



Weaviate

# Vector Search for Data Scientists

A Case Study with **Twitter Analytics**



# Common questions in Data Science

---

#1 - How is my data distributed?

#2 - Are there outliers in my data?

#3 - Are my variables correlated with each other?



# Vector Search

---

#1 - Can we capture the semantics in vector representations?

#2 - What can we learn about our data from semantic clusters?



# Social Media Clicks and Twitter Analytics

---

Connor Shorten  
@CShorten30

I just published "ANN Benchmarks with Etienne Dilocker -- Weaviate Podcast #16" on Medium! 🎉

This article breaks down the technical details to make the podcast more digestable and beginner friendly!

link:

connorshorten300.medium.com  
ANN Benchmarks with Etienne Dilocker—Weaviate podcast #16  
Written summaries of discussed topics in Approximate Nearest Neighbor (ANN) benchmarking! Billion-Scale Vector Search!

9:40 AM · May 27, 2022 · Twitter Web App

| | View Tweet analytics

3 Retweets 15 Likes

Connor Shorten  
@CShorten30

Approximate Nearest Neighbor algorithms allow us to Vector Search in massive datasets! 🔎

But which ANN configuration is right for your data? 🤔

Really happy to publish this podcast with @etiennedilocker discussing the new ANN benchmarks on @weaviate\_io!

youtube.com  
Weaviate Podcast #16 • ANN Benchmarks with Etienne Dil...  
ANN Benchmarks are a tool for evaluating the performance of in-memory approximate nearest neighbor algorithms. ...

9:13 AM · May 24, 2022 · Twitter Web App

| | View Tweet analytics

14 Retweets 50 Likes



# Twitter Analytics

## Tweet activity

Your Tweets earned **65.0K impressions** over this **28 day** period



Tweets

Top Tweets

Tweets and replies

Promoted

Impressions

Engagements

Engagement rate



Connor Shorten @CShorten30 · 24h

New Weaviate Podcast with Kyle Lo (@kylelostata)!! 🎧

Kyle is a leader in Scientific Literature Mining with projects such as TLDR, QASPER, CORD19, and many more! ↗

This is one of my favorite applications for Vector Search, I hope you enjoy the podcast:

[youtube.com/watch?v=kUjhCs...](https://youtube.com/watch?v=kUjhCs...)

[View Tweet activity](#)

Last 28 Days

Export data

YOUR TWEETS  
During this 28 day period, you earned **2.3K impressions** per day.

## Engagements

Showing 28 days with daily frequency

Engagement rate

**3.0%**

Jun 1  
**2.6% engagement rate**



Link clicks  
**145**

Jun 1  
**2 link clicks**

# Twitter Analytics CSV Data

---

Tweet Text	Time	Impressions	Engagements	Engagement Rate	Retweets	Replies	Likes	User Profile Clicks	Url Clicks
I just published "ANN Benchmarks with Etienne Dilcoke -- Weaviate Podcast #16 on Medium..	May 27th, 1:34pm	1905	50	2.6%	3	1	15	2	18
Approximate Nearest Neighbor algorithms allow us to Vector Search in massive datasets! ...	May 24th, 1:13pm	7182	252	3.5%	14	1	50	27	36



**Feature Engineering:**  
Contains Emoji?  
Character Count?  
Word Count?  
Contains "Weaviate"?



# Key Takeaways: **“Vector Search for Data Scientists”**

Slides, Colab Notebook, Video Presentation available on:  
[github.com/CShorten/Vector-Search-for-Data-Scientists](https://github.com/CShorten/Vector-Search-for-Data-Scientists)

1. Segmentation in Data Science
2. Vector Representations of Data
3. Vector Segmentation
4. Weaviate for Twitter Analytics
5. Research Questions and Discussion

