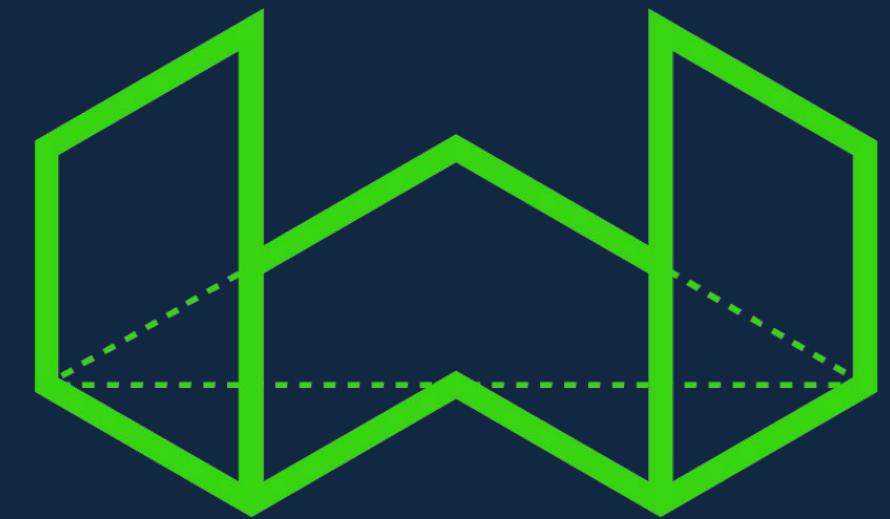




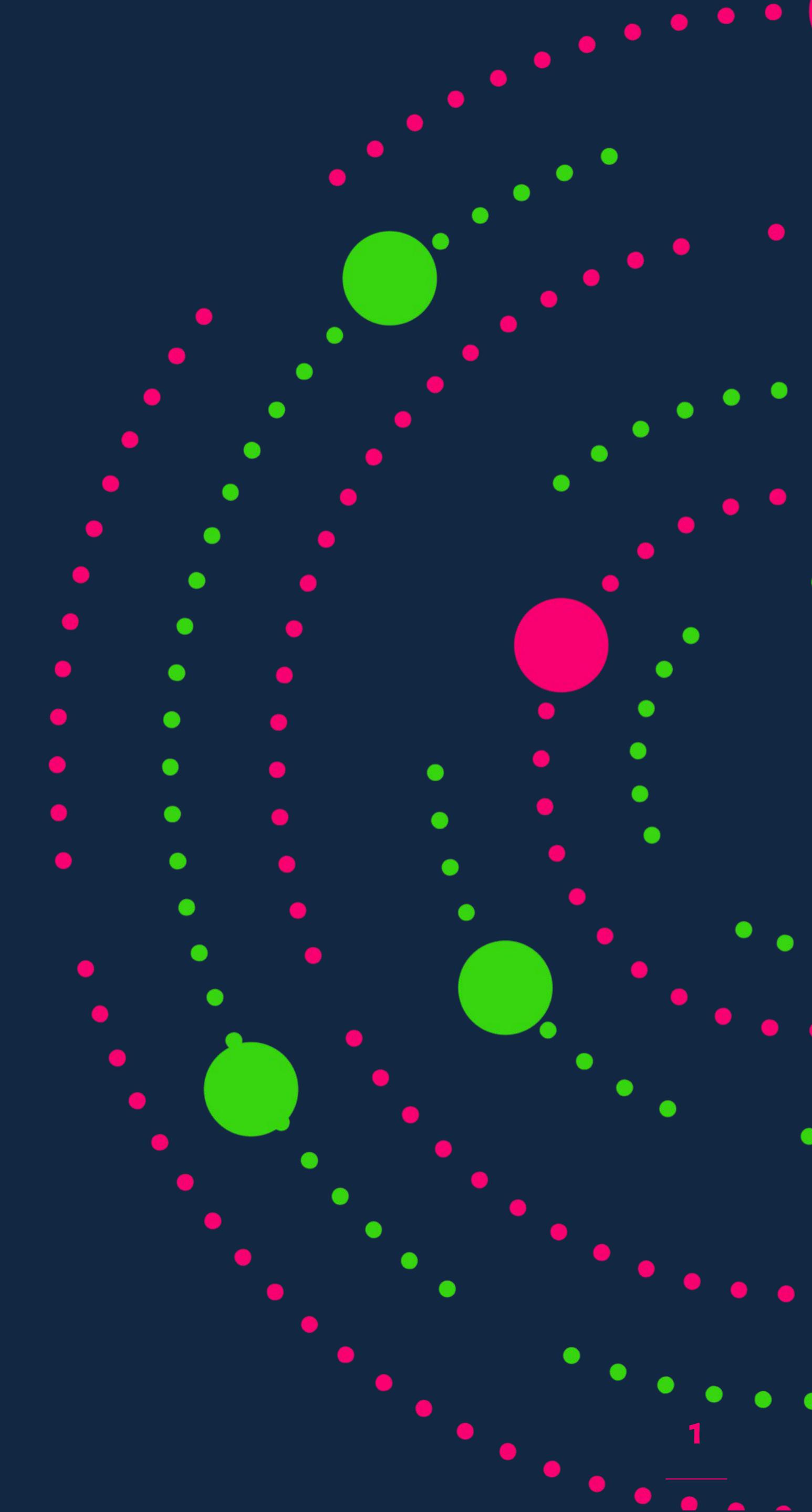
SeMI Technologies



# Weaviate

Introduction to Vector Search

@Conf42 - 19 May 2022 - by Laura Ham



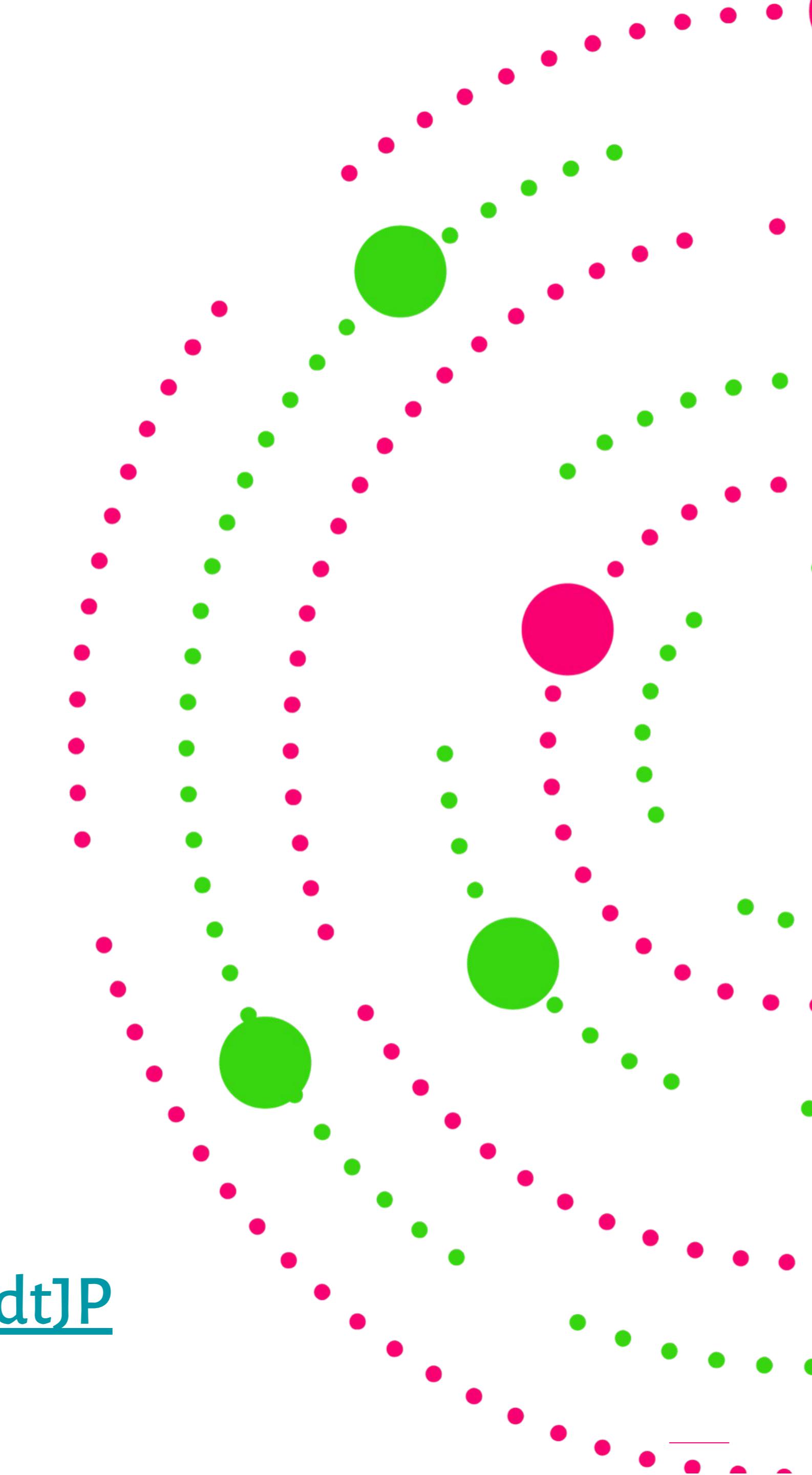


# SeMI Technologies

## I'm Laura, nice to meet you!

Let's meet on:

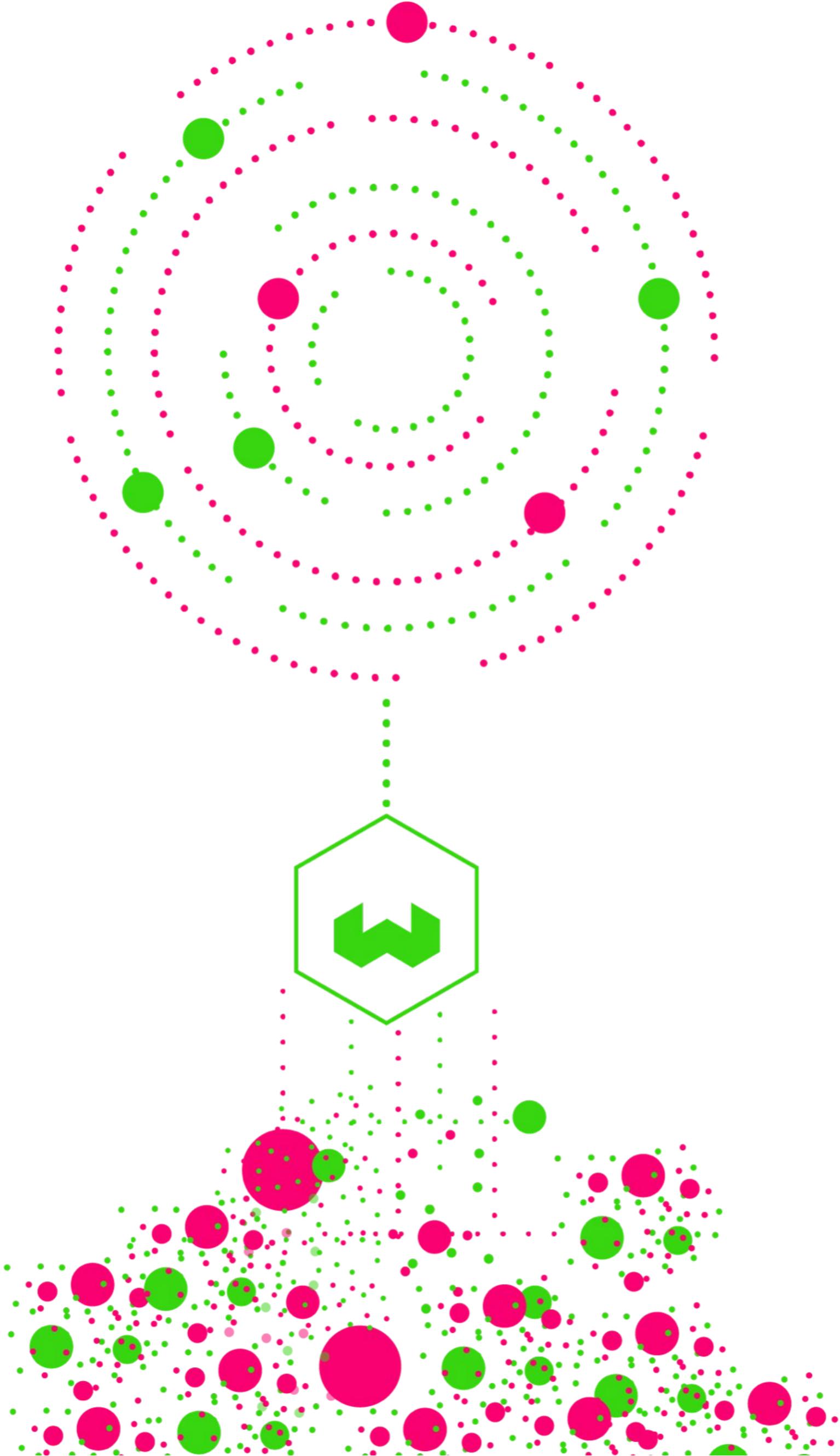
- LinkedIn: [Laura Ham](#)
- Email: [laura@semi.technology](mailto:laura@semi.technology)
- Slack: <https://link.semi.technology/3vbEtKN>
- Download slides at: <https://link.semi.technology/3kUdtJP>



# I'm going to talk about:

---

1. Structured vs. unstructured data
2. What is a vector database and how does it work?
3. Vector search with Weaviate: features and live demos :)
4. Q&A



---

# 1. What is a Vector Database?

---

What's so **difficult** about  
unstructured data?

