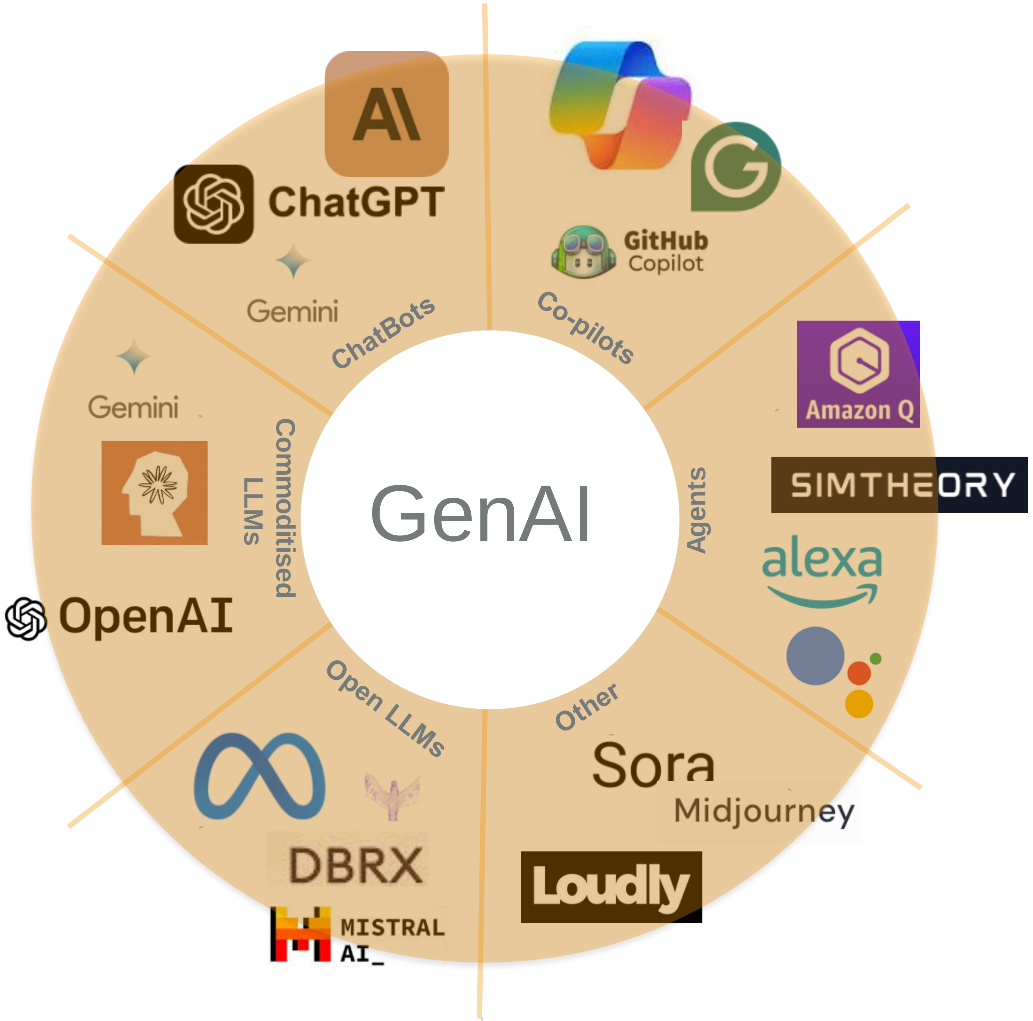
Pistoia Alliance 2024 London conference

Ontologies and LLMs: *Democratising Scientific Data*

Joe Mullen, PhD

24th April 2024

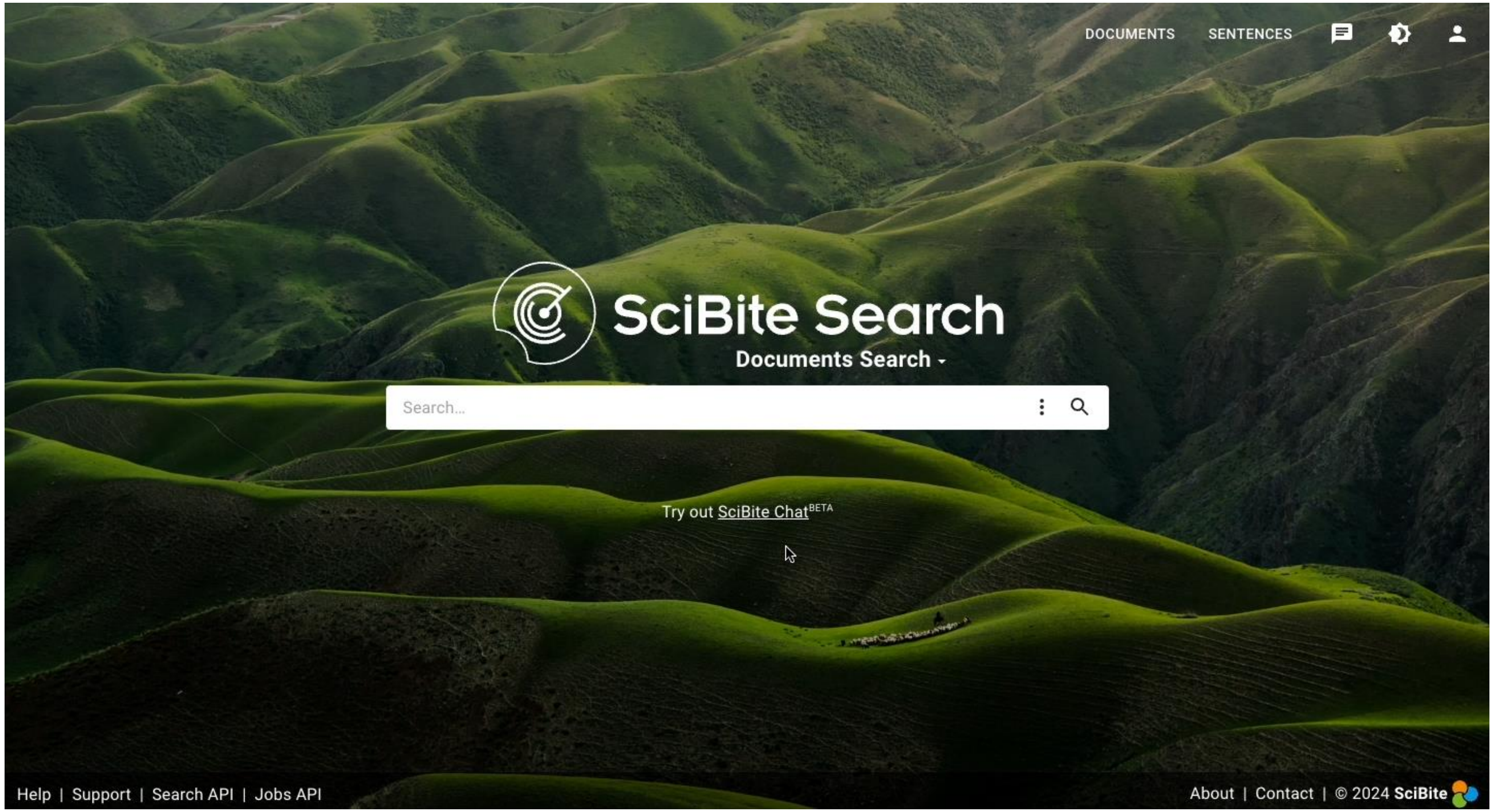

GenAI
ChatBots
ChatGPT
Gemini
Co-pilots
GitHub Copilot
Agents
Amazon Q
SIMTHEORY
alexa
Other
Sora
Midjourney
Loudly
Open LLMs
DBRX
MISTRAL AI_
Commoditised LLMs
Gemini
OpenAI

Roll out and adoption of GenAI brings changes to **user expectations** when interfacing with technology