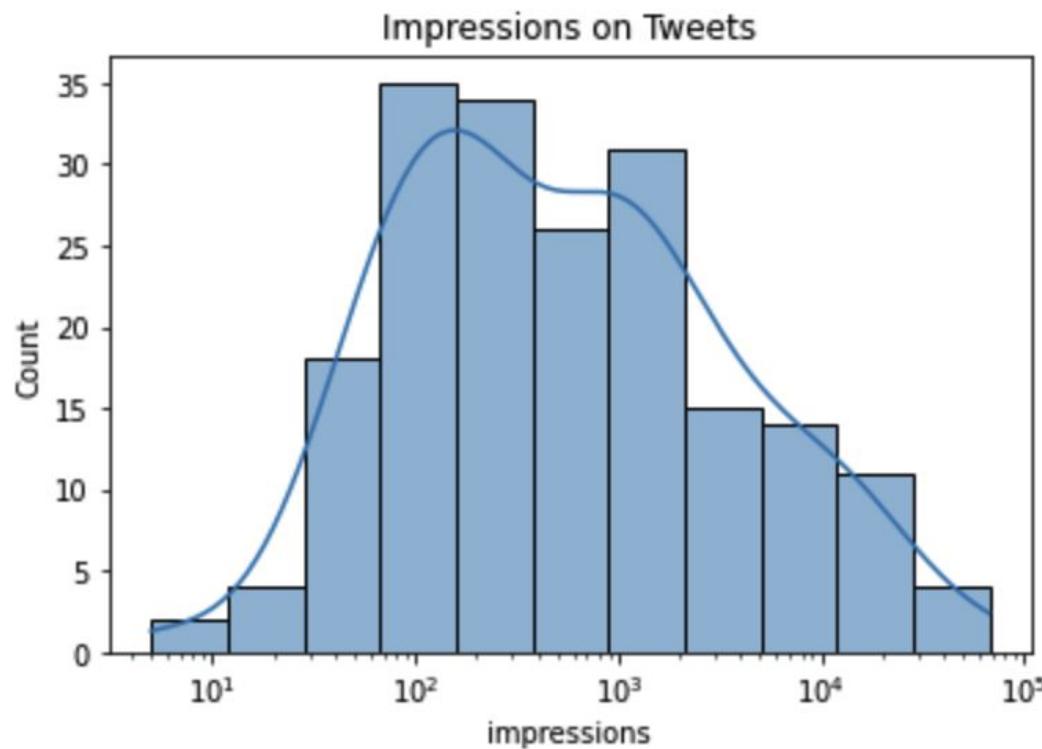


# Key Takeaway #1 - **Segmentation in Data Science**



# Visualizing Distributions of Values

---

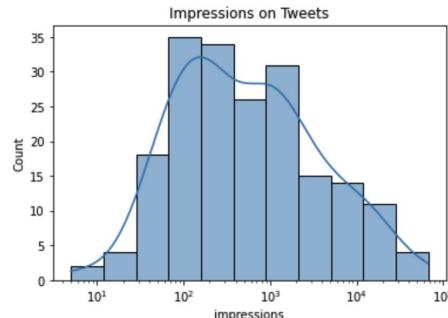


# Segmentation in Data Science

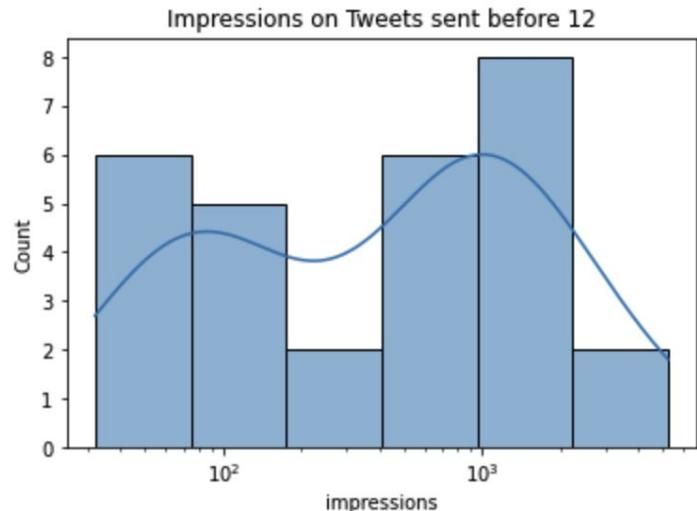
- What Time was the Tweet sent?
- Is there a URL Link in the Tweet?
- Symbolic vs. Vector Segmentation



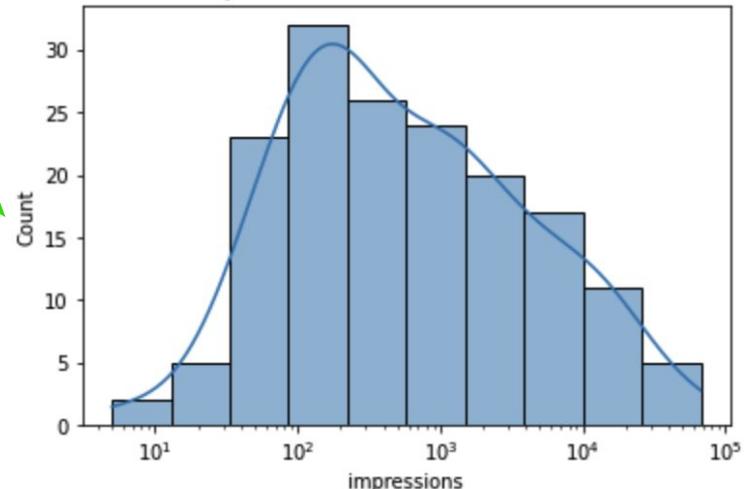
# What Time was the Tweet sent?