# Structured vs. Unstructured data

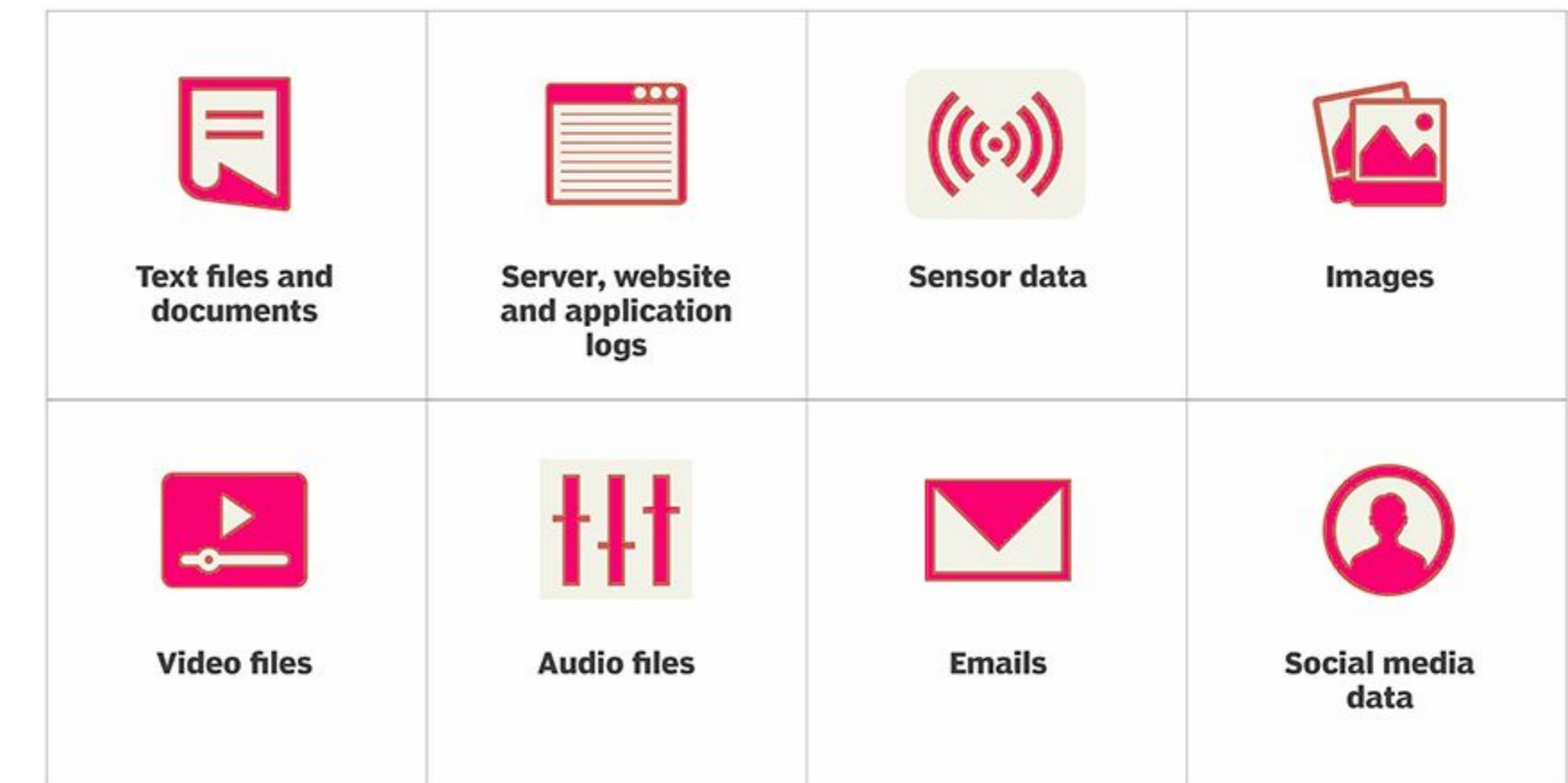
## Structured data

- What you find in a typical database

| ID | Name    | City       |
|----|---------|------------|
| 1  | Alice   | Amsterdam  |
| 2  | Bob     | Berlin     |
| 3  | Charlie | Copenhagen |

## Unstructured data

- What you find in the 'wild'



# Example situation: search engine



{"article": "The origins of dogs - How dogs were domesticated"}

"Animals"

No articles found ...

*Traditional search engine*

"Animals"

The origins of dogs  
- How dogs were  
domesticated

*Vector search engine*

# This can be solved with Vector Search!

Google is a vector search engine too!

Thanks to this, you can use Google Search to find answer in a vast amount of webpages. But did you know that it is estimated that Google only indexes about 0.004% of unstructured data? That's because companies like yours have the rest! What if you could use similar search technology for your data in a reliable and secure way?

The screenshot shows a Google search results page. The search bar at the top contains the query "What color of wine is Chardonnay?". A red arrow points from a callout box containing the text "The question is very abstract." to the search bar. Below the search bar, the Google logo is visible. The search interface includes standard filters: All, Images, Shopping, News, Videos, More, Settings, and Tools. The search results section displays the following information:

- About 12.800.000 results (0,81 seconds)
- Chardonnay / Wine color
- White Wine**
- A small image of a glass of white wine.
- A quote: "Chardonnay is the most compelling and popular **white wine** in the world, because it is the red wine of whites," Ramey said. "It's so complex, so interesting. And it's the red wine of whites for two reasons: barrel fermentation and malolactic." 26 Jul 2019

# This can be solved with Vector Search!

Google is a vector search engine too!

Thanks to this, you can use Google Search to find answer in a vast amount of webpages. But did you know that it is estimated that Google only indexes about 0.004% of unstructured data? That's because companies like yours have the rest! What if you could use similar search technology for your data in a reliable and secure way?

The screenshot shows a Google search results page. The search query "What color of wine is Chardonnay?" is entered in the search bar. Below the search bar, there are navigation links for All, Images, Shopping, News, Videos, and More. The text "About 12.800.000 results (0,81 seconds)" is displayed. The first result is titled "Chardonnay / Wine color" and features the text "White Wine". A callout box with a pink border and arrow points from the text "The question is very abstract." to the search query. Another callout box with a pink border and arrow points from the text "Finds a concrete answer from unstructured data" to the result "White Wine". To the right of the search results, there is a small image of a glass of white wine.

The question is very **abstract**.

Finds a **concrete answer** from **unstructured data**

# This can be solved with Vector Search!

Google is a vector search engine too!

Thanks to this, you can use Google Search to find answer in a vast amount of webpages. But did you know that it is estimated that Google only indexes about 0.004% of unstructured data? That's because companies like yours have the rest! What if you could use similar search technology for your data in a reliable and secure way?

The screenshot shows a Google search results page. The search bar at the top contains the query "What color of wine is Chardonnay?". Below the search bar, there are several navigation links: All, Images, Shopping, News, Videos, More, Settings, and Tools. The "All" link is highlighted. The search results section starts with the text "About 12.800.000 results (0,81 seconds)". The first result is titled "Chardonnay / Wine color" and features the text "White Wine". A callout bubble with a pink border and arrow points from the text "The question is very abstract." to the word "Abstract" in the result title. Another callout bubble with a pink border and arrow points from the text "And it answers fast!" to the word "fast" in the result title. A third callout bubble with a pink border and arrow points from the text "Finds a concrete answer from unstructured data" to the word "concrete" in the result title. To the right of the search results, there is a small image of a glass of white wine.

The question is very **abstract**.

What color of wine is Chardonnay?

All Images Shopping News Videos More Settings Tools

About 12.800.000 results (0,81 seconds)

Chardonnay / Wine color

White Wine

And it answers **fast**!

Finds a **concrete answer** from **unstructured data**

"Chardonnay is the most compelling and popular **white wine** in the world, because it is the red wine of whites," Ramey said. "It's so complex, so interesting. And it's the red wine of whites for two reasons: barrel fermentation and malolactic." 26 Jul 2019

**Google** is good at working with unstructured data on the public web (which is < 0.01% of data available)

---

**Google** is good at working with unstructured data on the public web (which is < 0.01% of data available)

---

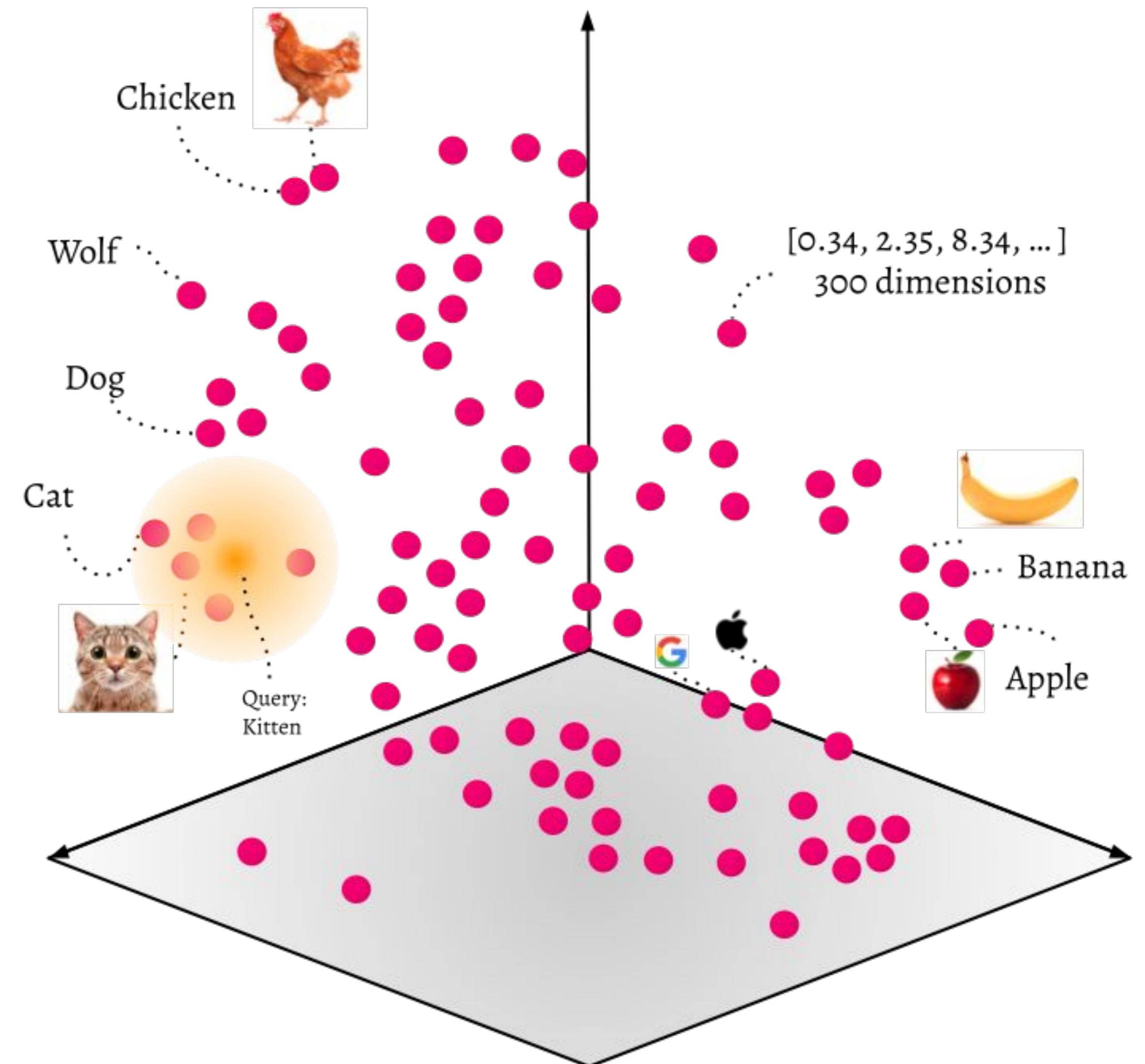
**What if you could do the same  
with your own data in a simple  
and secure way?**

# What is vector search?

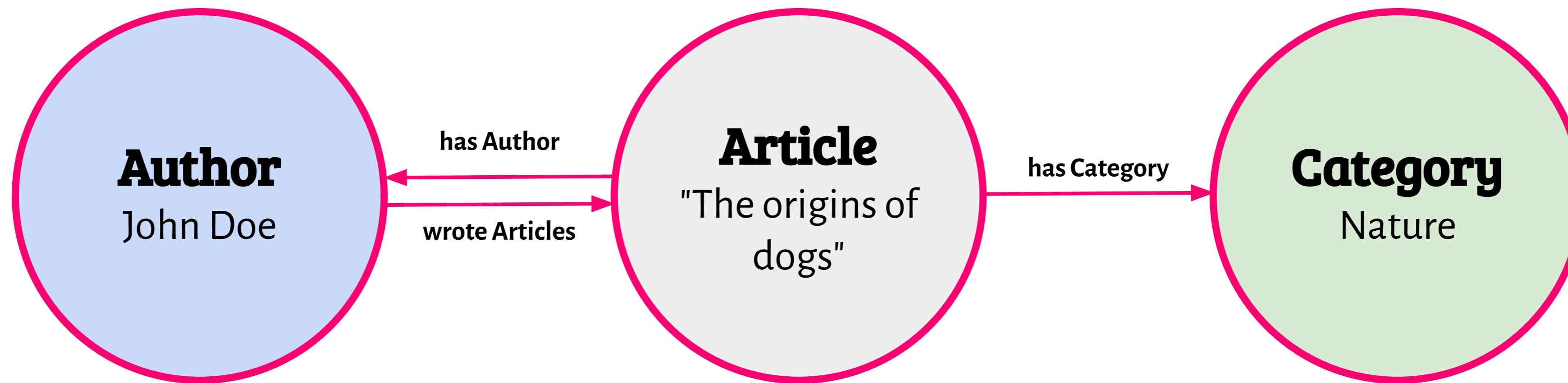
Weaviate is a *vector search engine*

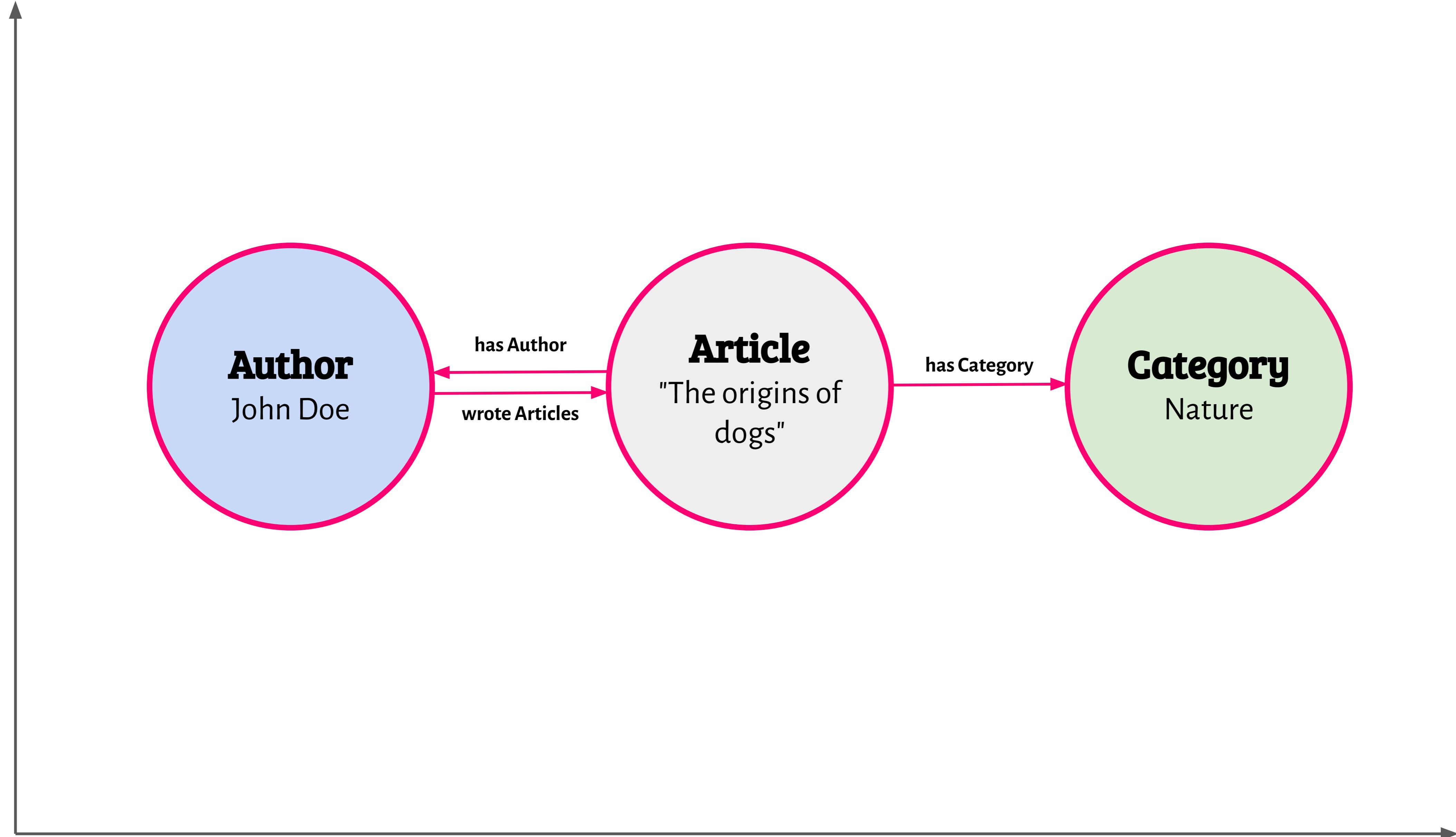
Instead of just storing raw data like traditional databases do, Weaviate leverages the power of machine learning (ML) models to *vectorize* the data. What this means, is that the ML-models try to understand your data while storing it. This allows Weaviate to search, discover and classify similar results in your dataset.