How can GenAI enable insight within life sciences?

SciBite Chat

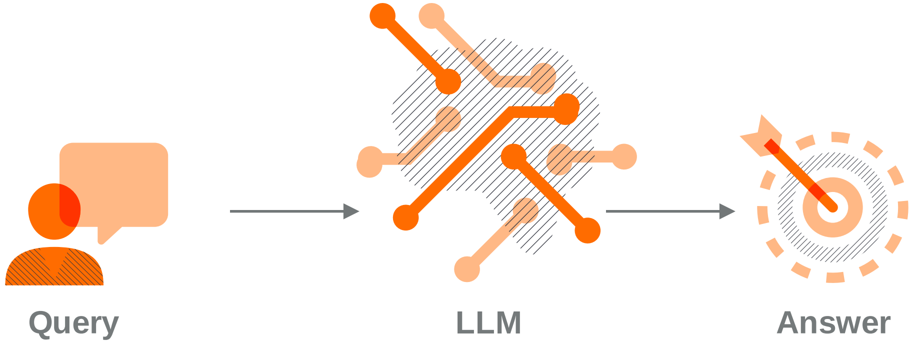
Launched April '24!


DOCUMENTS
SENTENCES
SciBite Search
Documents Search
Search...
Help | Support | Search API | Jobs API
About | Contact | © 2024 SciBite

Try out SciBite Chat BETA

Democratising data; speeding up insight

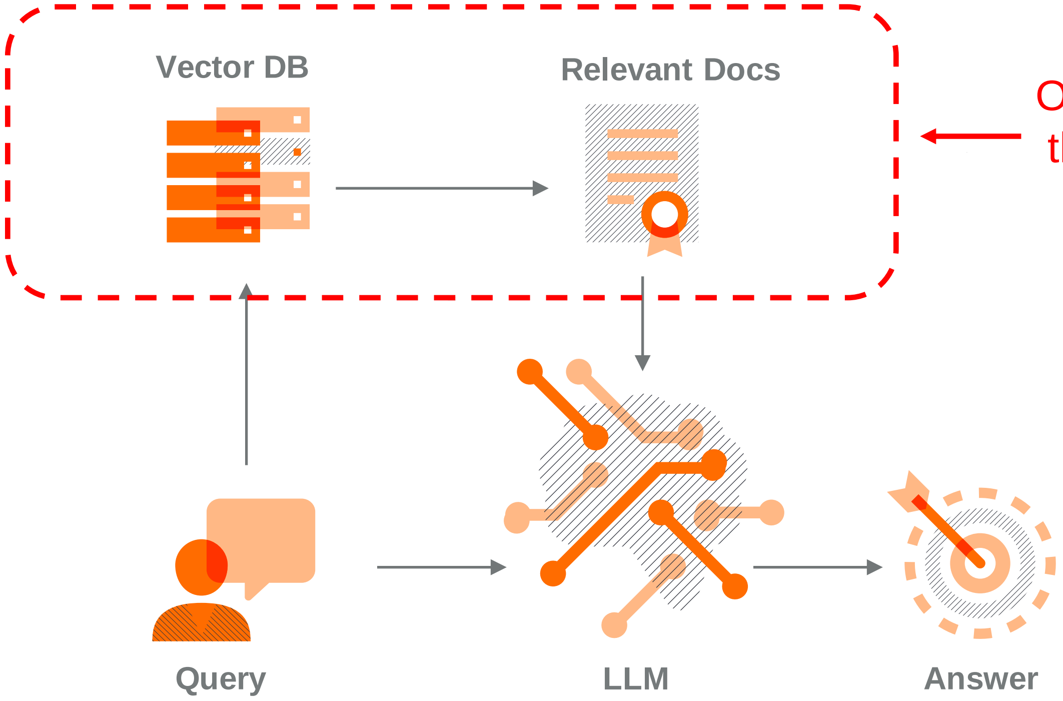
But how did we get here?


