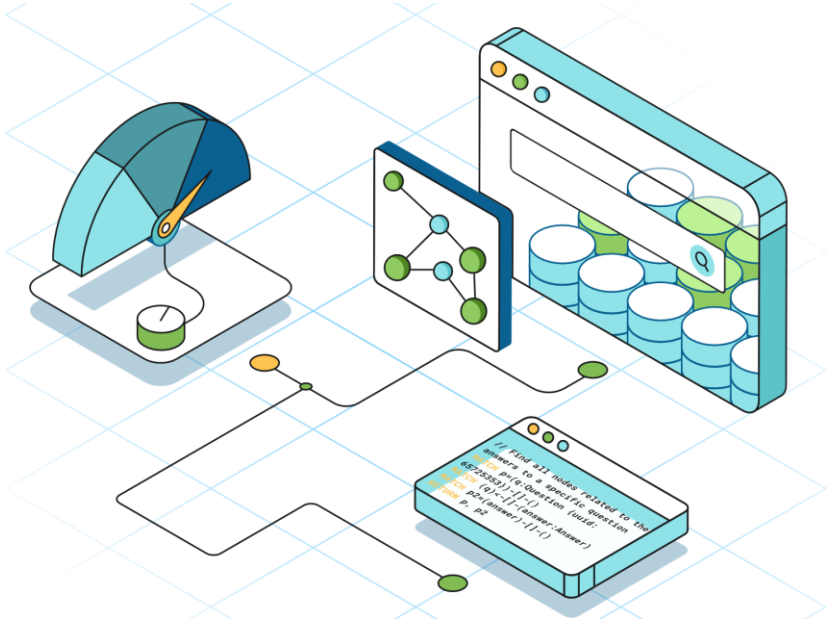


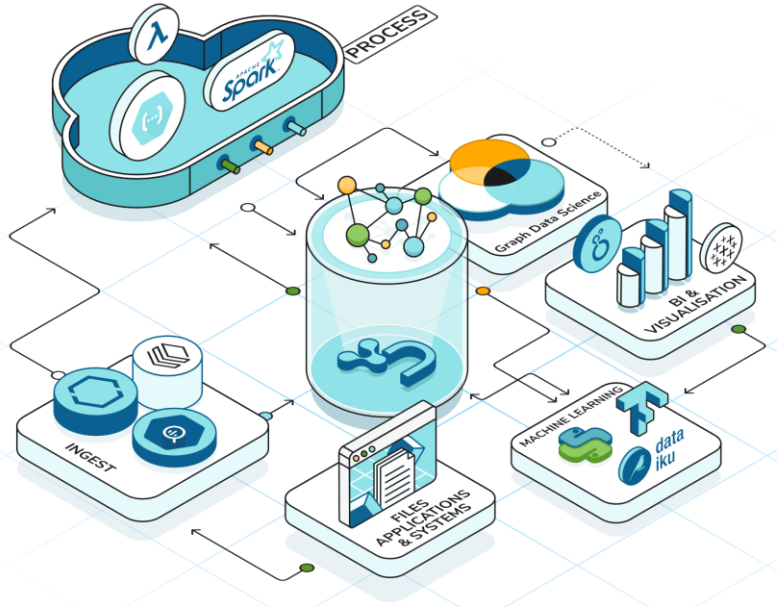
Как быстро собрать самому?

- Выбрать стек
- Вюбрать
  - Хранилище
  - Экстракторы
  - Валидаторы
  - Индексаторы
- Найти туториал
- Организовать датапровод и пространство для экспериментов
- Проверить гипотезу о полноте и согласованности эксплицитного знания для конкретных задач

Главная открытая СУБЗ

- ***Neo4j*** *Graph Database* (http://neo4j.com/product/neo4j-graph-database/)
  - ***Enterprise Edition*** *for scalability, availability, and fine-grained access controls!*
  - *+ Neo4j Graph Data Science* (http://neo4j.com/product/graph-data-science/)
    - *Analytics and machine learning (ML) solution that analyzes relationships in data to improve predictions and discover insights*