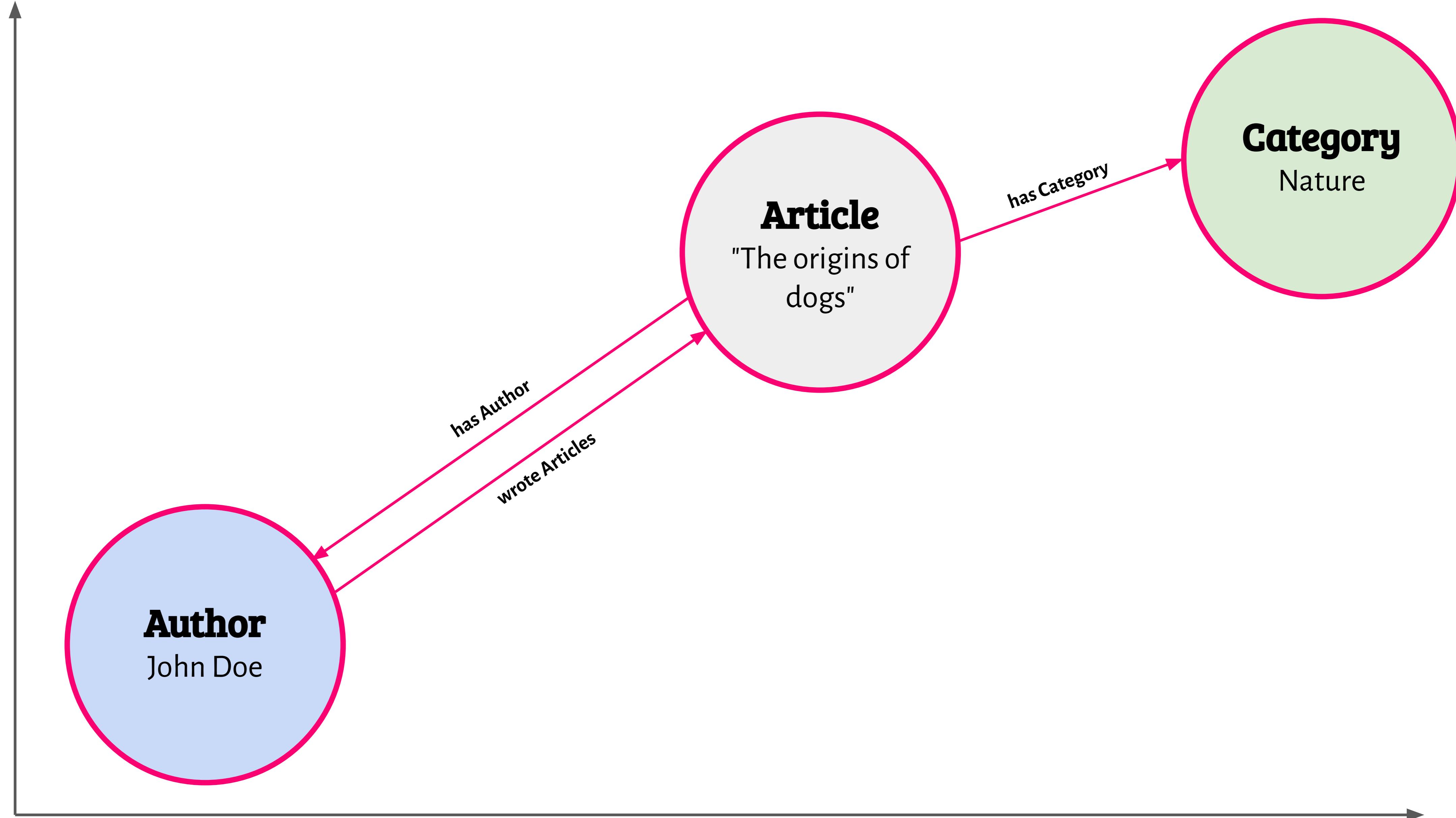
Simply put, **Weaviate is a database that understands your data.**

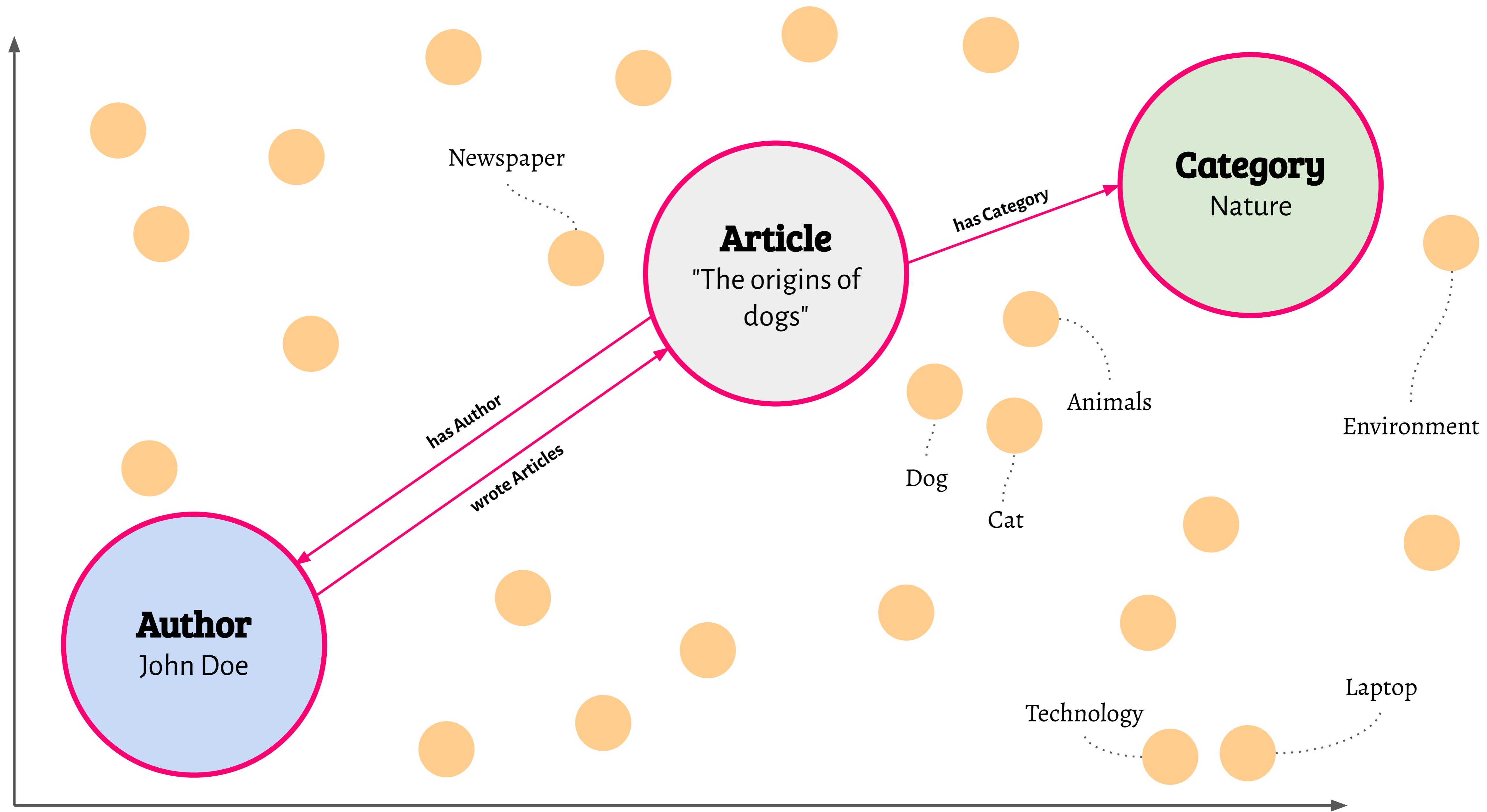


# From Graph Database to Vector Database







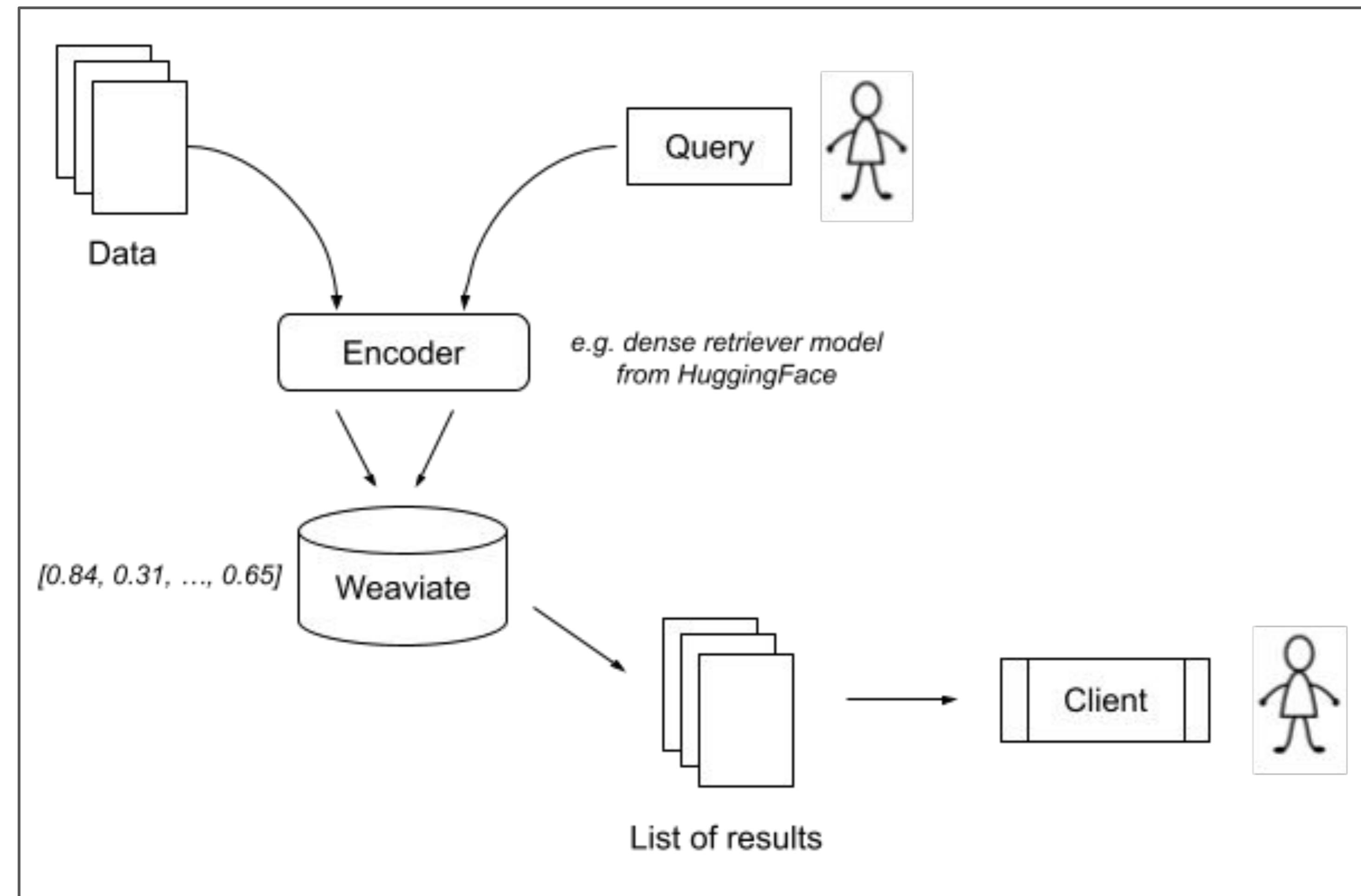


---

# How does vector search **work**?

# How does vector search work?

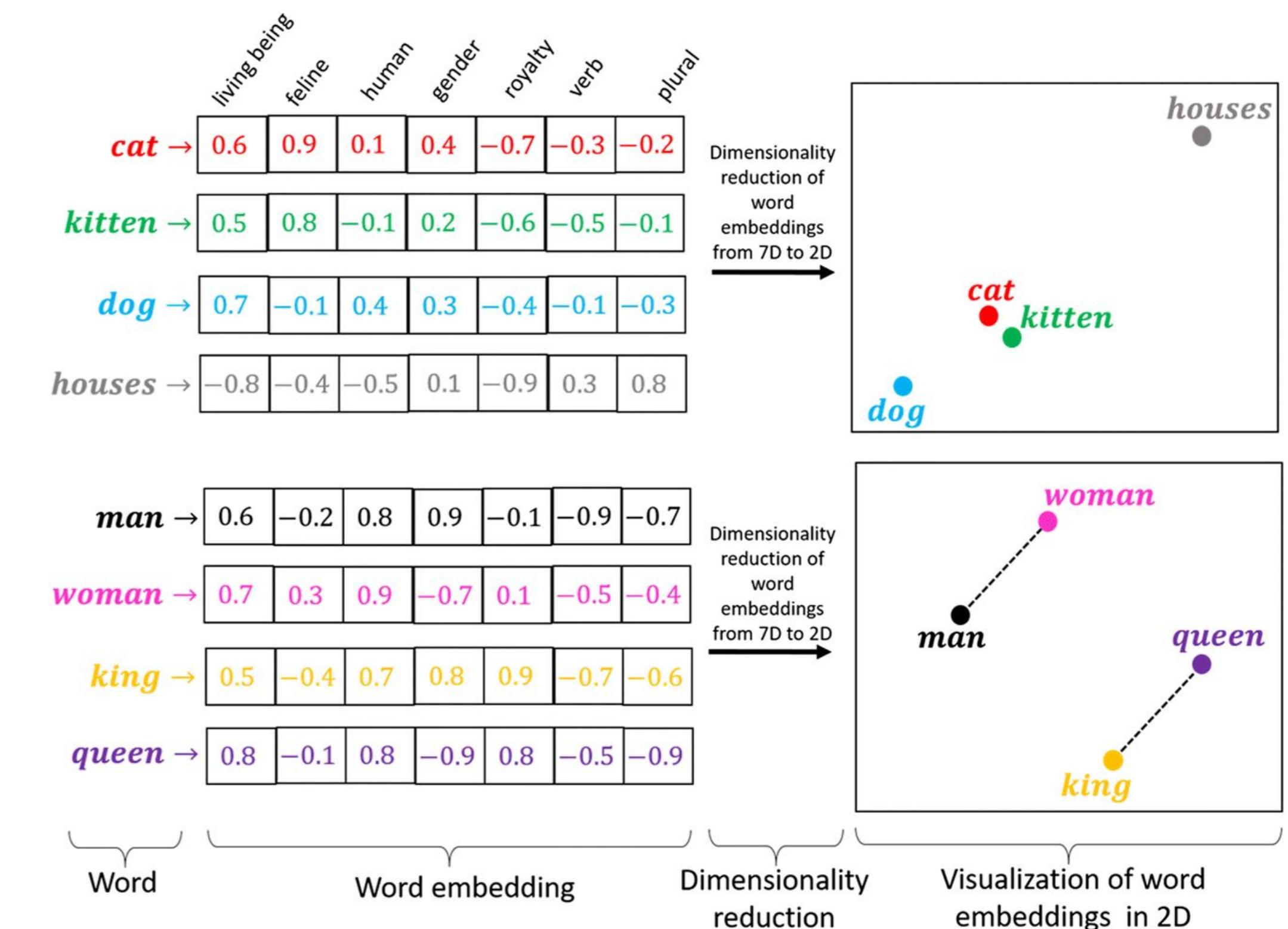
---



# How does vector search work?

## 1. Choose an encoder model

- Transforms data into vectors
- Also called *retriever models*
- Dense retrievers:
  - Embeddings calculated by deep neural networks
  - Language models represent words and concepts in hyperspace
  - Examples: BERT, Sentence Transformers, ResNet50
- Alternative to neural networks:
  - Sparse retrievers: TF-IDF or BM25