Query
LLM
Answer

Did we just ask an LLM direct?

accuracy ✗ | transparency ✗ | flexibility ✗

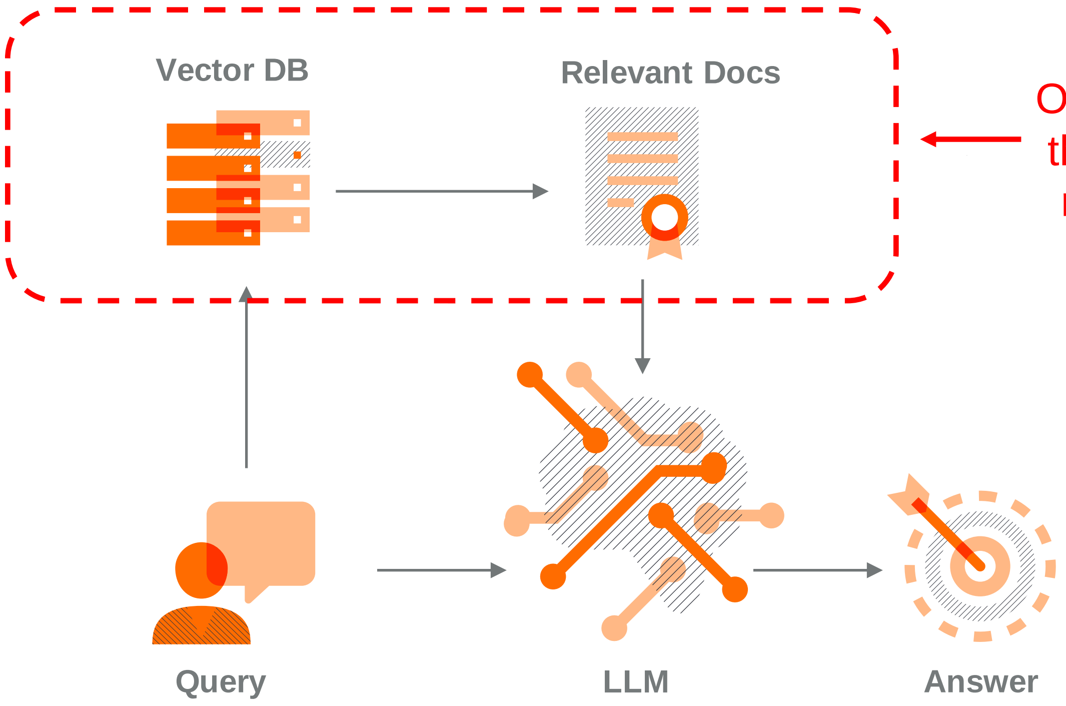
Not quite....


Vector DB
Relevant Docs
Query
LLM
Answer

Only as good as the information retrieval step!

RAG architecture

*Retrieval Augmented Generation


Vector DB
Relevant Docs
Query
LLM
Answer

Only as good as the information retrieval step!

Did we make use of a RAG architecture with **vector-based** information retrieval?

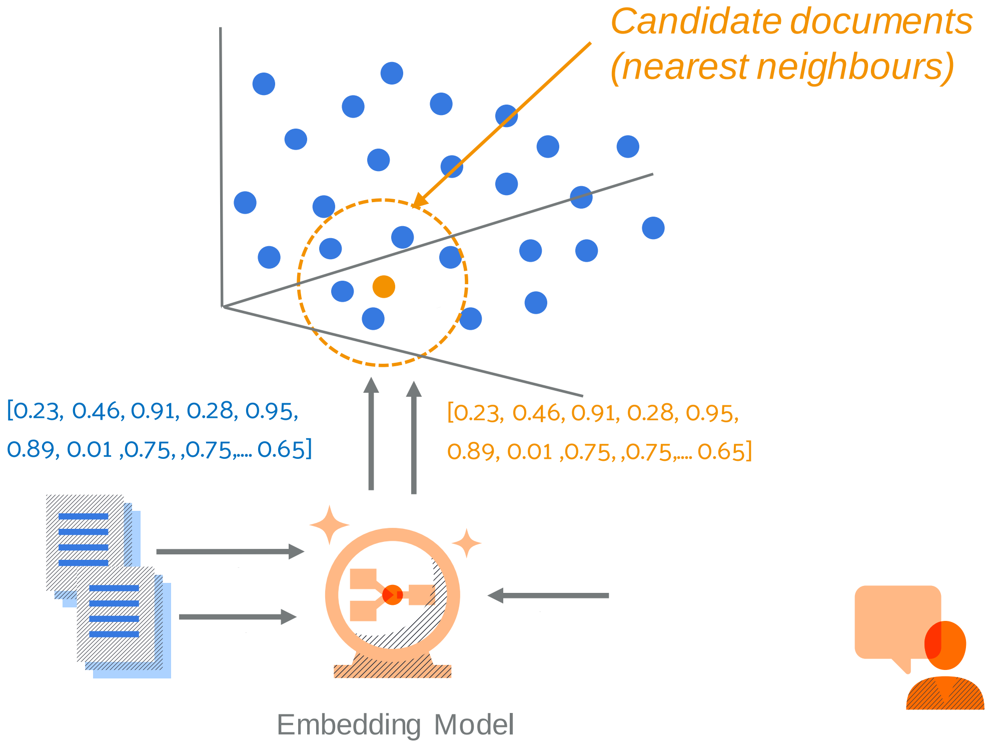
accuracy	transparency	flexibility
?	✗	✓

Not quite....