Impressions on Tweets sent before 12

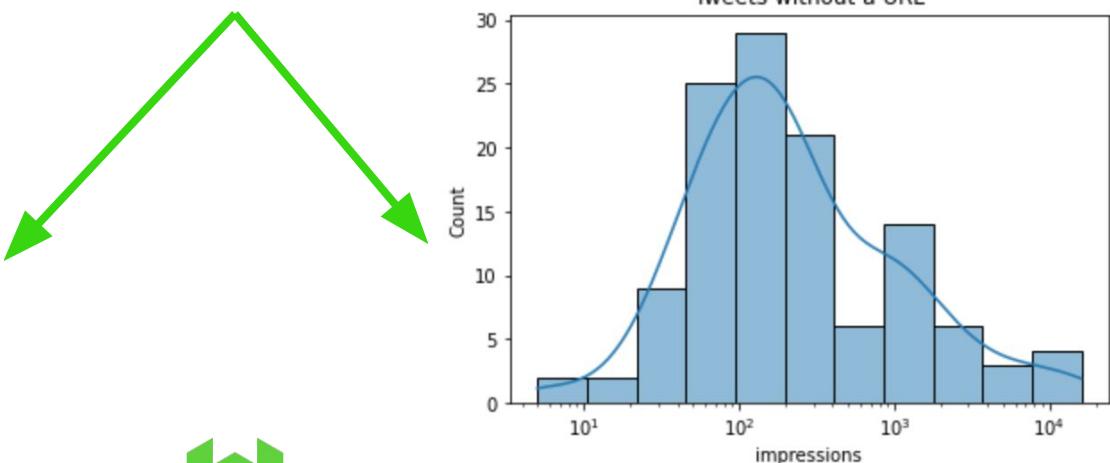
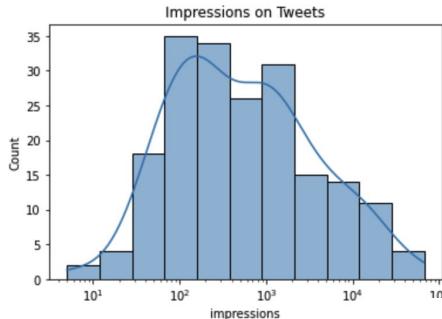
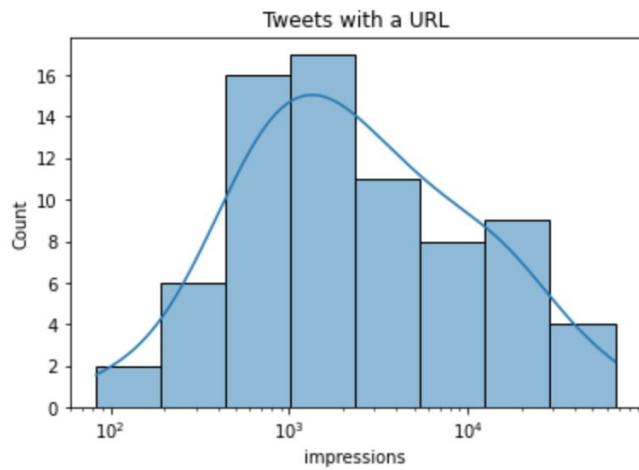


Impressions on Tweets sent after 12



# Is there a URL Link in the Tweet?

---



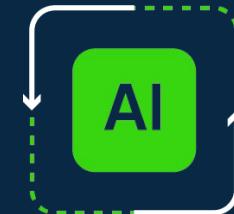
# Can we split Impressions based on the Semantics of the content?



Weaviate Podcast



Weaviate Tutorial



AI Weekly Update



# How can we segment analytics based on the semantics of...

- Text
- Images
- Code
- Audio
- Video
- Graph-Structure
- Biological Sequences
- ... !



# Summary of Takeaway #1 **Segmentation in Data Science**

We visualize the **Distribution** of our data to get a sense of it.

For example we see that **Impressions** are somewhat **Normally Distributed**.

Is that also true for Tweets sent at **3 AM**?

What about Tweets related to **Deep Learning for Robotics**?



# Key Takeaway #2 - **Vector Representations of Data**



# Symbols compared to Vectors

---

## Symbols

Category - [0, 1, 0, 0, 0, 0]

Numeric - 52

Boolean - True

## Vectors

[0.1, 0.8, 0.34, 0.8, ... 0.2]



# Vector Representations of Data

---



Photo by [Shayna Douglas](#) on [Unsplash](#)

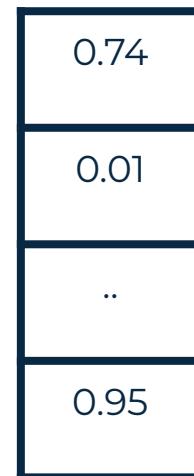


Photo by [Bill Stephan](#) on [Unsplash](#)



# Are these puppies similar? Let's ask Vector Distance!

$$L_2 \text{ Distance} = \sum || a_i - b_i ||^2$$

Vector Name	Value 1	Value 2	Value 3
Puppy1	4	8	10
Puppy2	2	9	11
Airplane	1	20	20

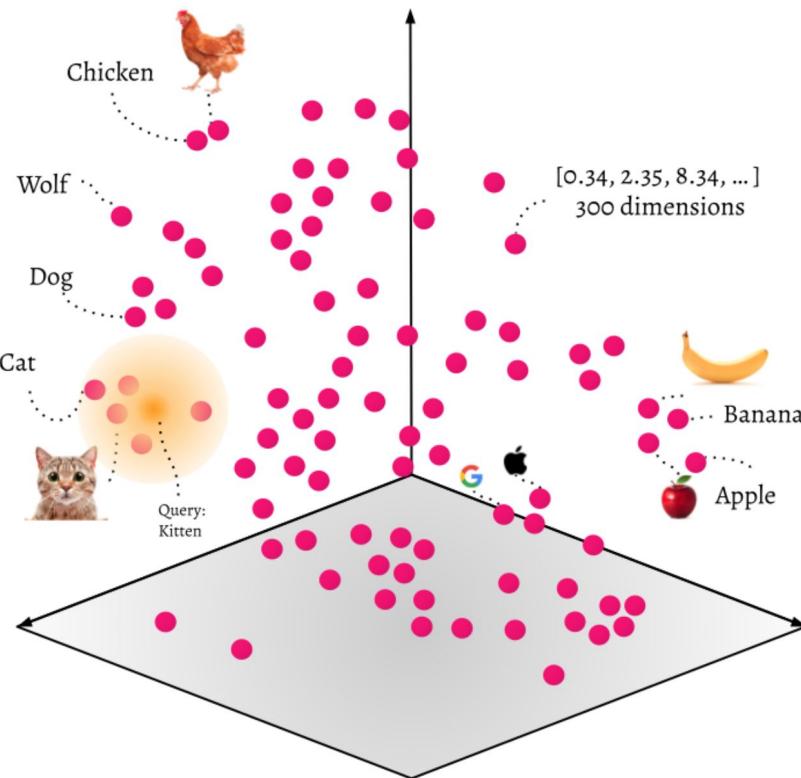
$$L_2 \text{ Distance} (\text{Puppy1}, \text{Puppy2}) = (4-2)^2 + (8-9)^2 + (10-11)^2 = 6$$

$$L_2 \text{ Distance} (\text{Puppy1}, \text{Airplane}) = (4-1)^2 + (8-20)^2 + (10-20)^2 = 253$$