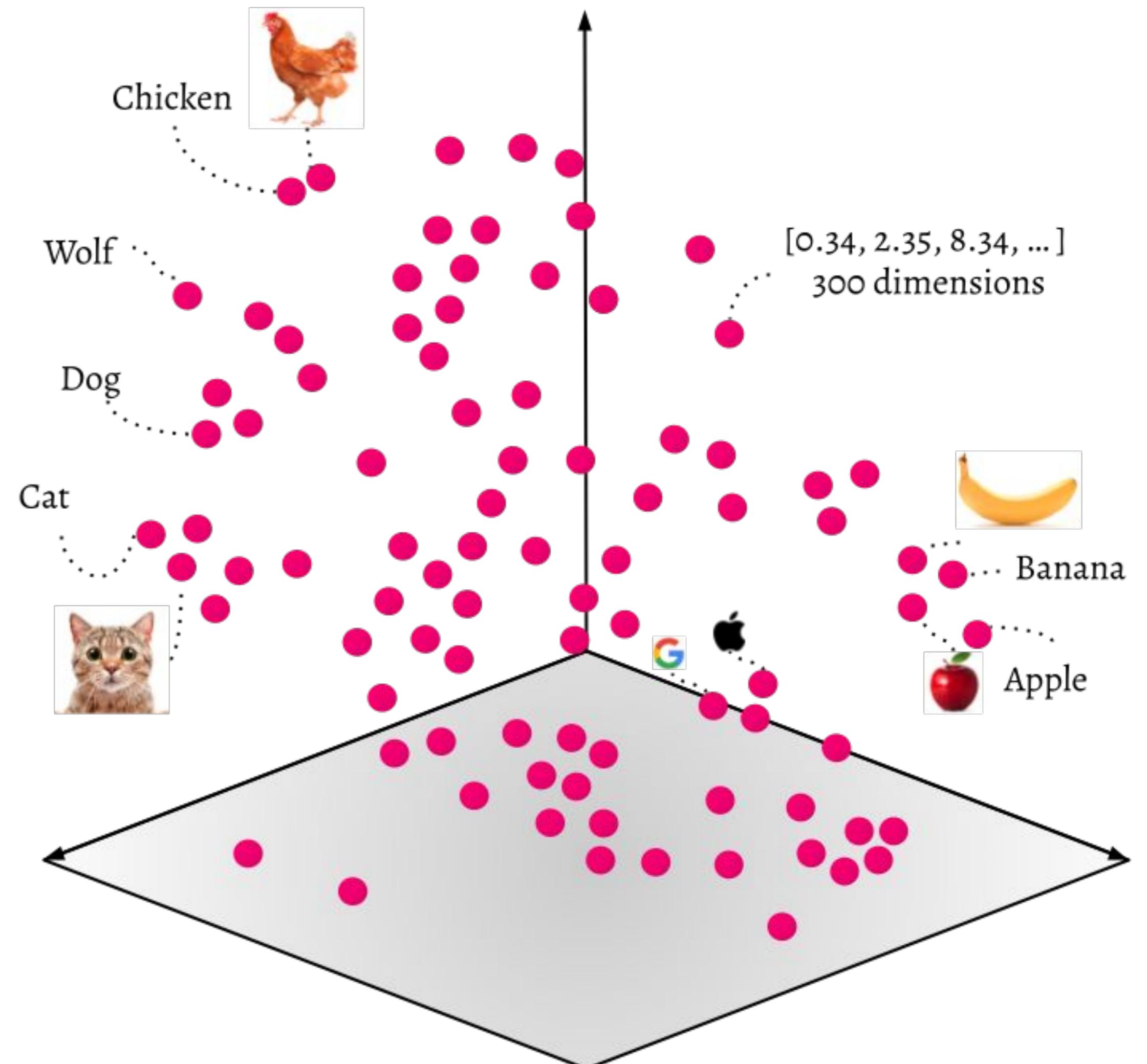


[Source](#)

# How does vector search work?

## 2. Automatically vectorize and index your data

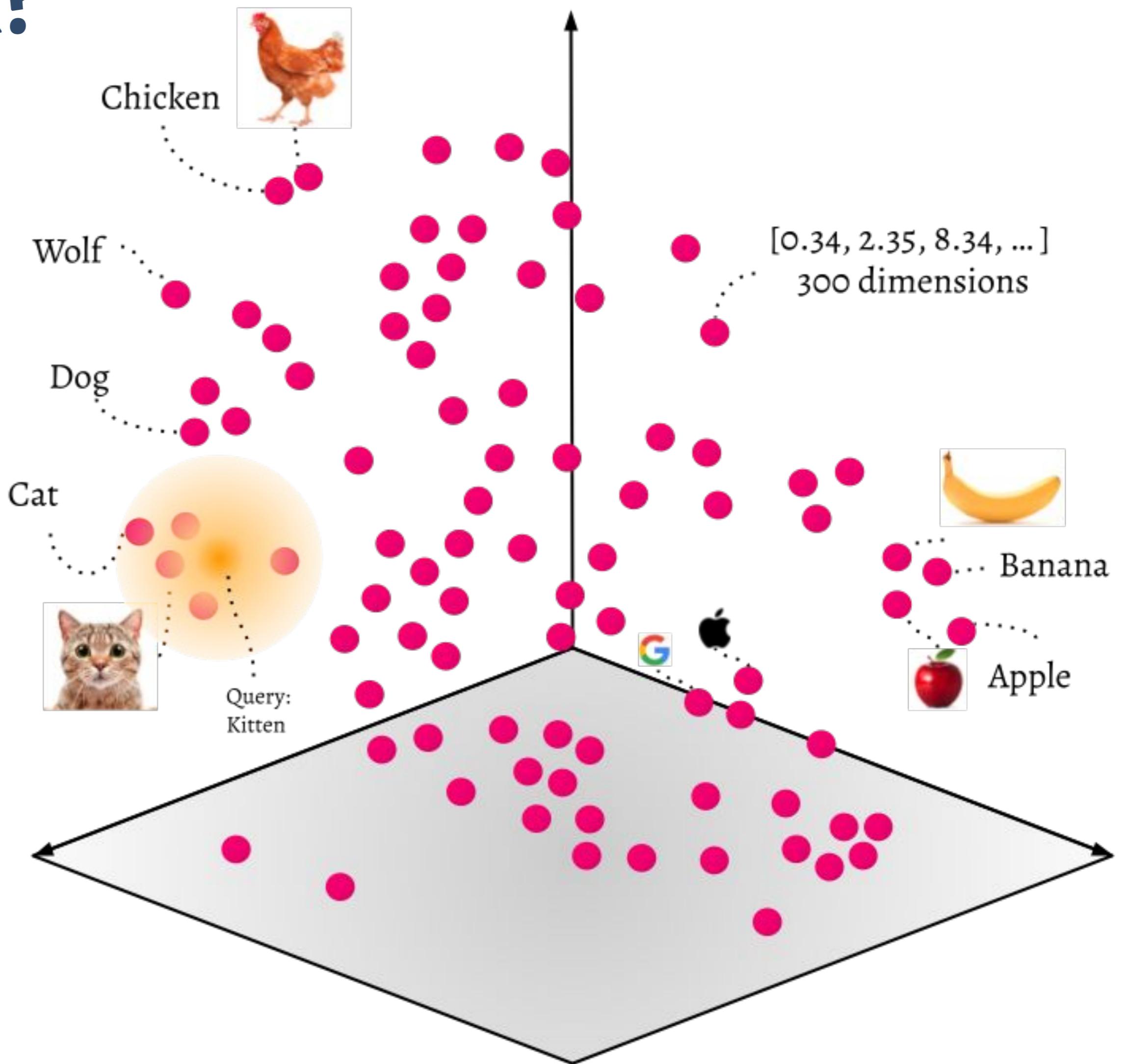
- When you import data, Weaviate looks at the data object and uses the ML model to vectorize the data
- Weaviate *understands* your data, it will be placed in the hyperspace (e.g. 300 dimensions)
- E.g. a *Cat* is closely related to *Dog*, *Animal* and the *image of a Cat*, but far away from *Apple* and *Banana*



# How does vector search work?

## 3. Search query

- Your search queries in natural language will also be vectorized and understood by the machine learning module of Weaviate (*retriever model*)
- It is placed close to the words and data objects that are semantically related to the query
- E.g. "Kitten" is semantically close to "Cat"

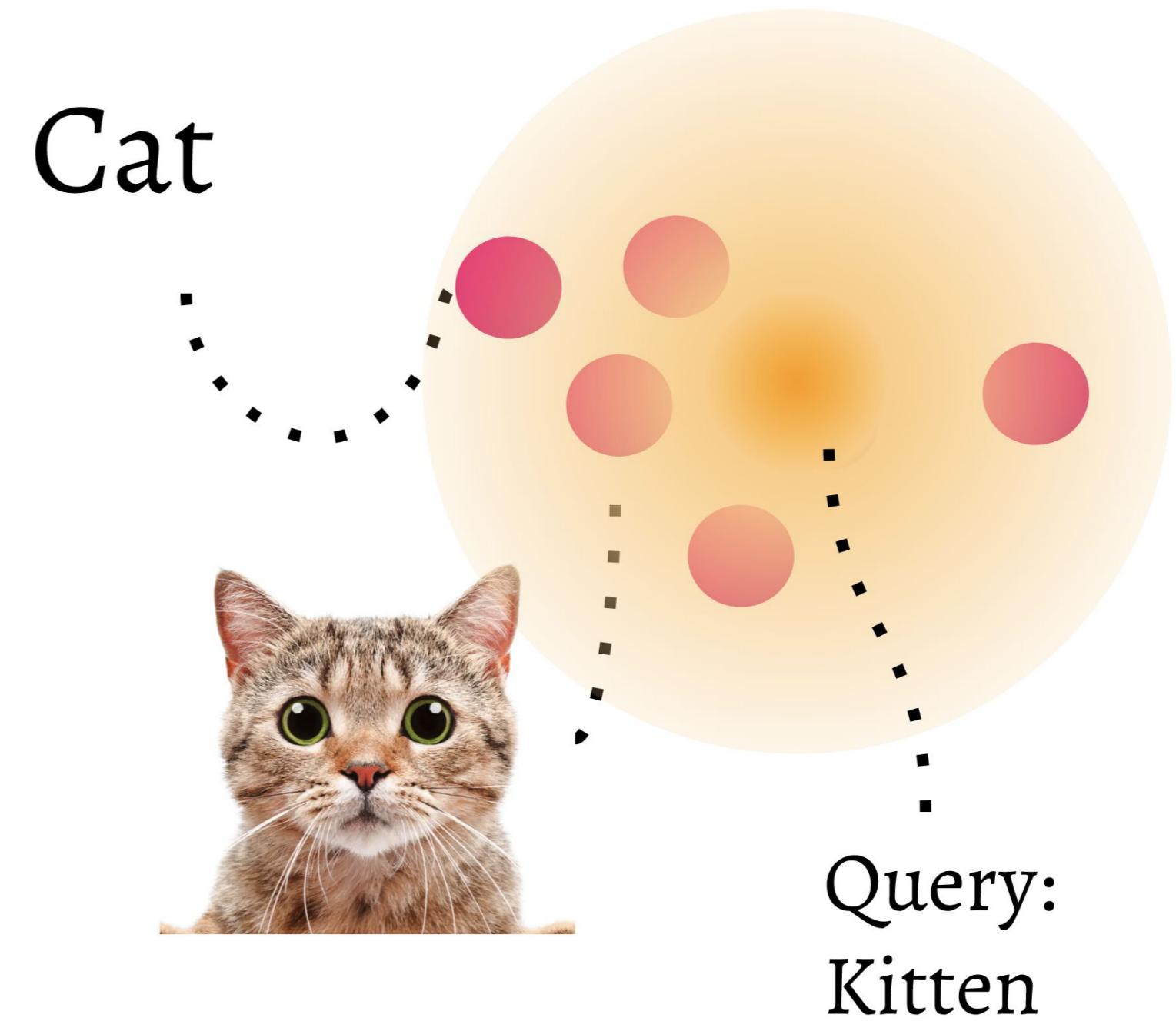


# How does vector search work?

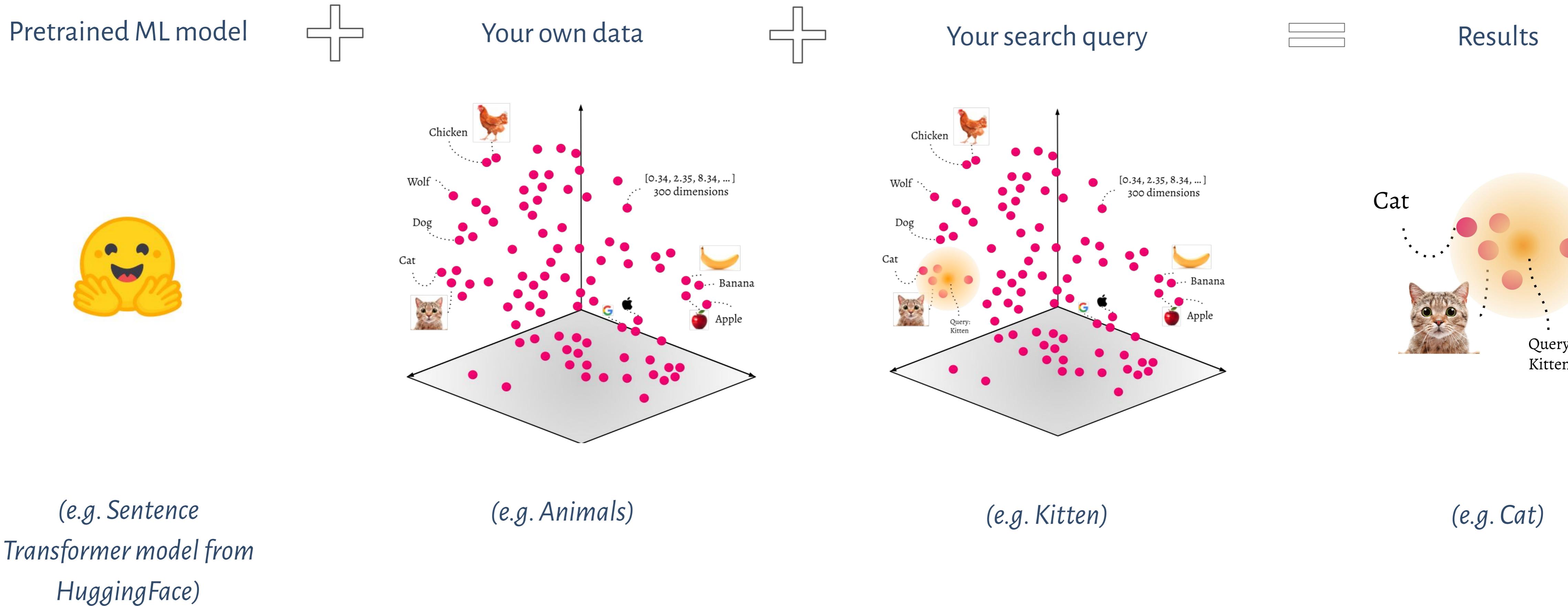
---

## 4. Results

- The data objects that are closest to the search query are retrieved from the dataset
- ANN (Approximate Nearest Neighbor) search, using e.g. cosine distance
- Efficient & fast retrieval using HNSW