Neo4j Neosemantics

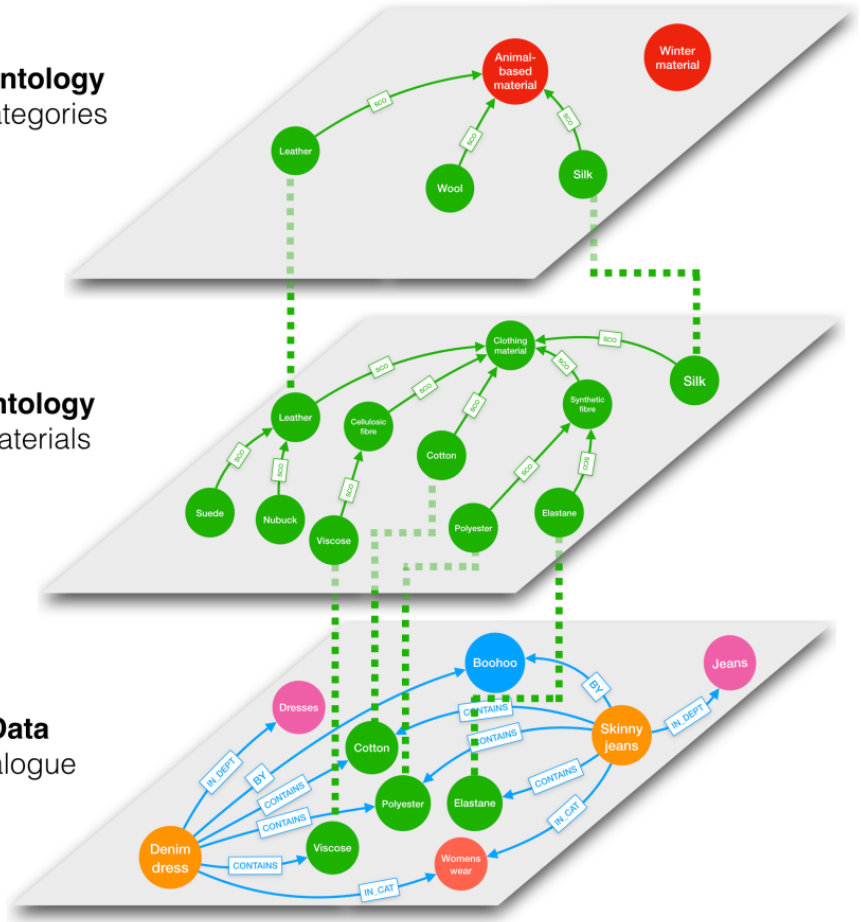
- ***neosemantics*** (*n10s*): *Neo4j RDF & Semantics toolkit* (http://neo4j.com/labs/neosemantics/)
  - *Plugin that enables the use of RDF and its associated vocabularies like (OWL,RDFS,SKOS and others) in Neo4j*
  - *Barrasa J. QuickGraph#9 The fashion Knowledge Graph. Inferencing with Ontologies in Neo4j*. 2019 (http://jbarrasa.com/2019/11/25/quickgraph9-the-fashion-knowledge-graph-inferencing-with-ontologies-in-neo4j/)