6 << 253, Puppy1 is thus much more semantically similar to Puppy2 than Airplane

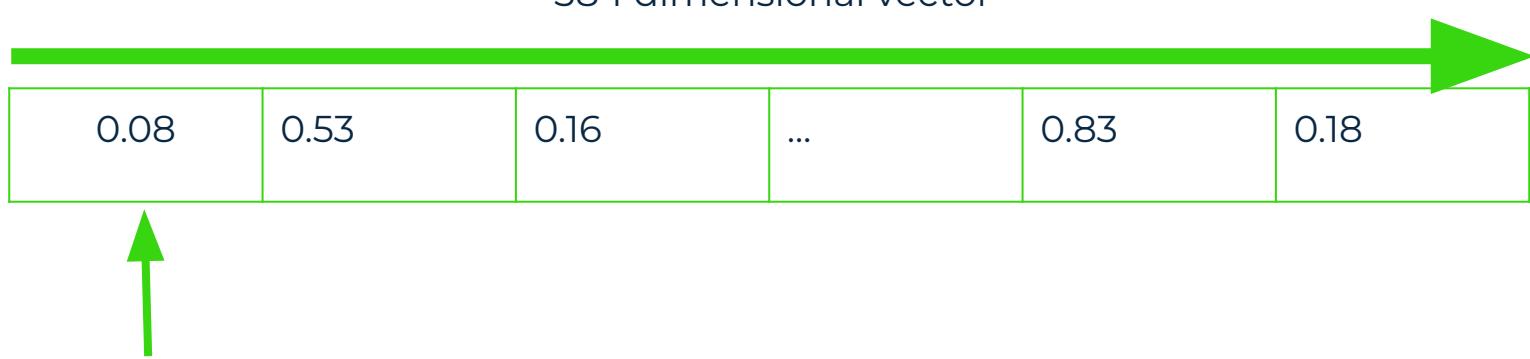
# Capturing Semantics in Vector Representations

---



# How do Vectors represent real-world objects?

---



Does this represent how much of a “brand” this is?

We aren't sure! But there are research fields such as “Multimodal Neurons” from OpenAI, and the general field of **Disentangled Representation Learning** that are making great strides in understanding this.



# Can we compress these vectors?

---



Sometimes!

Ideas like Binary Passage Retrieval (shown above) - fp32 to Binary values

Ideas like Product Quantization - 384-d vector mapped to 32-d



# Semantic Similarity with Vector Representations

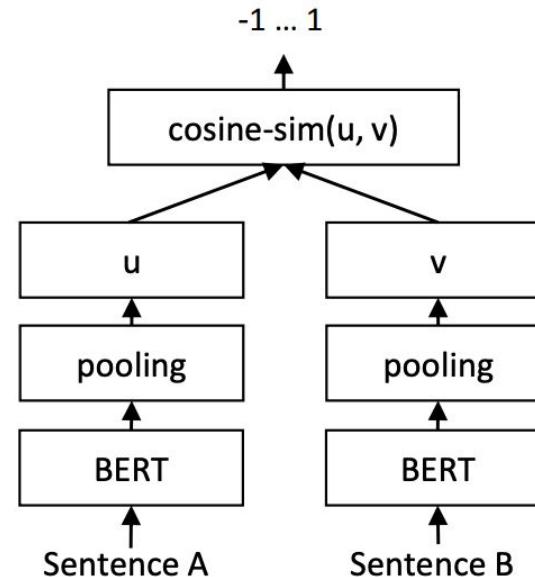
---

Sentence-BERT:  
Sentence Embeddings  
using Siamese  
BERT-Networks

**Authored by**

Nils Reimers and Iryna  
Gurevych

Published 2019





# Positive and Negative Pair Sampling

WIKIPEDIA  
The Free Encyclopedia

## Query Point

The **Miami Heat** are an American professional basketball team based in [Miami](#). The Heat compete in the [National Basketball Association](#) (NBA) as a member of the league's [Eastern Conference Southeast Division](#). The club plays its home games at [FTX Arena](#), and has won three [NBA championships](#).

The franchise began play in the [1988–89 season](#) as an [expansion team](#). After a period of mediocrity, the Heat gained relevance in the mid-1990s when [Pat Riley](#) became team president and head coach. Riley constructed the trades of [Alonzo Mourning](#) and [Tim Hardaway](#), which propelled the team into [playoff](#) contention. Mourning and Hardaway led the Heat to four consecutive division titles prior to their departures in 2001 and 2002, respectively. The team also experienced success after drafting [Dwyane Wade](#) in 2003.