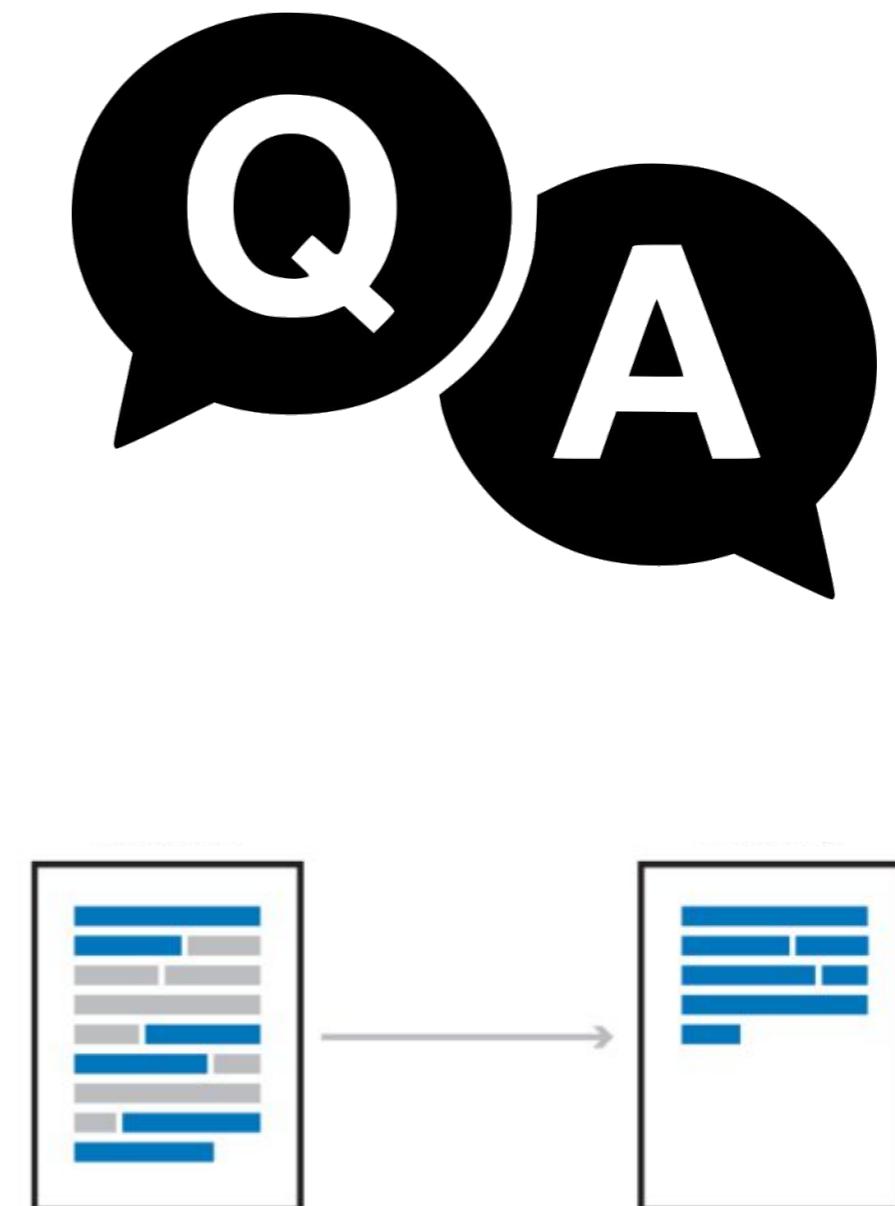
# How does vector search work?



# And there's more!

---

- **Retriever** models are used with an ANN algorithm to retrieve relevant data items
- You can extend the vector search pipeline by **Reader** or **Generator** models
  - **Reader** models extract information from the retrieved data objects. For example
    - Question Answering
    - Named Entity Recognition
  - **Generator** models use language generation to generate an answer from the retrieved data objects. For example
    - Summarization



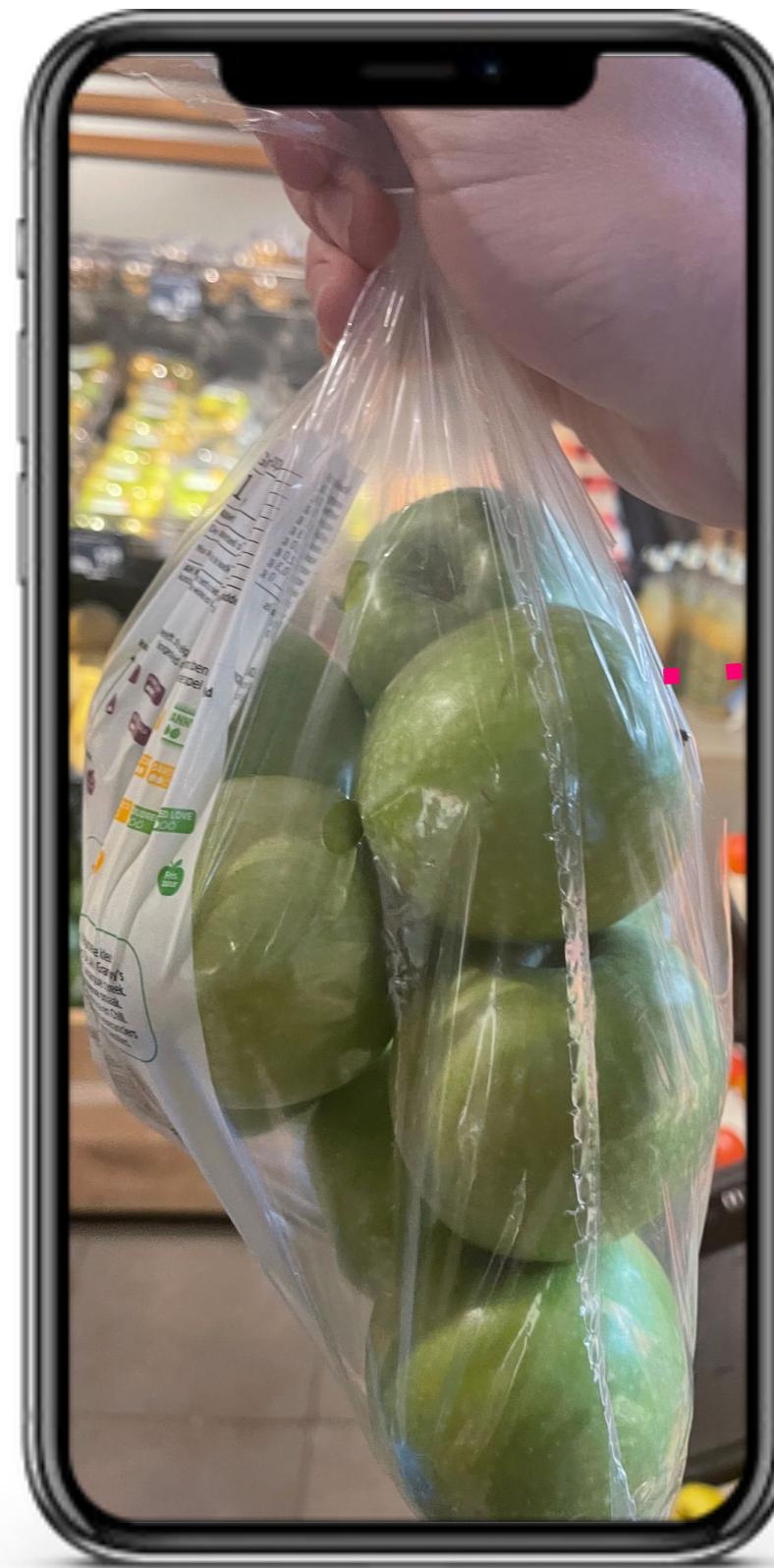
# How do I interact with the vector database Weaviate?

---

API endpoints:

- RESTful API endpoints for CRUD operations
- **GraphQL API** for intuitive querying, e.g.
  - Retrieve data objects (*Get*)
  - Semantic search (*nearText* argument)
  - Question Answering (*ask* argument)
- Demo time!
  - [~3500 news articles](#)
  - [Complete English Wikipedia](#)

# Mix media-types within Weaviate (multi-modal search)



Weaviate is able to mix multiple media-types including (but not limited to) images and text

The screenshot shows the Weaviate web interface. At the top, there are navigation icons and a "Log Out" button. Below that is a toolbar with buttons for "Prettify", "Merge", "Copy", "History", and "Share this query". The main area displays a GraphQL query on the left and its results on the right. A pink arrow points from the image on the smartphone to the "Photo" field in the query results. Another pink arrow points from the text description in the query results to the text "GRANNY SMITH APPLES FRESH PRODUCE FRUIT 3 POUND BAG" in the results pane.

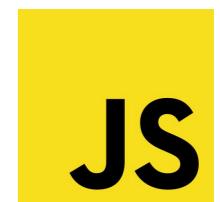
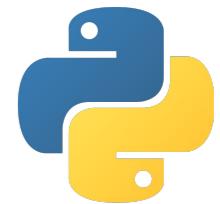
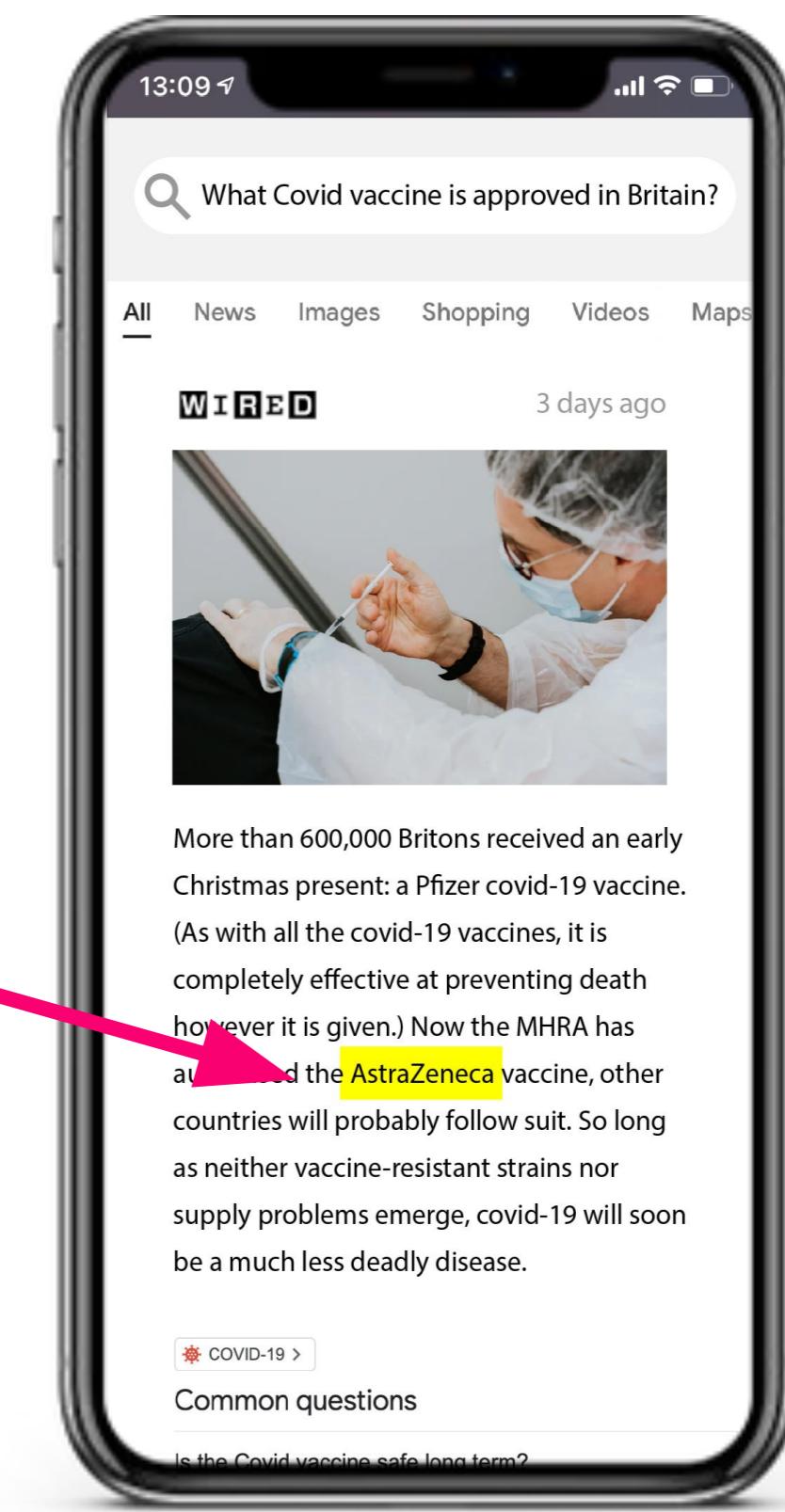
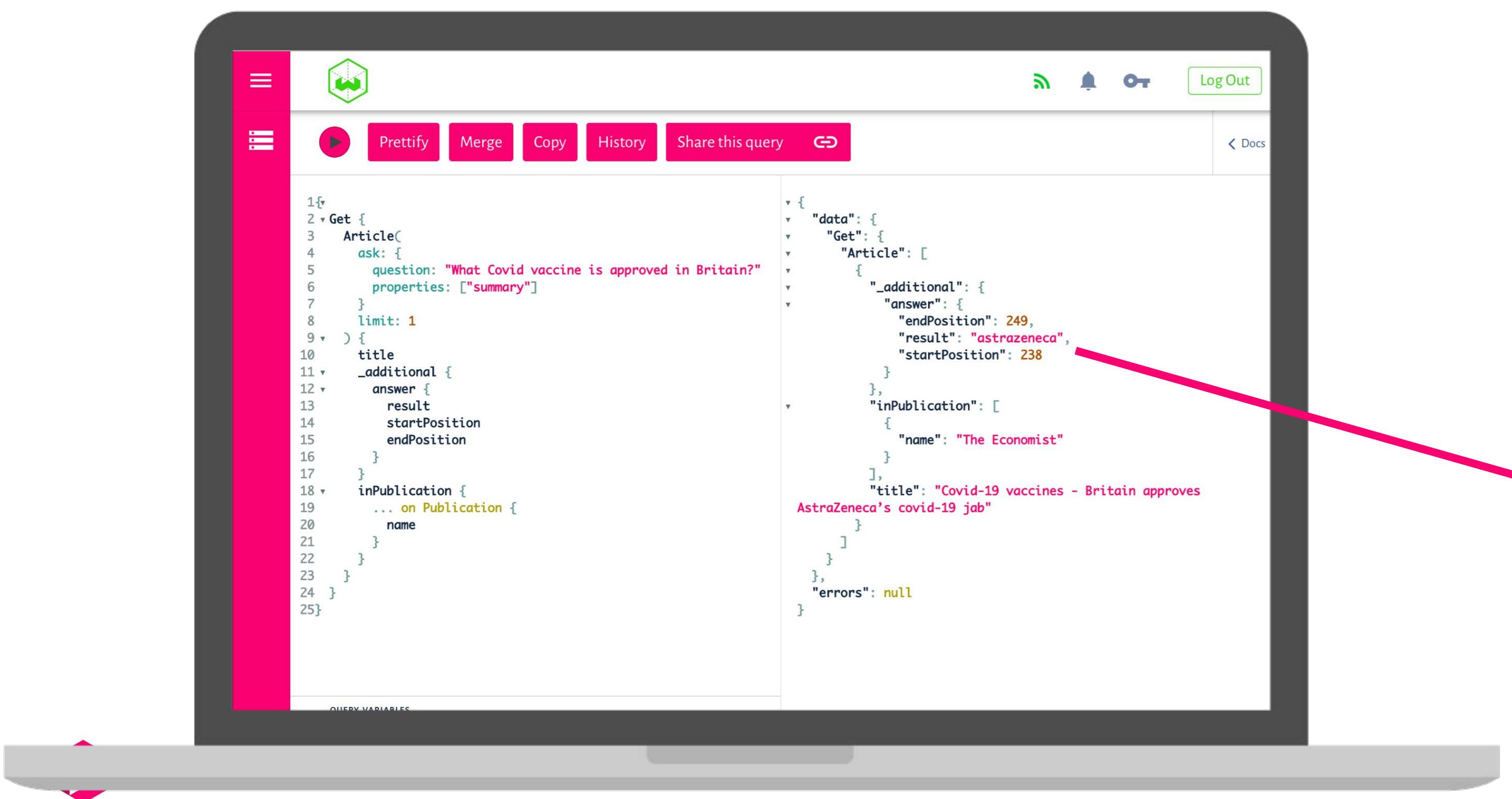
```
1 <|> {  
2   Get {  
3     Photo{  
4       nearObject:{  
5         id: "013d889b-4d56-3fe8-a82c-9e1c10b922e7"  
6         certainty: 1.0  
7       }  
8     }  
9     ofProduct {  
10       ... on Product {  
11         name  
12       }  
13     }  
14     _additional {  
15       certainty  
16     }  
17   }  
18 }  
19 }
```