*Retrieval Augmented Generation

transparency **human explainability**

*Vector-based retrieval – **what** is related, not **how***



Candidate documents (nearest neighbours)

[0.23, 0.46, 0.91, 0.28, 0.95, 0.89, 0.01 ,0.75, ,0.75,.... 0.65]

[0.23, 0.46, 0.91, 0.28, 0.95, 0.89, 0.01 ,0.75, ,0.75,.... 0.65]

Embedding Model

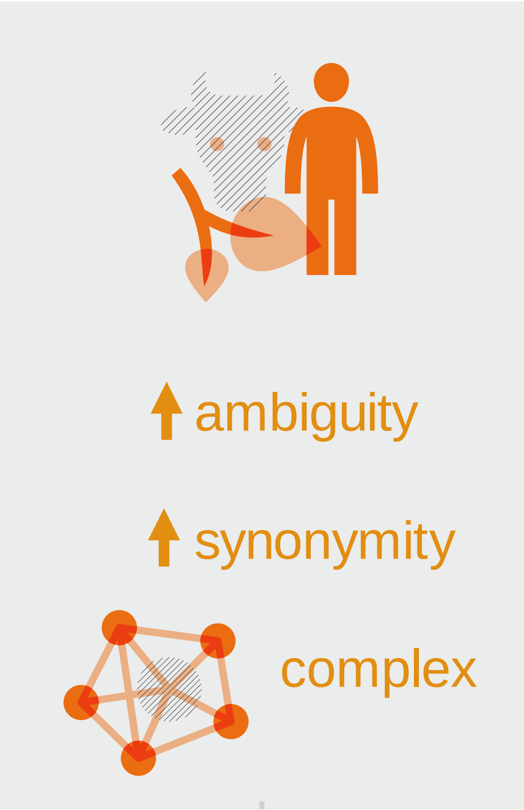
Amazing technology, almost like magic.... **but** how do you explain magic?

There is another way...

Building on domain knowledge

We’ve spent **a long** time organising the chaos in data....