# Positive (Semantically Similar)

---

Led by Wade and, following a trade for former NBA Most Valuable Player (MVP) Shaquille O'Neal, the Heat won their first NBA title in 2006, after Riley named himself head coach for a second stint. After the departure of O'Neal two years later, the team struggled for the remainder of the 2000s. Riley remained team president, but was replaced as head coach by Erik Spoelstra. In 2010, the Heat signed former league MVP LeBron James and NBA All-Star Chris Bosh, creating the "Big Three" along with Wade. During their four years together, Spoelstra, James, Wade, and Bosh led the Heat to the NBA Finals in every season, culminating in back-to-back championships in 2012 and 2013. All three departed by 2016, and the team entered a period of rebuilding. After acquiring All-Star Jimmy Butler in 2019, the Heat returned to the NBA Finals in 2020. The Heat acquired six-time NBA All-Star Kyle Lowry in 2021.

The Heat hold the record for the NBA's third-longest winning streak, 27 straight games, set during the 2012–13 season. Six Hall of Famers have played for Miami, and James won two consecutive NBA MVP Awards while playing for the team.



# Negative (Semantically Different)

---

**Deep learning** (also known as **deep structured learning**) is part of a broader family of **machine learning** methods based on **artificial neural networks** with **representation learning**. Learning can be **supervised**, **semi-supervised** or **unsupervised**.<sup>[2]</sup>

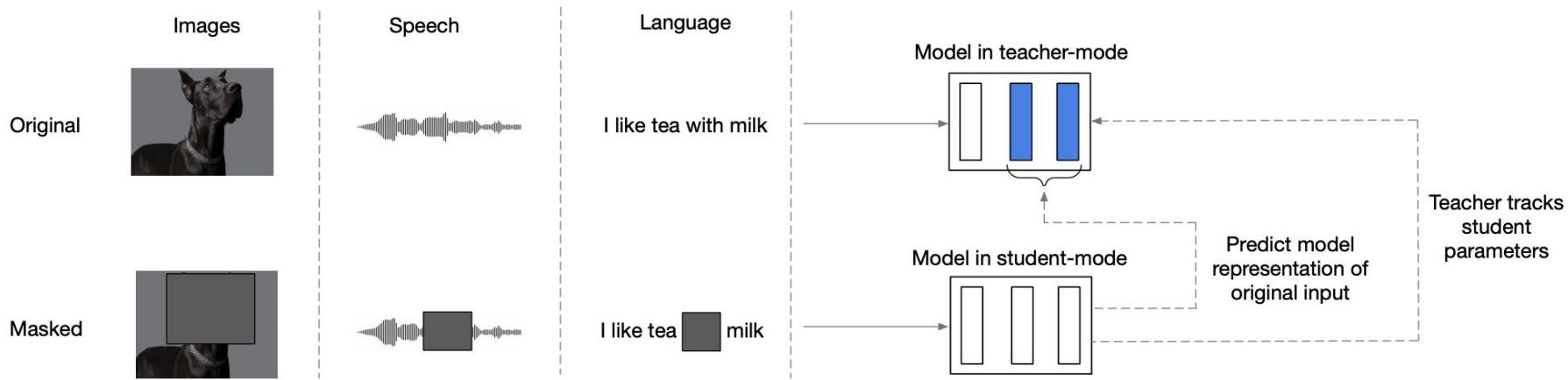
Deep-learning architectures such as **deep neural networks**, **deep belief networks**, **deep reinforcement learning**, **recurrent neural networks** and **convolutional neural networks** have been applied to fields including **computer vision**, **speech recognition**, **natural language processing**, **machine translation**, **bioinformatics**, **drug design**, **medical image analysis**, **climate science**, **material inspection** and **board game programs**, where they have produced results comparable to and in some cases surpassing human expert performance.<sup>[3][4][5]</sup>

**Artificial neural networks** (ANNs) were inspired by information processing and distributed communication nodes in **biological systems**. ANNs have various differences from biological **brains**. Specifically, artificial neural networks tend to be static and symbolic, while the biological brain of most living organisms is dynamic (plastic) and analogue.<sup>[6][7]</sup>



# Another strategy - Data2Vec, Baevski et al. 2022

## data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language



Do we need to train our own  
models?

**No! There are many pre-trained  
models that work very well for  
a broad range of data!**



# Great place to get started: Sentence Transformers



Hugging Face

Search models, datasets, users...



Sentence Transformers University

https://www.SBERT.net nreimers

Models 124

↑↓ Sort: Most Downloads

sentence-transformers/bert-base-nli-me...

Sentence Similarity • Updated Aug 5, 2021 • ↓ 2.72M • 6

sentence-transformers/all-MiniLM-L6-v2

Sentence Similarity • Updated Aug 30, 2021 • ↓ 1.88M • 35

sentence-transformers/paraphrase-multi...

Sentence Similarity • Updated Nov 2, 2021 • ↓ 643k • 38

sentence-transformers/all-MiniLM-L12-v2

Sentence Similarity • Updated Aug 30, 2021 • ↓ 503k • 3

sentence-transformers/msmarco-distilber...

Sentence Similarity • Updated Aug 5, 2021 • ↓ 449k • 1

sentence-transformers/paraphrase-MiniL...

Sentence Similarity • Updated Aug 30, 2021 • ↓ 2.07M • 11

sentence-transformers/all-mpnet-base-v2

Sentence Similarity • Updated Oct 15, 2021 • ↓ 695k • 30

sentence-transformers/all-distilrobert...

Sentence Similarity • Updated Aug 30, 2021 • ↓ 523k • 4

sentence-transformers/paraphrase-mpnet...

Sentence Similarity • Updated Aug 31, 2021 • ↓ 491k • 5

sentence-transformers/paraphrase-xlm-r...

Sentence Similarity • Updated Aug 5, 2021 • ↓ 369k • 31

▼ Expand 124 models



# Summary of Takeaway #2

## Vector Representations of Data

Data such as Images, Text, Code, ... can be represented as **Vectors** with Deep Learning models.