data: {  
Get: {  
Photo: [  
{  
\_additional: {  
certainty: 0.76  
},  
ofProduct: [  
{  
name: "GRANNY SMITH APPLES FRESH  
PRODUCE FRUIT 3 POUND BAG"  
}  
]  
}  
]  
}

# Discovery within Weaviate (Question Answering)

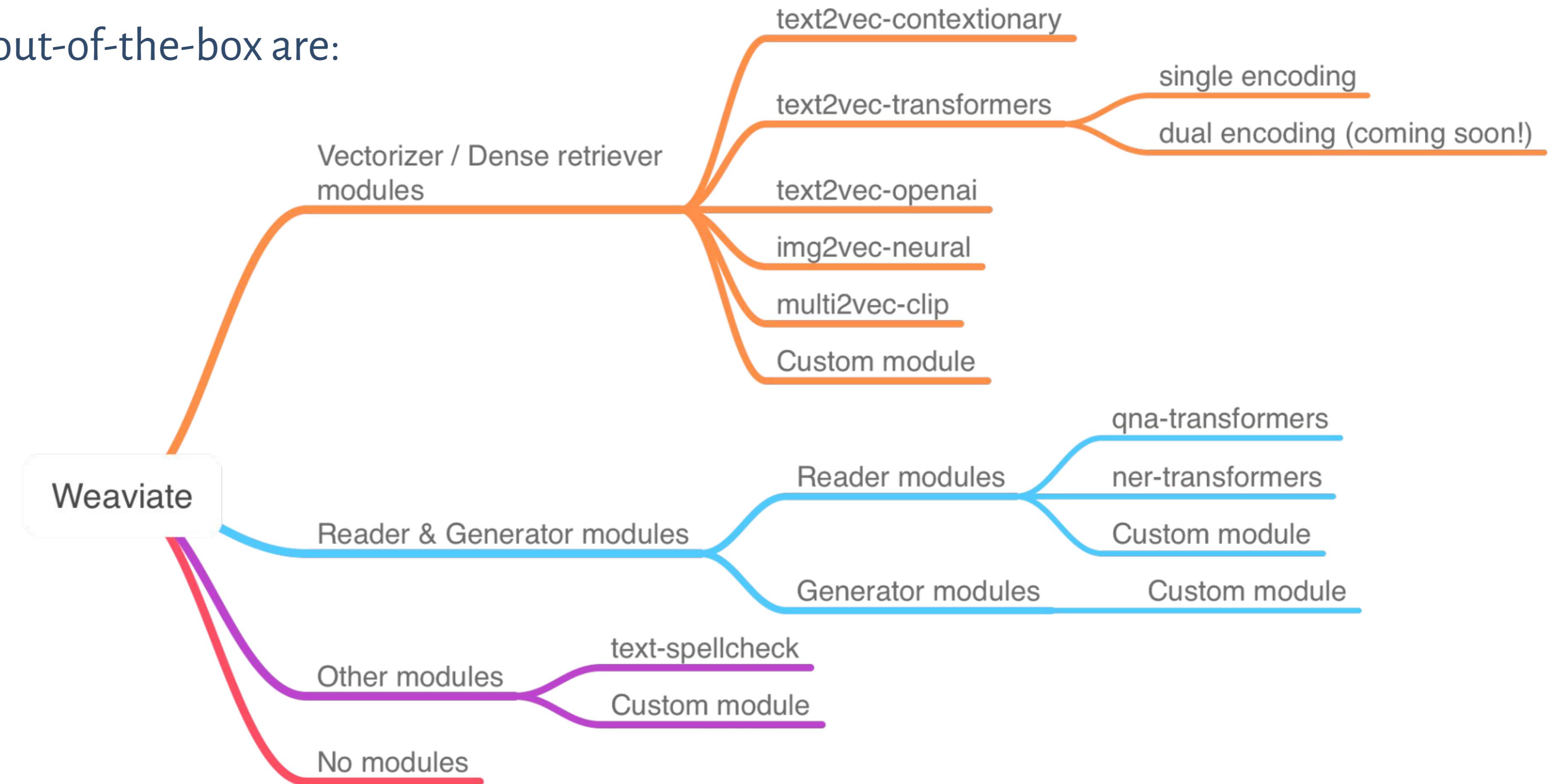
In the example below you see how Weaviate's API can be used to run **question answering** queries on a news article dataset. Weaviate show the answer to a question and the article that contains the answer. This example is also available as a live demo.

Weaviate can be integrated into almost any platform thanks to its APIs. For example in a web-app.



# Custom ML models

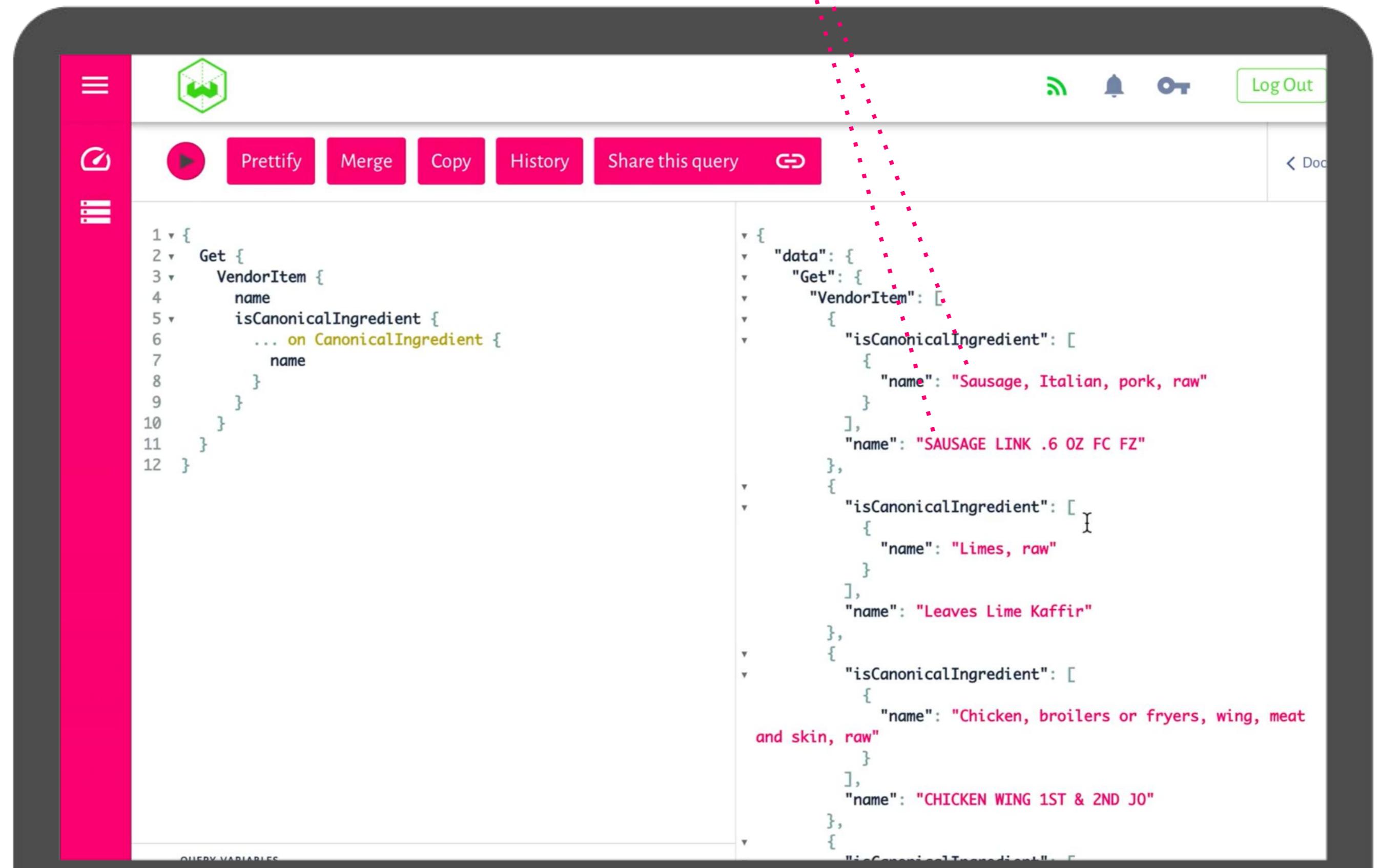
- Available models out-of-the-box are:



# Classification within Weaviate

Weaviate allows automatic classification based on your data. You can classify based on existing data (e.g., data object A is similar to data object B, **kNN classification**) or based on the meaning of the data (e.g., the "Product Fuji apple" classifies as "fruit", **zero-shot classification**).

The line item are automatically classified to generic data definitions



The screenshot shows the Weaviate web interface. On the left, a code editor displays a GraphQL query:

```
1 v {  
2   Get {  
3     VendorItem {  
4       name  
5       isCanonicalIngredient {  
6         ... on CanonicalIngredient {  
7           name  
8         }  
9       }  
10     }  
11   }  
12 }
```

On the right, the results of the query are shown in a tree-view format. The results are as follows:

- VendorItem:
  - isCanonicalIngredient:
    - name: "Sausage, Italian, pork, raw"
    - name: "SAUSAGE LINK .6 OZ FC FZ"
    - name: "Limes, raw"
    - name: "Leaves Lime Kaffir"
    - name: "Chicken, broilers or fryers, wing, meat and skin, raw"
    - name: "CHICKEN WING 1ST & 2ND JO"



# Questions?



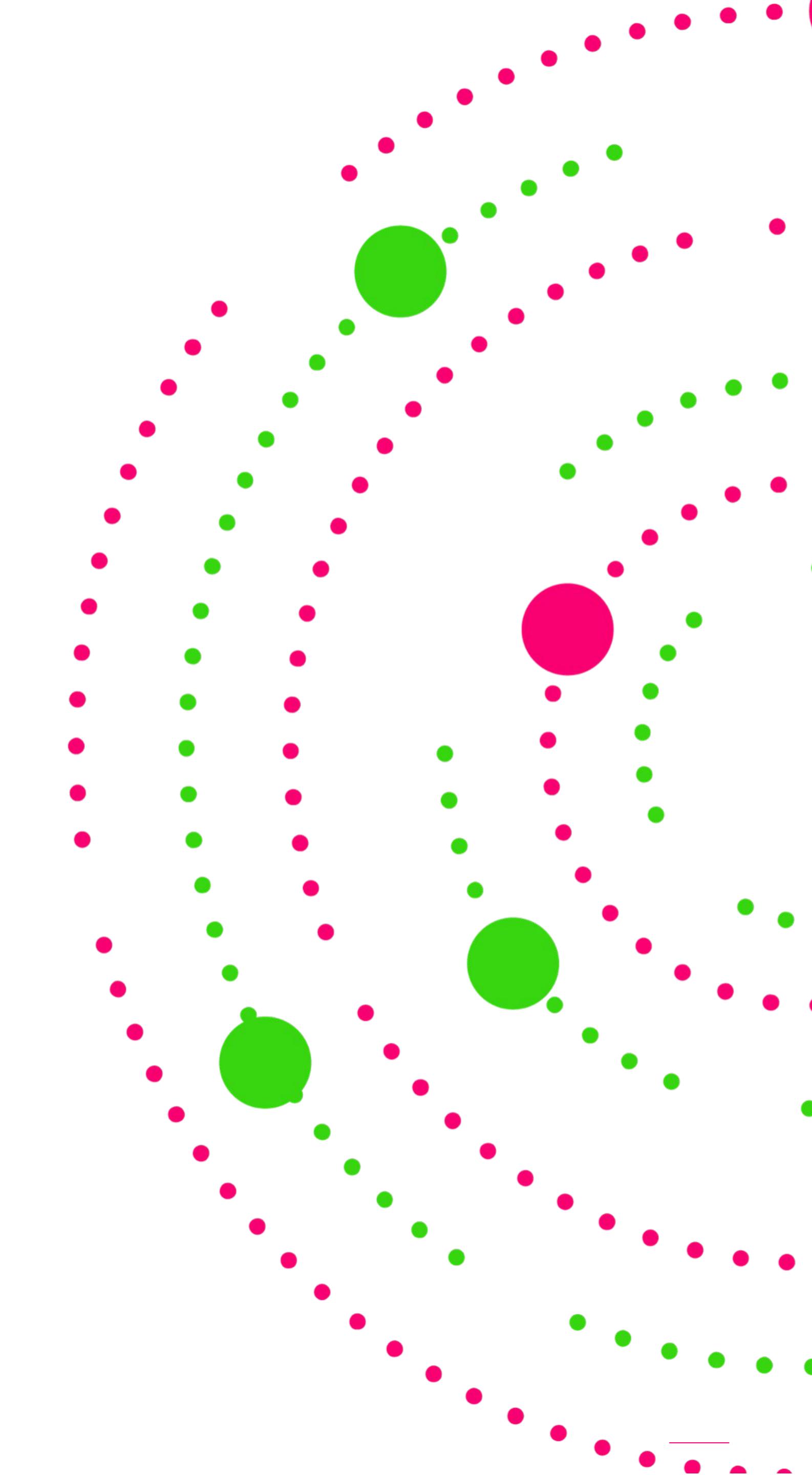
**Laura Ham**

Email: [laura@semi.technology](mailto:laura@semi.technology)