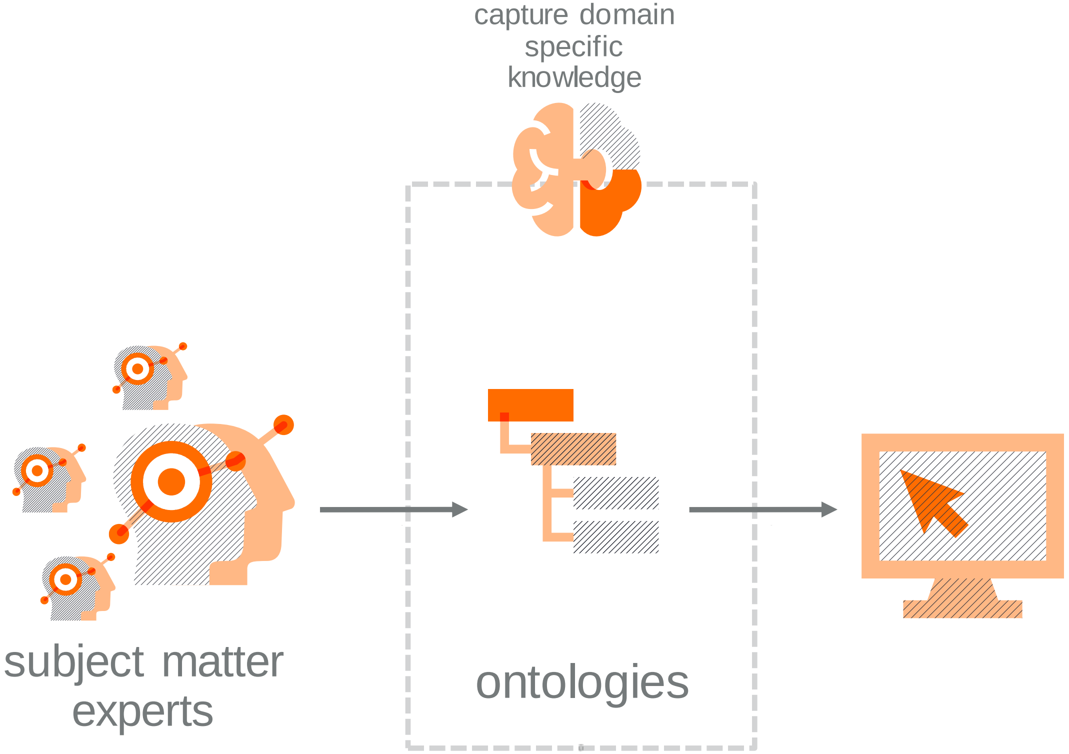
Life Sciences



↑ ambiguity

↑ synonymity

complex



capture domain specific knowledge

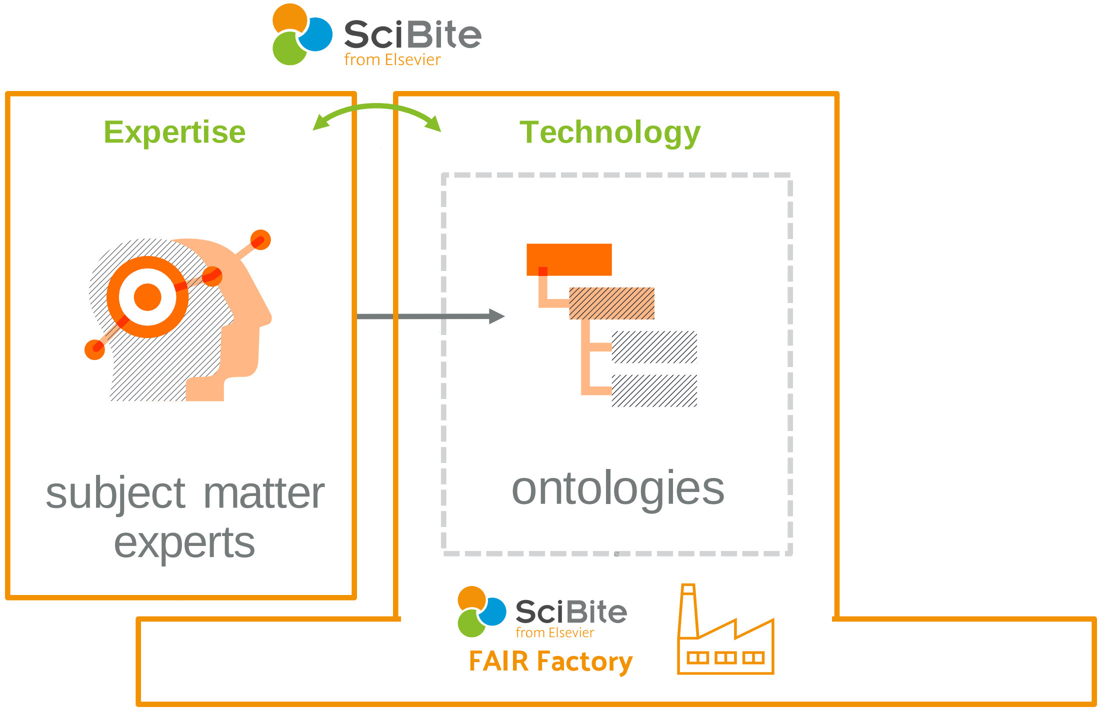
subject matter experts

ontologies

internal
external
collaborative

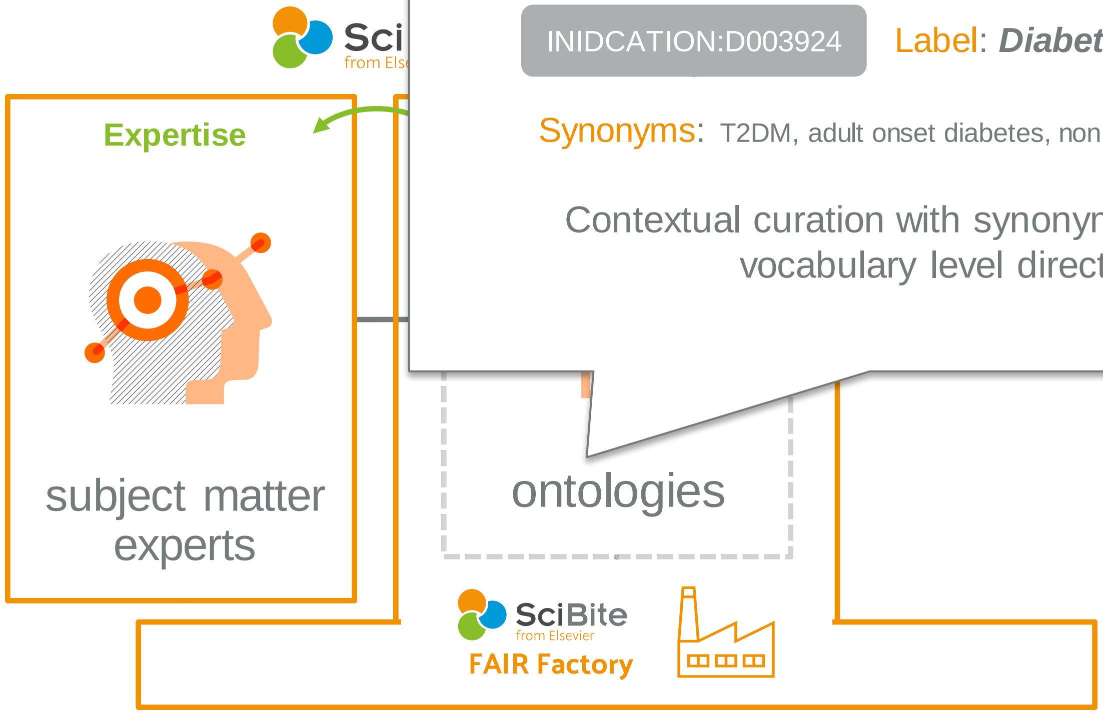
Building on domain knowledge

We've spent a **long** time organising the chaos in data....


SciBite
from Elsevier
Expertise
Technology
subject matter experts
ontologies
SciBite
from Elsevier
FAIR Factory

Building on domain knowledge

We’ve spent a **long** time organising the chaos i


Sci
from Else
Expertise
subject matter experts
ontologies
SciBite
from Elsevier
FAIR Factory

transforming ontologies for named entity recognition

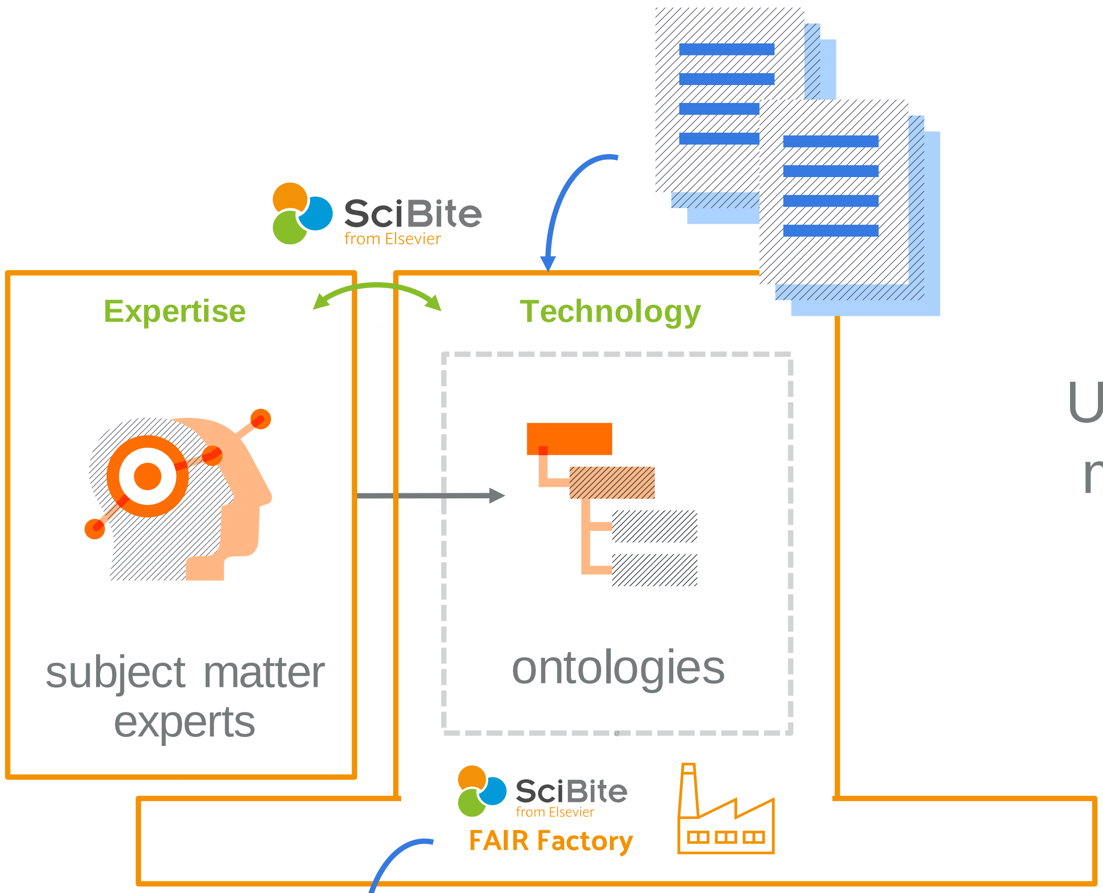
INIDCATION:D003924 Label: ***Diabetes Mellitus, Type 2***