These models are trained to maximize semantic similarity with **massive** collections of data.

We often do not need to train the models ourselves for **particular data domains** to reach reasonable performance.



# Key Takeaway #3 - **Vector Segmentation**



# We can segment analytics based on the semantics of...

- Text
- Images
- Code
- Audio
- Video
- Graph-Structure
- Biological Sequences
- ... !



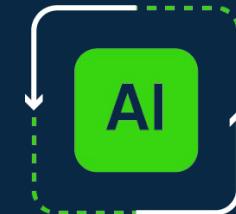
# Can we split Impressions based on the Semantics of the content?



Weaviate Podcast



Weaviate Tutorial



AI Weekly Update



# More Examples



# House Hunting

---

Symbols: # of bedrooms, # of bathrooms, square feet, city

→ With Vectors we can encode:

- Visual style
- Neighborhood structure
- *More flexible interface to define features with text*



# e-Commerce Products

---

Symbols: “Shoes”, “T-Shirt”, “Pants” or colors

→ With Vectors we can encode visual styles



# Movies

---

Symbols can differentiate between genres like “Children”, “Action”, or “Sci-Fi”

→ With Vectors we can encode:

- Themes
- Characters
- Storylines



# Scientific Papers

---

Symbols: “Biology”, “Machine Learning”

→ With Vectors we can encode

- Nuance of the ideas
- Writing style



# Music

---

Symbols can differentiate between genres like “Hip Hop”, “Dance”

→ With Vectors we can encode:

- Tone
- Lyrics
- Instruments



**“That’s the magic of deep learning:  
turning meaning into vectors, then into geometric  
spaces, and then incrementally learning complex  
geometric transformations that map one space to  
another. All you need are spaces of sufficiently high  
dimensionality in order to capture the full scope of  
the relationships found in the original data.”**

- Francois Chollet, Deep Learning with Python, 2nd edition



# Summary of Takeaway #3

## Vector Segmentation

Vector representations, also known as embeddings, enable an **Interface** to split analytics based on the **Semantics** of the content.

This content could be **Text, Images, Code, Audio, Videos, ...**



# Key Takeaway #4 - **Weaviate for Twitter Analytics**



# Twitter Analytics

---

Tweet Text	Time	Impressions	Engagements	Engagement Rate	Retweets	Replies	Likes	User Profile Clicks	Url Clicks
I just published "ANN Benchmarks with Etienne Dilcoker -- Weaviate Podcast #16 on Medium..	May 27th, 1:34pm	1905	50	2.6%	3	1	15	2	18
Approximate Nearest Neighbor algorithms allow us to Vector Search in massive datasets! ...	May 24th, 1:13pm	7182	252	3.5%	14	1	50	27	36

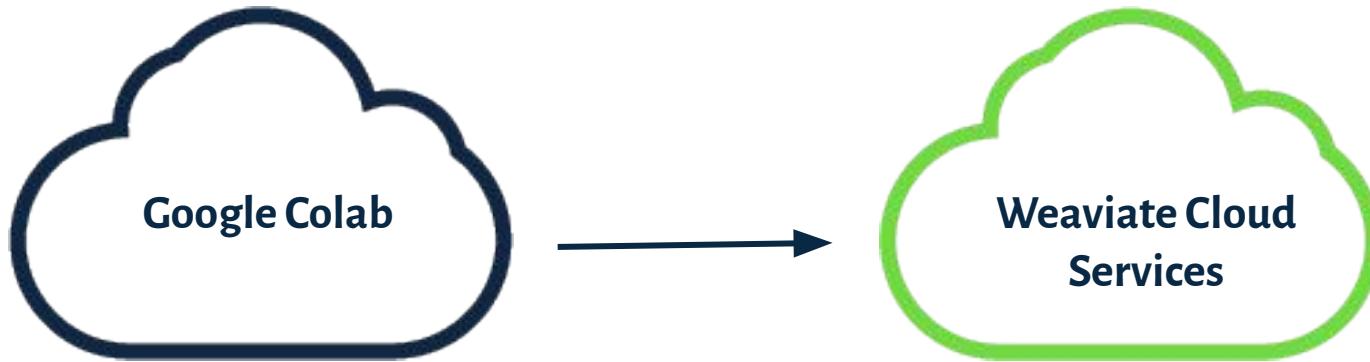


```
weaviate_schema = {
    "classes": [
        {
            "class": "Tweet",
            "description": "Tweet Analytics",
            "properties": [
                {
                    "name": "tweet_text",
                    "dataType": ["text"],
                    "description": "The text in the Tweet.",
                    "moduleConfig": {
                        "text2vec-transformers": {
                            "skip": False,
                            "vectorizePropertyName": False
                        }
                    },
                    ...
                },
                {
                    "name": "author",
                    "dataType": ["text"],
                    ...
                }
            ]
        }
    ]
}
```