Join our **Slack channel** for  
communication and questions  
<https://link.semi.technology/3vbEtKN>



<https://weaviate.io/>



---

# Additional slides & resources

# Weaviate: ML-first Vector Search Engine

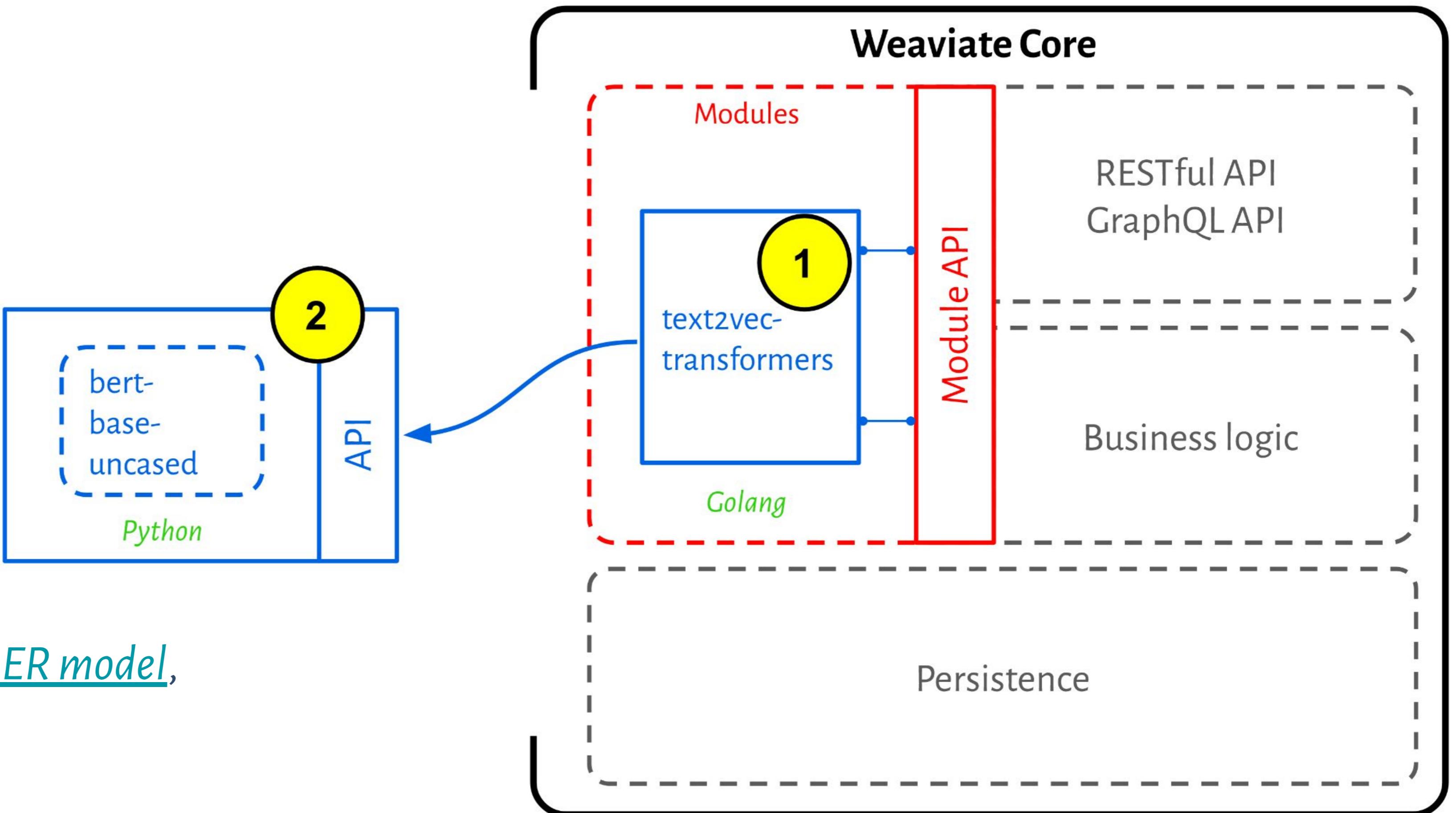
---

- Vector database and search engine with full CRUD support:
  - Data is stored at **vectors** (long arrays of numbers, aka coordinates in a high-dimensional space), allowing for **context-based** search and **automatic classification**
- Combine vector and scalar search
- Graph-connections between objects
- Fast queries (with RESTful and GraphQL interface)
- Horizontal Scalability
- Modular architecture:
  - Store any type of media with **vectorizer modules** (e.g. FastText, Transformers like Bert, SpaCy, ResNet, etc)
  - Extend capabilities with **any** other **ML/NLP** module (e.g. Q&A, spellcheck, NER, etc)
  - Fully customizable setup

# Modular architecture

1. Weaviate module
  - a. Written in Go
  - b. GraphQL design
  - c. E.g. [NER module in Weaviate](#)

2. Inference service
  - a. (Containerized) application
  - b. Wraps an ML model
  - c. E.g. [HuggingFace Transformer NER model, example app script](#)



# How to get started?

---

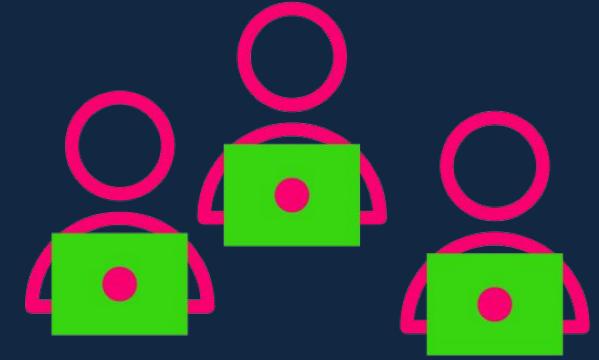
- Weaviate introduction ([link](#))
- Getting started guide ([link](#))
- Videos ([link](#))
- Tutorials ([link](#))
  - Google Colabs ([link](#))
- Use client libraries (Python, JS, Go, Java)
- [Weaviate Slack Channel](#)



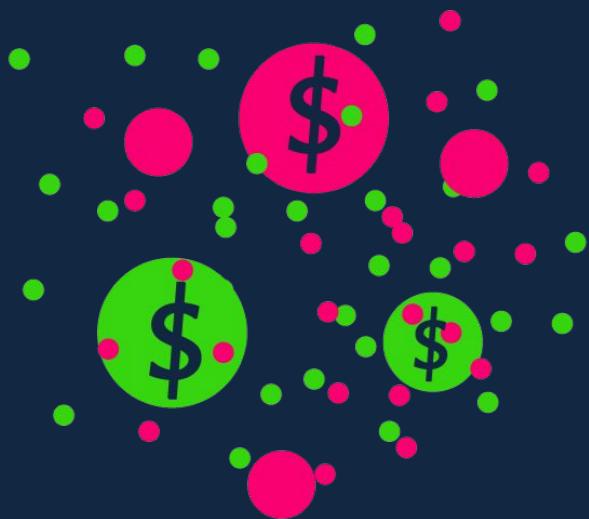
# Weaviate adds value in many industries

You can request an industry specific use-case presentation via [hello@semi.technology](mailto:hello@semi.technology) or [www.semi.technology](http://www.semi.technology)





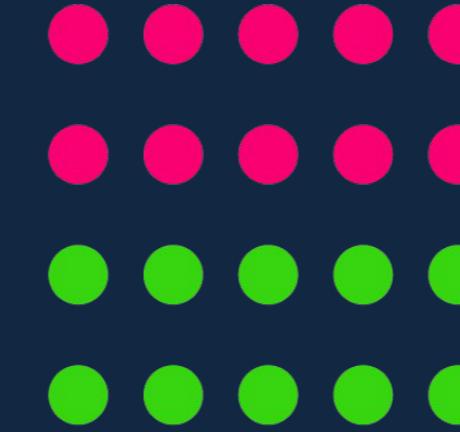
Most (unstructured) data is generated by humans.



The amount of data grows rapidly and especially the unstructured data contains the most value.



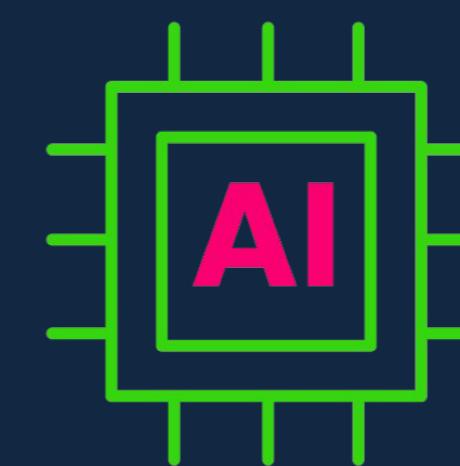
On top of this, you probably store data in many different silos.



Most existing database technology is focussed on structured data.