Synonyms: T2DM, adult onset diabetes, non insulin-dependent DM

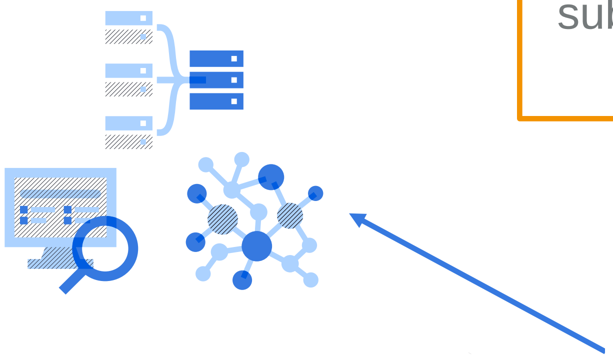
Contextual curation with synonym switches and vocabulary level directives

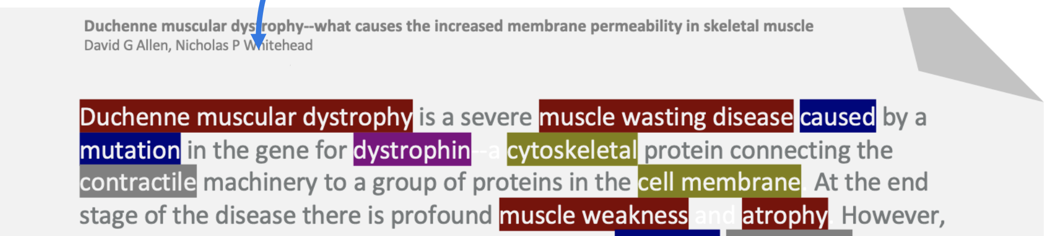
Building on domain knowledge

We've spent a **long** time organising the chaos in data....


SciBite
from Elsevier
Expertise
Technology
subject matter experts
ontologies
SciBite
from Elsevier
FAIR Factory

Unstructured text to machine readable, semantic data



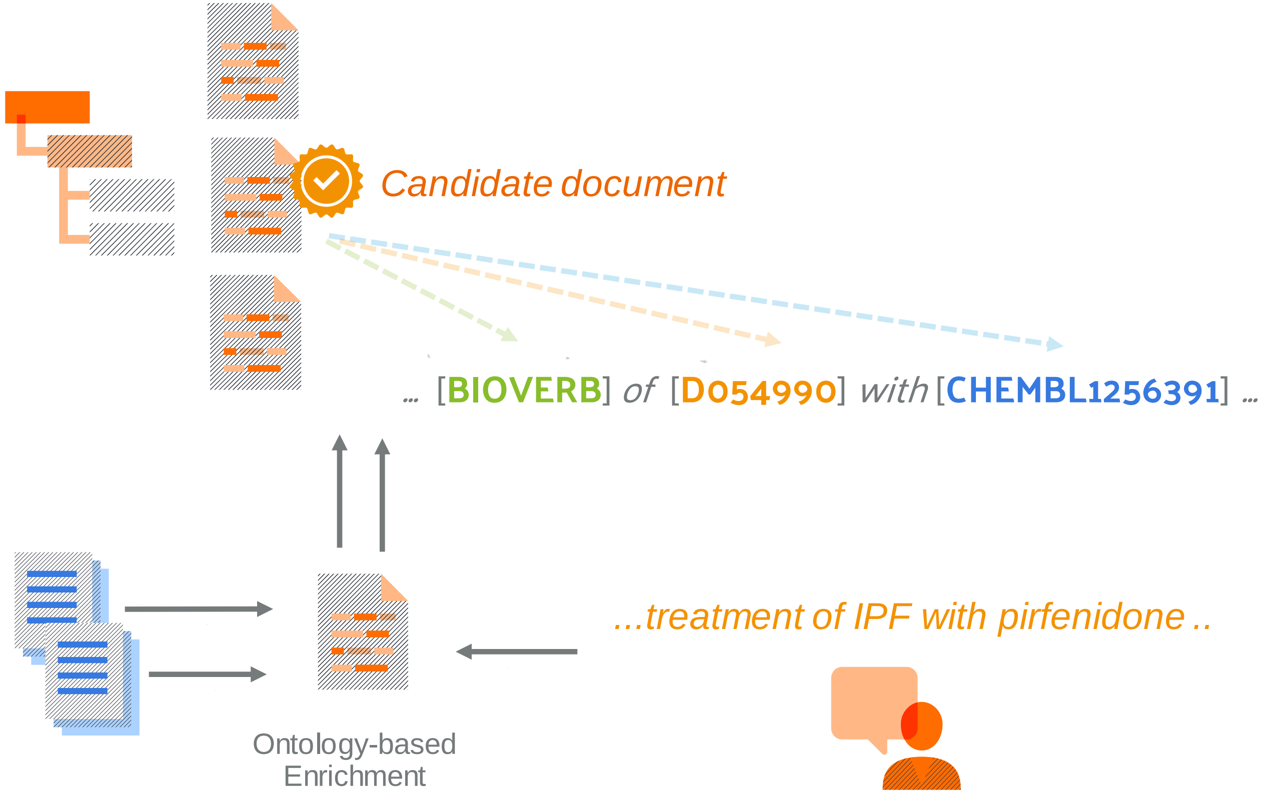


Duchenne muscular dystrophy--what causes the increased membrane permeability in skeletal muscle
David G Allen, Nicholas P Whitehead