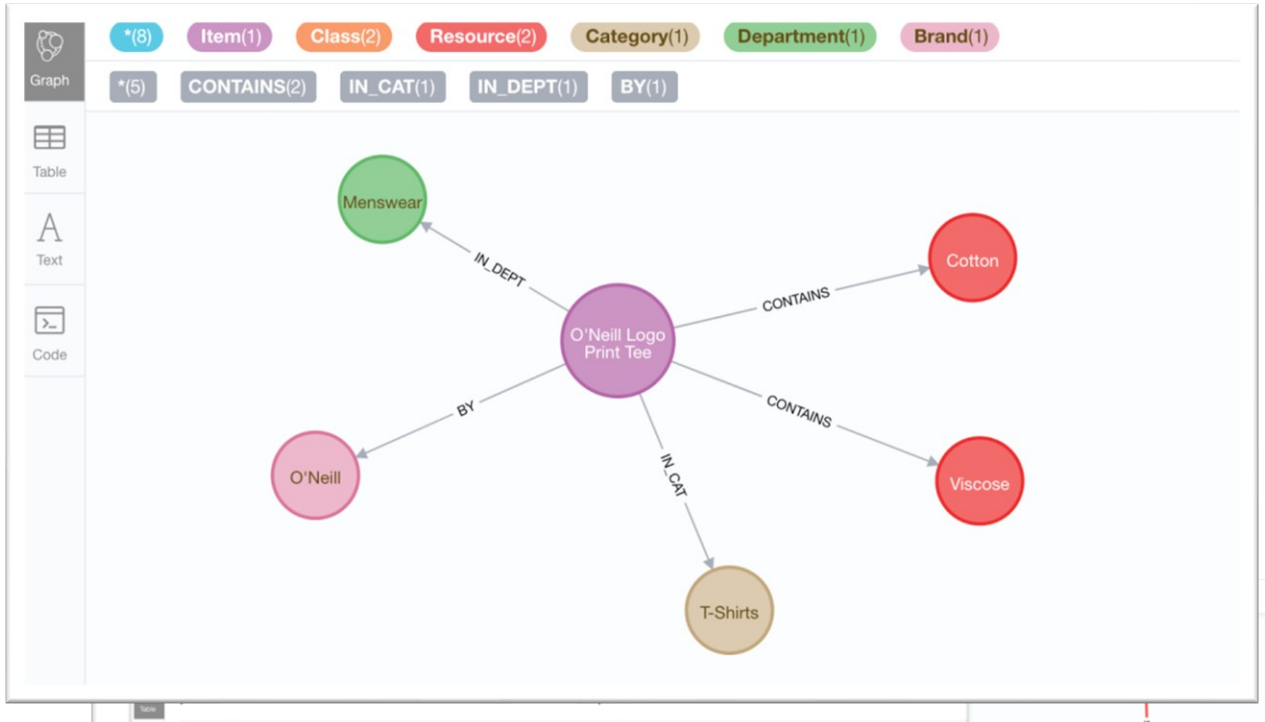
Example from QuickGraph#9

Custom Ontology
Custom Categories

Domain Ontology
Clothing Materials

Instance Data
Product Catalogue





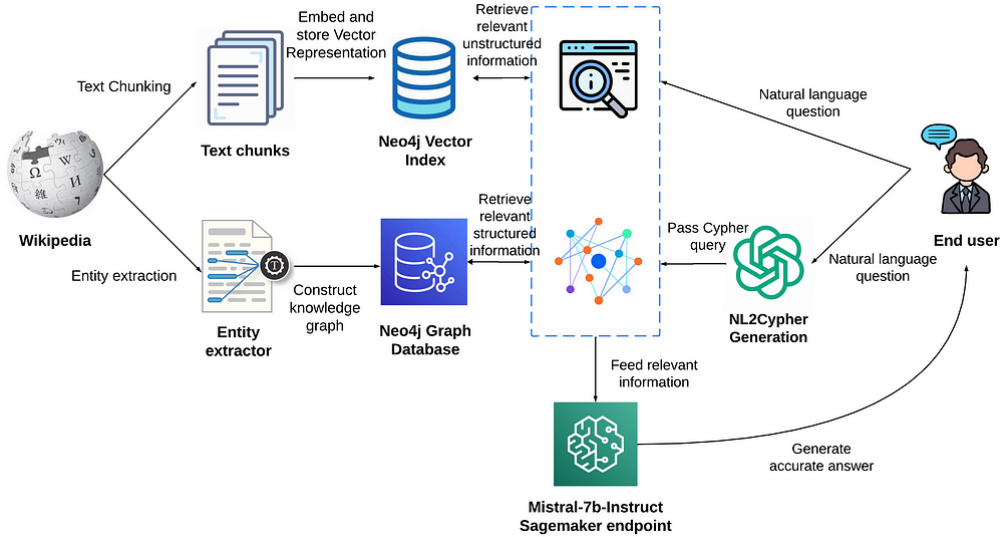
Объединение с «векторами» и документами

- *neo4j Vectors & Semantic Search*. 2023 (http://riino.site/blog/2023-11-26-neo4j_llm)
  - В одну сторону…
- *neo4j Knowledge Graph Builder App* (http://github.com/neo4j-labs/llm-graph-builder)
  - В другую сторону…
    - *Uses ML algorithms to turn unstructured data into actionable knowledge graphs*
  - Стыковка с метаонтологиями и слияние предметных онтологий – отдельная тема…
    - Но БЯМ уже здорово помогают
    - *Agent-OM: Leveraging LLM Agents for Ontology Matching.* 2025 (http://arxiv.org/abs/2312.00326v10)
- Отдельно – ***OntoGPT*** (http://github.com/monarch-initiative/ontogpt)

Neo4j vs. Elasticsearch

- *Neo4j vs. Elasticsearch: Vector Search, RAG, and LLM Integration*
  (http://www.langchain.ca/blog/neo4j-vs-elasticsearch-navigating-the-world-of-vector-search-and-llm-integration/)
  - *Neo4j: The Graph Database Powerhouse*
    - ***vs***
  - *Elasticsearch: The Search and Analytics Engine*

  - *Recommendations:*
    - *Choose Neo4j if your data is **highly interconnected** and you need to leverage complex relationships in your queries or recommendations*
    - *Opt for Elasticsearch if your primary focus is on full-text search, **document-based** recommendations, or handling large volumes of textual data*



Встраивание хранилища в удобный стек

- ***Haystack*** (http://docs.haystack.deepset.ai)
  - *“open-source framework for building production-ready LLM applications, retrieval-augmented generative pipelines and state-of-the-art search systems that work intelligently over large document collections”*
- ***neo4j-haystack*** (http://github.com/prosto/neo4j-haystack)
  - *Using Vector search index!*
  - *For documents collection*

Что в трендах

- Удобство валидации
  - Сейчас очень сложно автоматизируется бесплатными средствами
- Учёт динамики на основе темпоральной логики
  - Связь с временными рядами для поиска «событий»
- Бенчмаркинг
  - Как раз на хайпе

Но это ещё не конец?!

Вопросы? Замечания? Предложения?