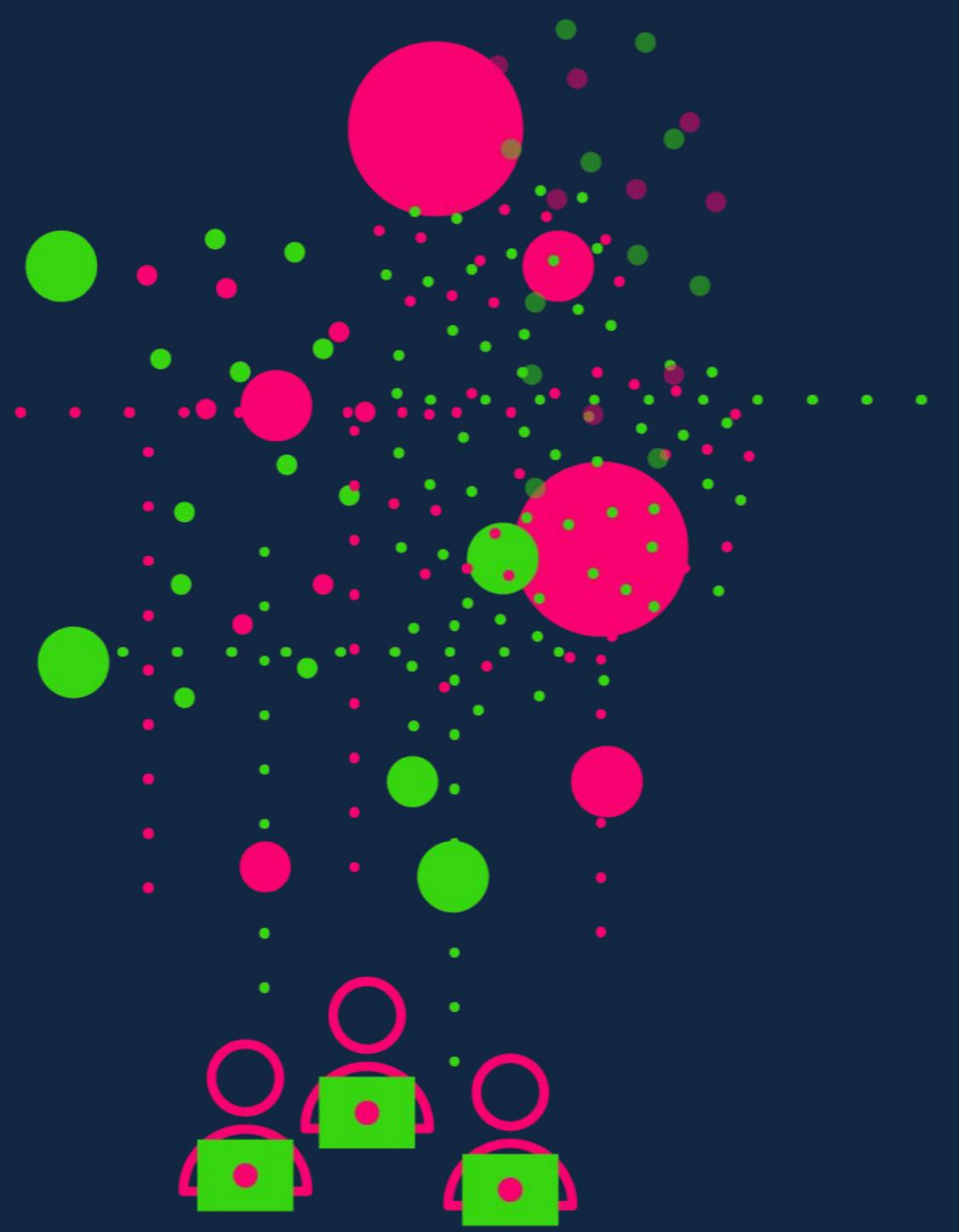


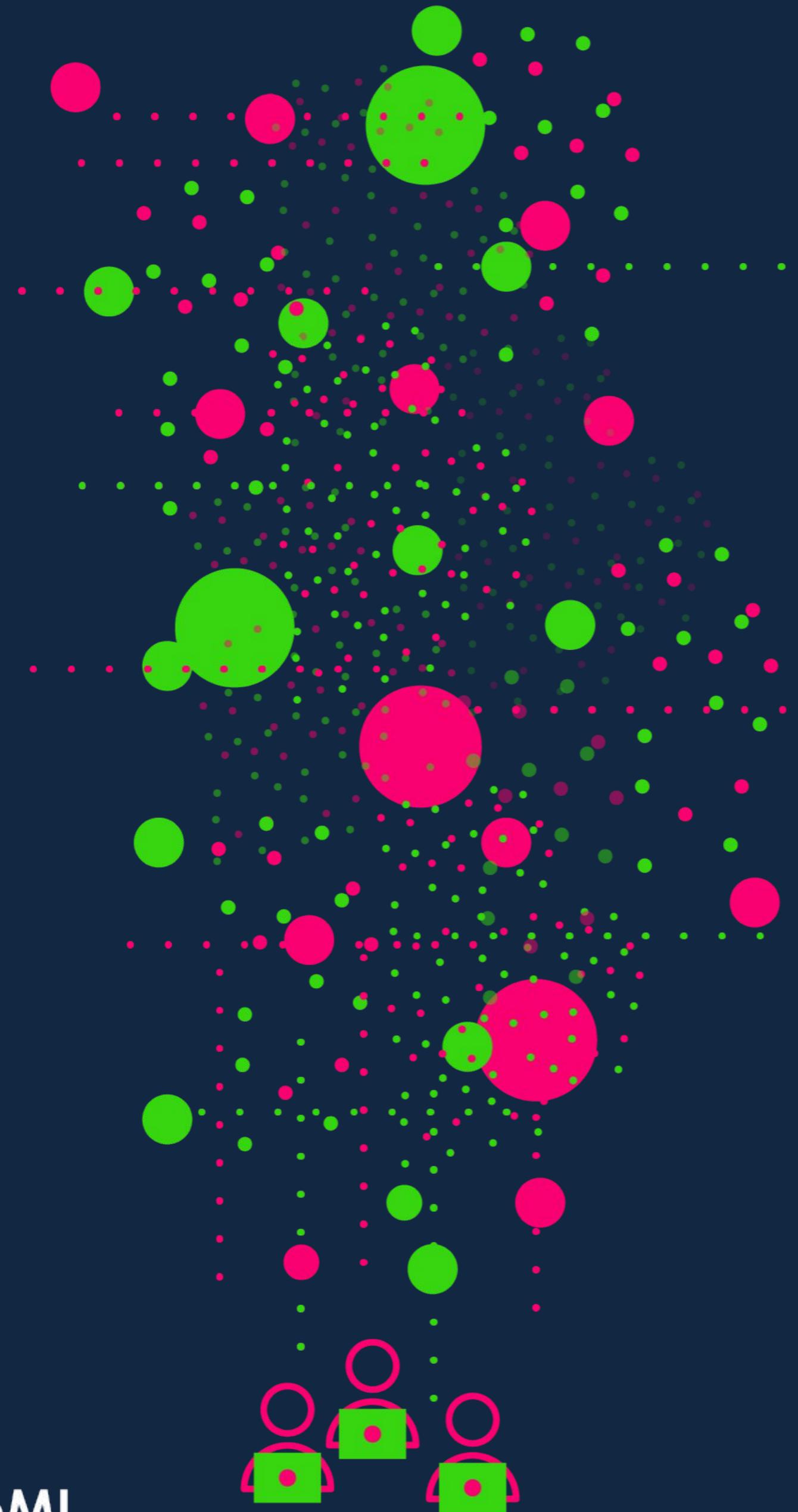
It's hard for machines to “understand” what your unstructured data means.



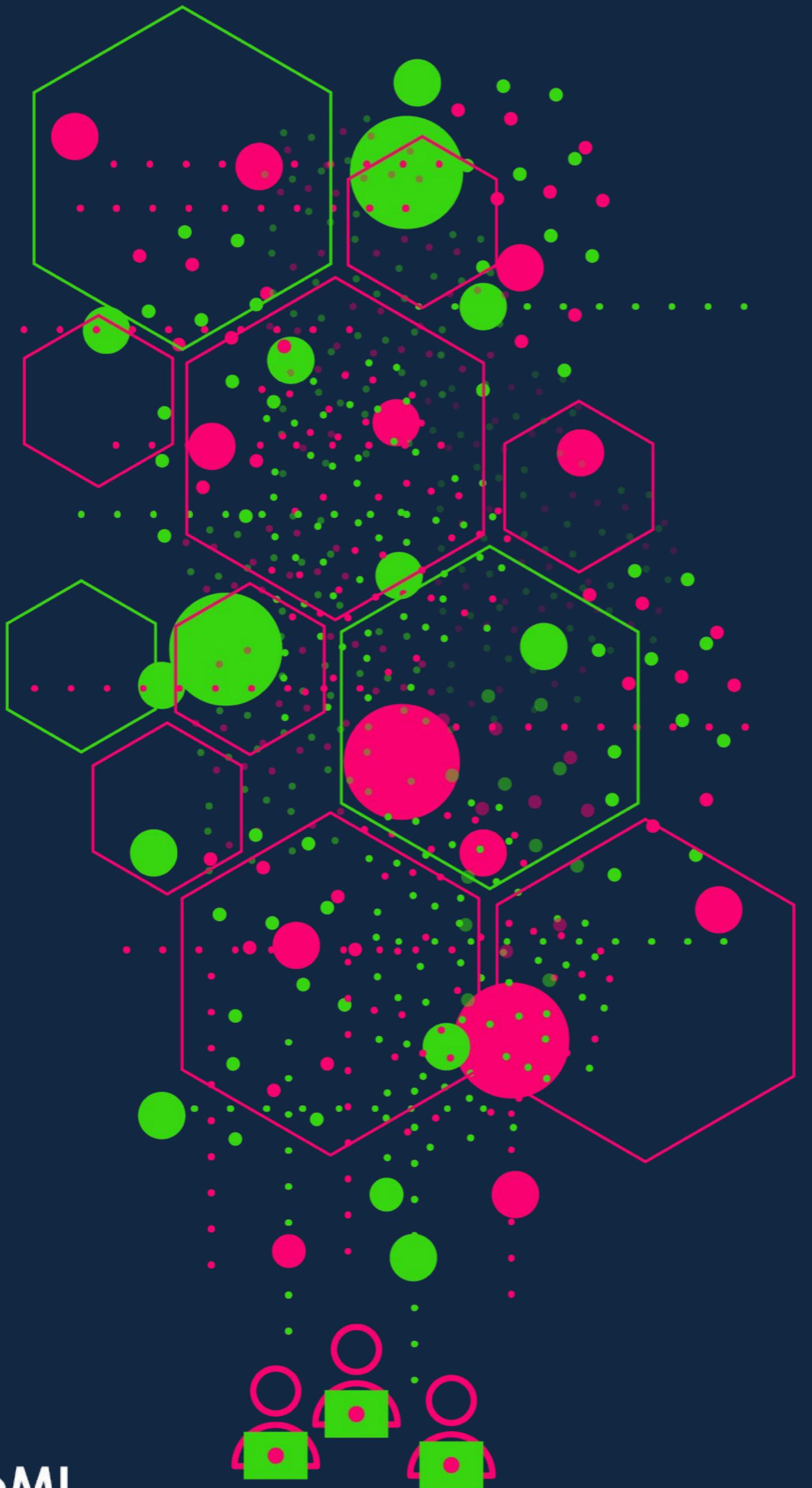
Machine learning can help in organizing this data, but processes to implement this are complex and costly.

- Most (unstructured) data is generated by humans.



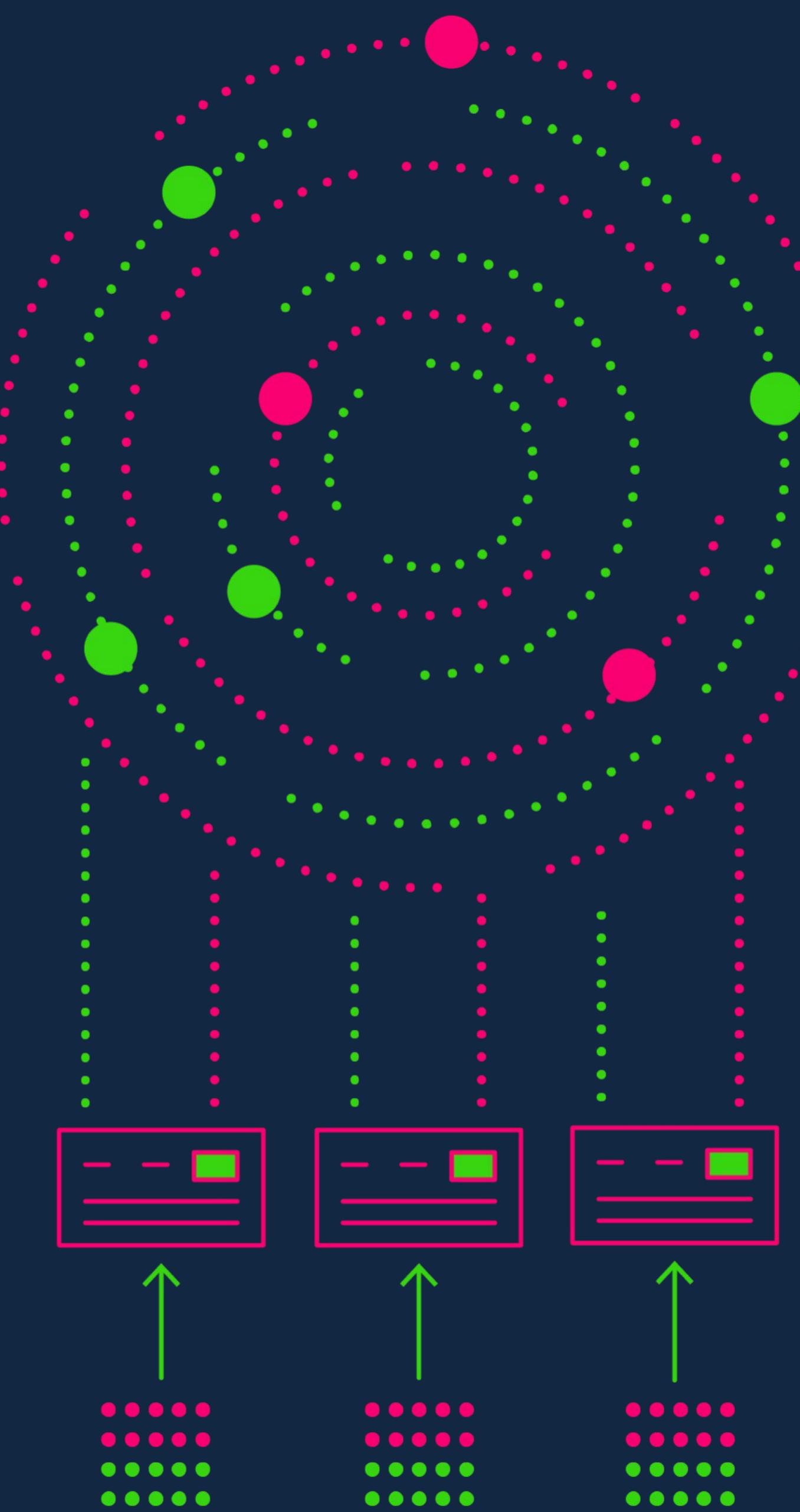


- Most (unstructured) data is generated by humans.
- The amount of data **grows rapidly** and especially the unstructured data contains the most value.



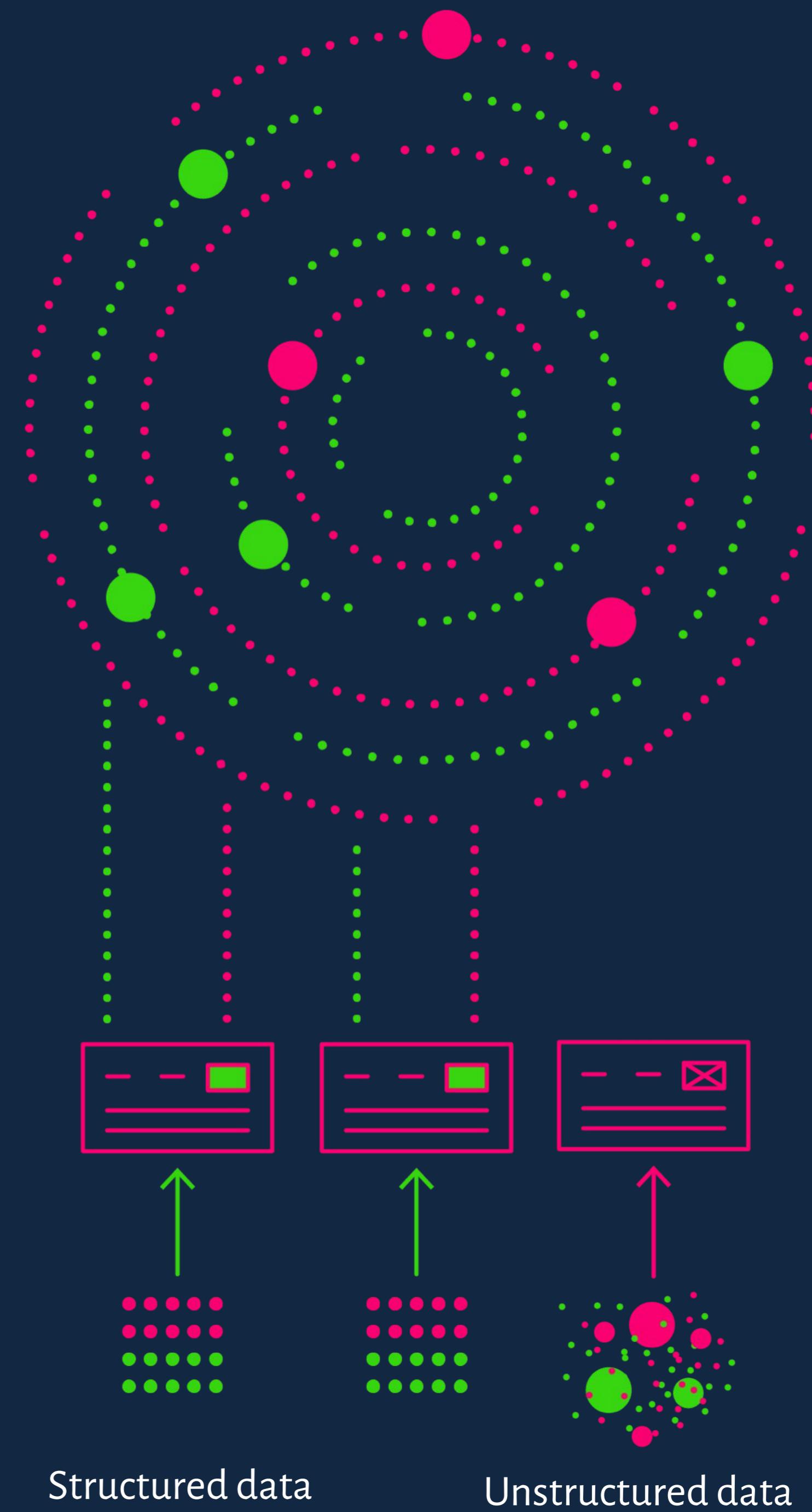
- Most (unstructured) data is generated by humans.
- The amount of data grows rapidly and especially the unstructured data contains the most value.
- On top of this, you probably store data in many different silos.

- Most existing database technology is focussed on structured data.



Structured data

- Most existing database technology is focussed on structured data.
- Machine's are very good at handling the structured part of data, but not the unstructured part. It's hard for machines to “understand” what your data means.



- Most existing database technology is focussed on structured data.
- Machine's are very good at handling the structured part of data, but not the unstructured part. It's hard for machines to “understand” what your data means.
- Machine learning is helping in organizing this data for you, but processes to implement this are often complex and costly because a lot of software needs to be created especially for your use case.