Duchenne muscular dystrophy is a severe muscle wasting disease caused by a mutation in the gene for dystrophin--a cytoskeletal protein connecting the contractile machinery to a group of proteins in the cell membrane. At the end stage of the disease there is profound muscle weakness and atrophy. However,

transparency **human explainability**

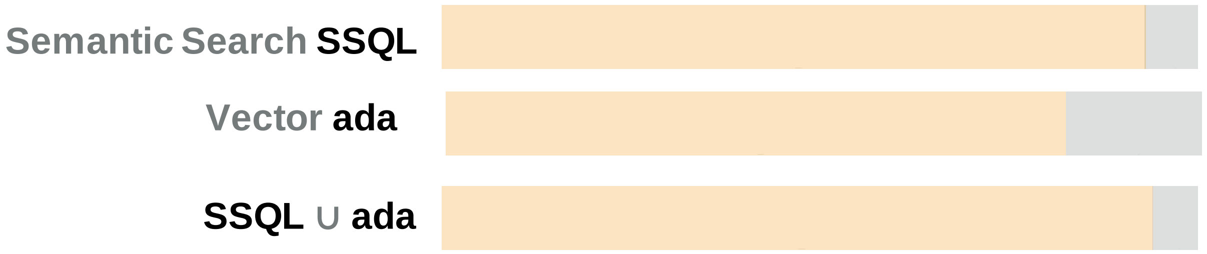
*ontology-based retrieval – what it **is**, **what** is related, **and** how*


Candidate document
… [BIOVERB] of [D054990] with [CHEMBL1256391] …
…treatment of IPF with pirfenidone ..
Ontology-based Enrichment

Looking at the accuracy of the information retrieval approaches

Information Retrieval

Reviewing accuracy, internal benchmark



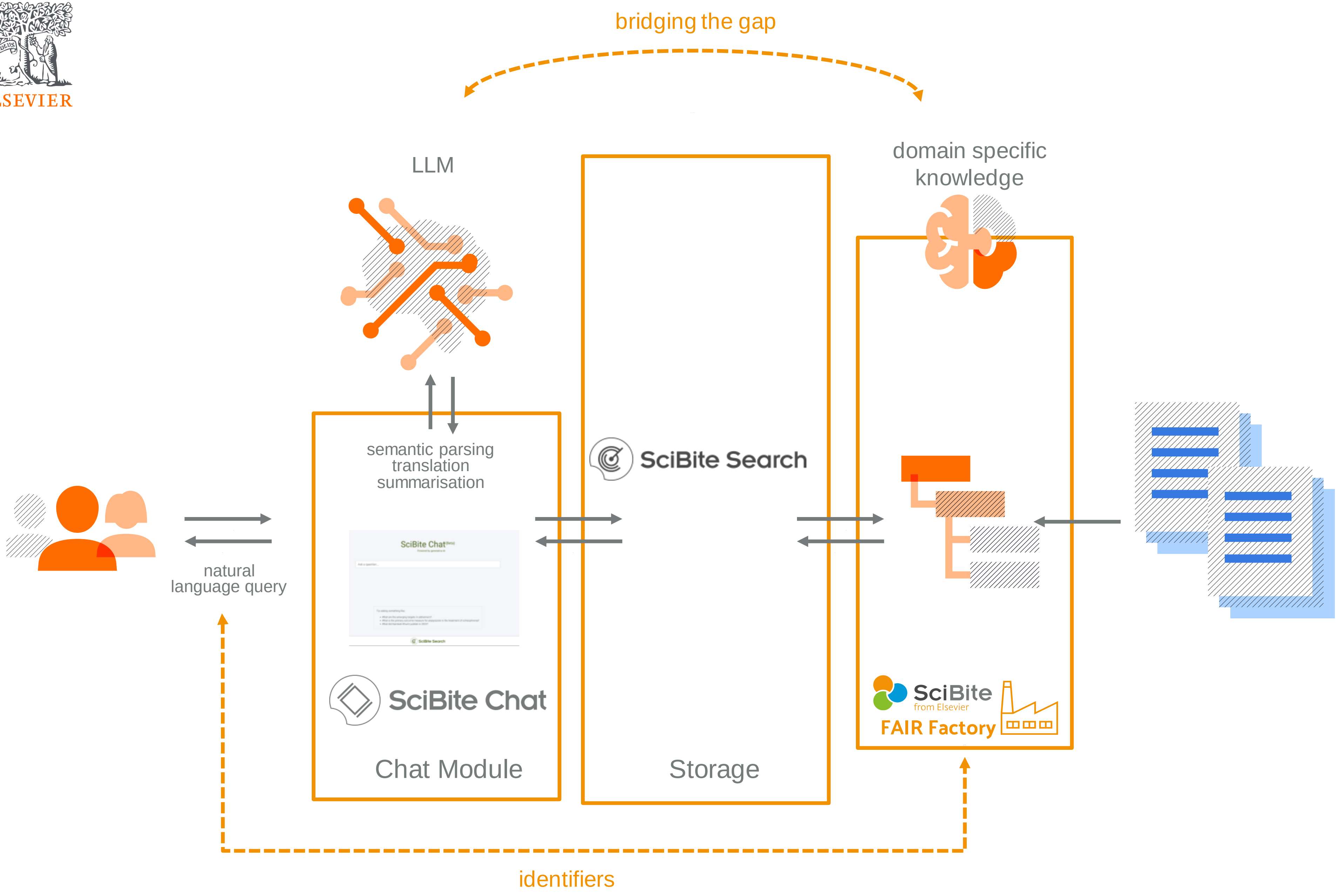
Questions with at least one answer match

We used a RAG architecture with SciBite ontology-based information retrieval....

ELSEVIER

Let’s take a closer look at

SciBite Chat



bridging the gap

LLM

domain specific
knowledge

semantic parsing
translation
summarisation

SciBite Chat

SciBite Search

natural
language query

SciBite Chat

Chat Module

Storage

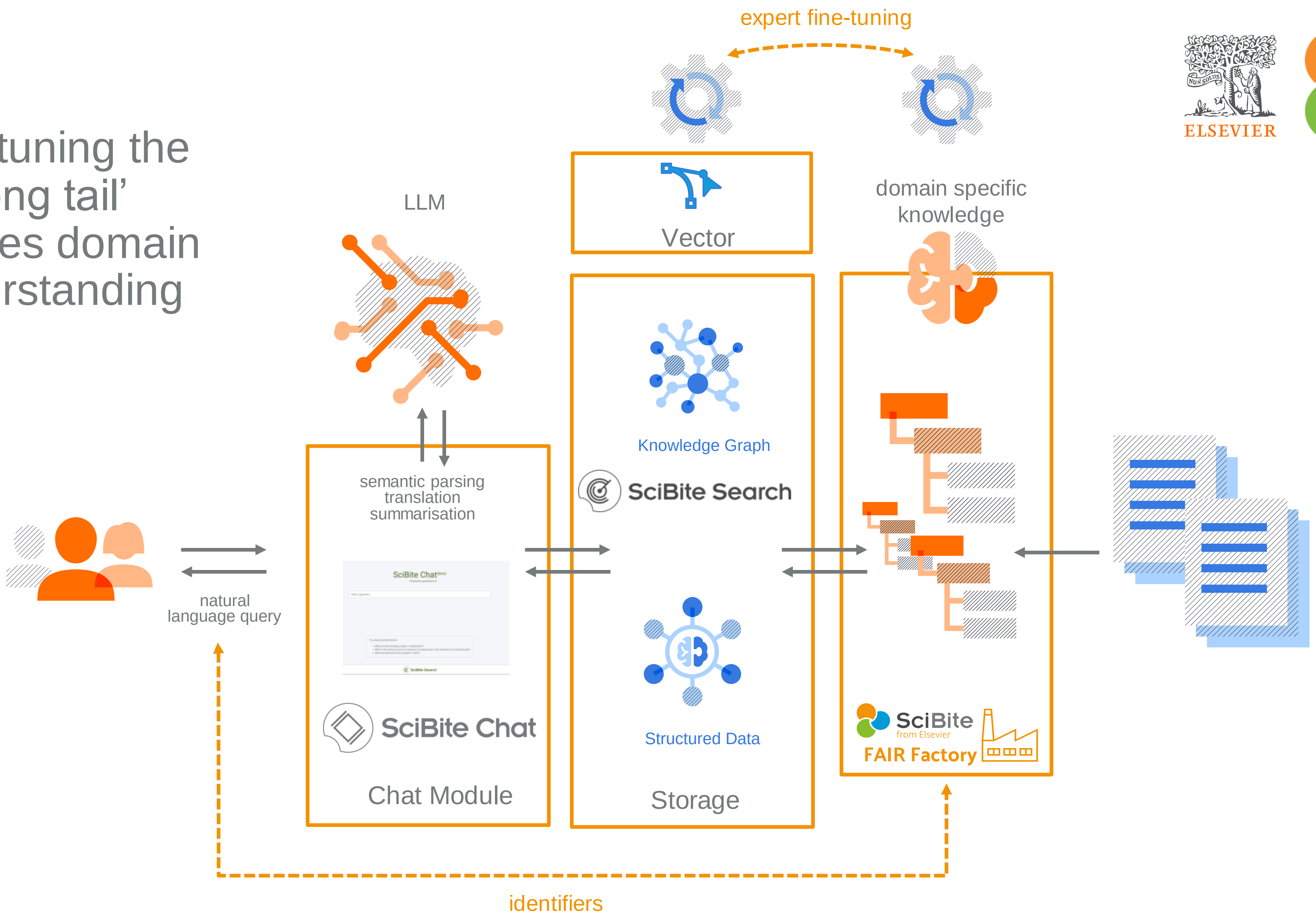
SciBite
from Elsevier
FAIR Factory

identifiers

Flexibility
Deployment
Security
Data
Ontologies

The vision

Fine-tuning the ‘long tail’ requires domain understanding


expert fine-tuning
LLM
Vector
domain specific knowledge
semantic parsing
translation
summarisation
SciBite Chat
Chat Module
natural language query
Knowledge Graph
SciBite Search
Structured Data
Storage
SciBite from Elsevier
FAIR Factory
identifiers

TL;DR

SciBite Chat

Data

Expertise

Technology

- accuracy
- transparency
- flexibility
- user expectations

We speak your language – hear what your data has to say!

SciBite Chat

Acknowledgements

SBS Team
Phil Verdemato
Rob Martin
Andrew Cowley
Dave Styles
Ian Hart
Alex Biddle
Svetlana Teneva

Data Science
Oliver Giles
Thom Pijnenburg

PMO
Tim Miller
George Jiang

TIO
Helen Potts
Divya Nair

SciBite
from Elsevier

Thanks for listening!

www.scibite.com
info@scibite.com
joe@scibite.com

SciBite Chat
Take the next step
Request a Demo

SciBite on LinkedIn

Come and talk to the team!

Joe Mullen
Director Data Science & Professional Services

Jane Lomax
Head of Ontologies

Rebecca Foulger
Senior Biocurator - Content Integration Manager

Romain Elleboode
Sales Director - EMEA

Jamie Smith
Sales Director - EMEA

Ajit Singh
Technical Consultant