# Cloud Data Upload

---



There are many other ways to do this as well



# GraphQL Live Demo



## 5 Nearest Neighbors to → “Weaviate Coding Tutorial”

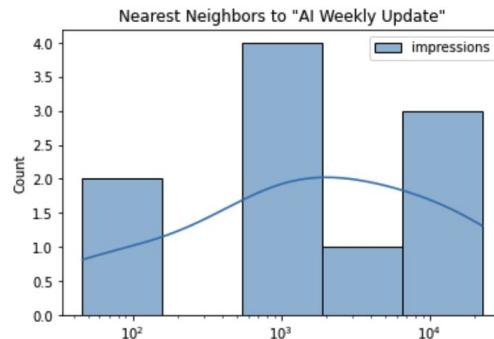
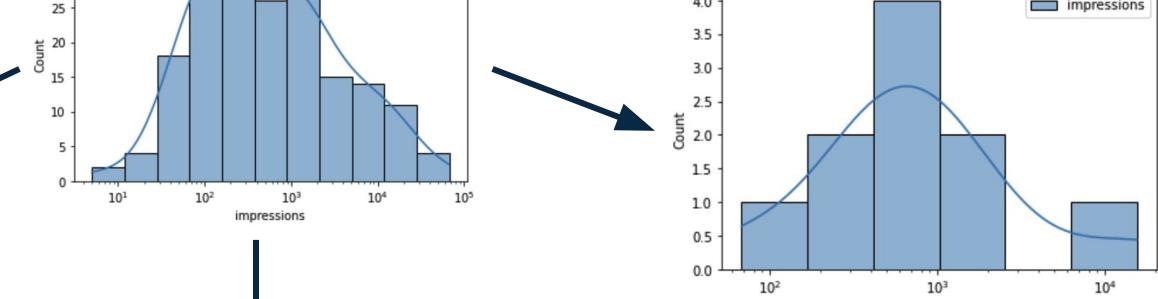
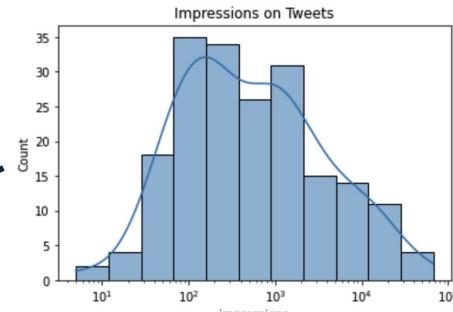
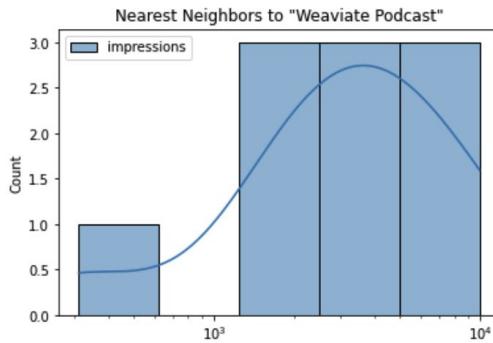
Content	Impressions
“We have 4 Weaviate Podcast Episodes so far [ ... ] <b>how to utilize the Weaviate Database as a Document Store in Haystack pipelines ...</b> ”	311
“We have 2 new coding tutorials on Weaviate YouTube...”	1144
“@weaviate_io Love the integration of this with the GraphQL API!”	378
“Here are some thoughts on combining Weaviate and Haystack! TLDR: Weaviate is a great Vector Search database...”	15563
“Weaviate (@weaviate_io) is also announcing a collaboration with Jina AI (@JinaAI_)! ...”	586



```
def nearText_wrapper(query):
    nearText = {
        "concepts": query
    }
    nearest_neighbor_search = client.query.get("Tweet", ["tweet_text", "impressions"])
    nearest_neighbor_search = nearest_neighbor_search.with_near_text(nearText).do()
    return nearest_neighbor_search["data"]["Get"]["Tweet"]
```



# What was the Tweet about?



# Have I tweeted something like this before?

The screenshot shows a database query interface with a pink header bar. The top navigation bar includes icons for a profile picture, a gear, and a search bar, followed by links for 'Log Out' and 'Docs'. Below the header is a toolbar with buttons for 'Prettify', 'Merge', 'Copy', 'History', 'Schema', 'Share this query', and a refresh icon.

The main area displays a code editor with a JSON-like query and its results. The query is:

```
1 v {
2   Get {
3     Tweet {nearText: {
4       concepts: ["This video explains some ideas around the OpenAI Embeddings API!\n\nI had the opportunity to interview Arvind Neelakantan (@arvind_io) from OpenAI about these ideas and this video summarizes my takeaways and provides background for each topic.\n\nhttps://t.co/rJymcSYx0t",
5       "New Weaviate podcast with Arvind Neelakantan (@arvind_io) about the OpenAI Embeddings API, covering many topics from:\n\n• What's new in Text Embeddings?\n• One model for all domains\n• Impact of Data Preprocessing\n• Large Embedding Vectors\n• Label Embeddings\n• and more!\n\nhttps://t.co/Tn7xH3Ppd",
6       "Our Weaviate Podcast with Arvind Neelakantan (@arvind_io) on the OpenAI Embeddings API and miscellaneous other topics has hit 500 views! 🎉\n\nThank you so much for the support on the Weaviate podcast, really looking forward to building this further!\n\nhttps://t.co/v92izE3J0r",
7       "New AI Weekly Update on Henry AI Labs for January 31st, 2022! 🎉\n\nOpenAI Embeddings\n• Training LMs to follow instructions\n• Natural Language Descriptions of Deep Visual Features (MILAN)\n• GreaselM\n• Synchromesh\n• and more!\n\nhttps://t.co/HoL508AGHQ"
8     }
9   }
10 }
```

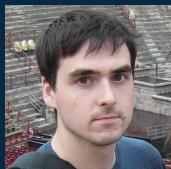
The results show a list of tweets with their text content, impressions, and URL clicks:

- "impressions": 66824, "tweet\_text": "This video explains some ideas around the OpenAI Embeddings API!\n\nI had the opportunity to interview Arvind Neelakantan (@arvind\_io) from OpenAI about these ideas and this video summarizes my takeaways and provides background for each topic.\n\nhttps://t.co/rJymcSYx0t", "url\_clicks": 211}
- "impressions": 2699, "tweet\_text": "New Weaviate podcast with Arvind Neelakantan (@arvind\_io) about the OpenAI Embeddings API, covering many topics from:\n\n• What's new in Text Embeddings?\n• One model for all domains\n• Impact of Data Preprocessing\n• Large Embedding Vectors\n• Label Embeddings\n• and more!\n\nhttps://t.co/Tn7xH3Ppd", "url\_clicks": 0}
- "impressions": 3521, "tweet\_text": "Our Weaviate Podcast with Arvind Neelakantan (@arvind\_io) on the OpenAI Embeddings API and miscellaneous other topics has hit 500 views! 🎉\n\nThank you so much for the support on the Weaviate podcast, really looking forward to building this further!\n\nhttps://t.co/v92izE3J0r", "url\_clicks": 20}
- "impressions": 4358, "tweet\_text": "New AI Weekly Update on Henry AI Labs for January 31st, 2022! 🎉\n\nOpenAI Embeddings\n• Training LMs to follow instructions\n• Natural Language Descriptions of Deep Visual Features (MILAN)\n• GreaselM\n• Synchromesh\n• and more!\n\nhttps://t.co/HoL508AGHQ", "url\_clicks": 18}

# Have any Weaviate Podcast guests tweeted something like this recently?



Weaviate  
podcast





```
import tweepy

auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token_key, access_token_secret)

api = tweepy.API(auth)

this_persons_tweets = api.user_timeline(username, count=100)
```





Tweet, Author, Likes



Weaviate

# GraphQL Live Demo

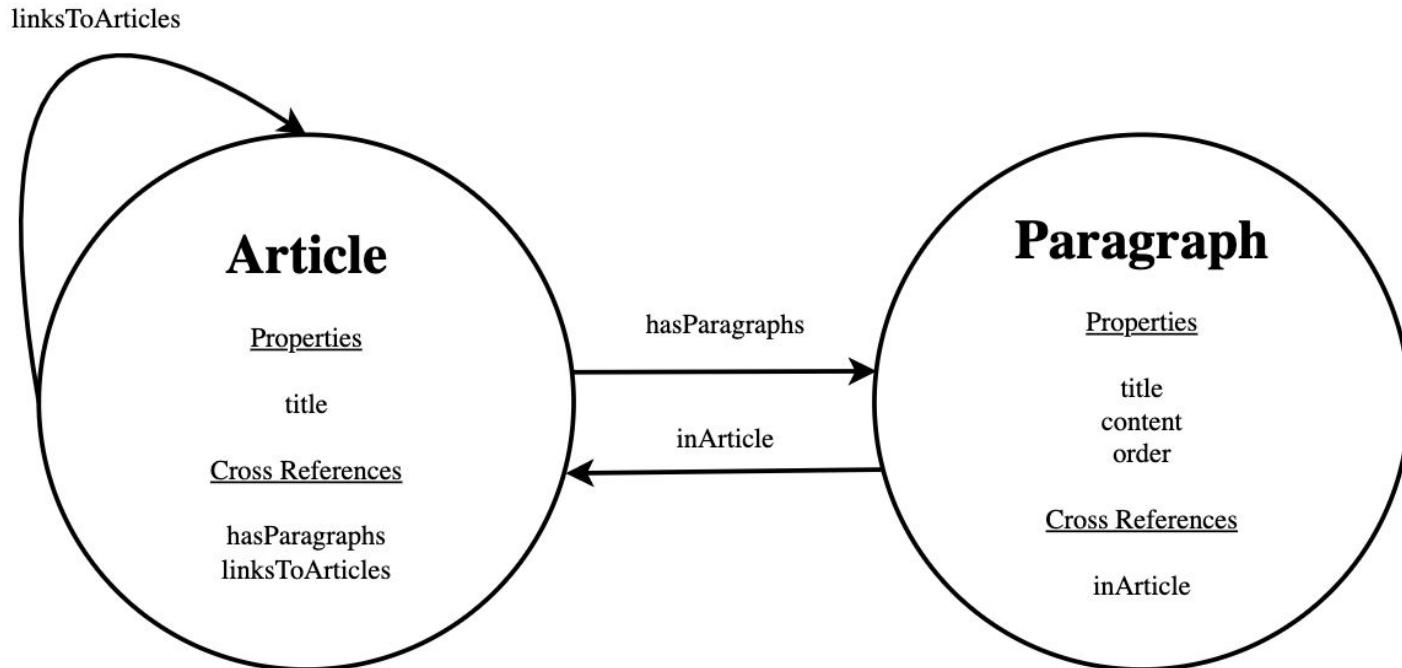


# GraphQL Wikipedia Demo



WIKIPEDIA  
The Free Encyclopedia

# Wikipedia Live Demo - Graph Data Model



# GraphQL Wikipedia Demo



**WIKIPEDIA**  
The Free Encyclopedia



 SCAN ME



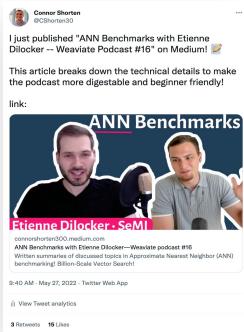
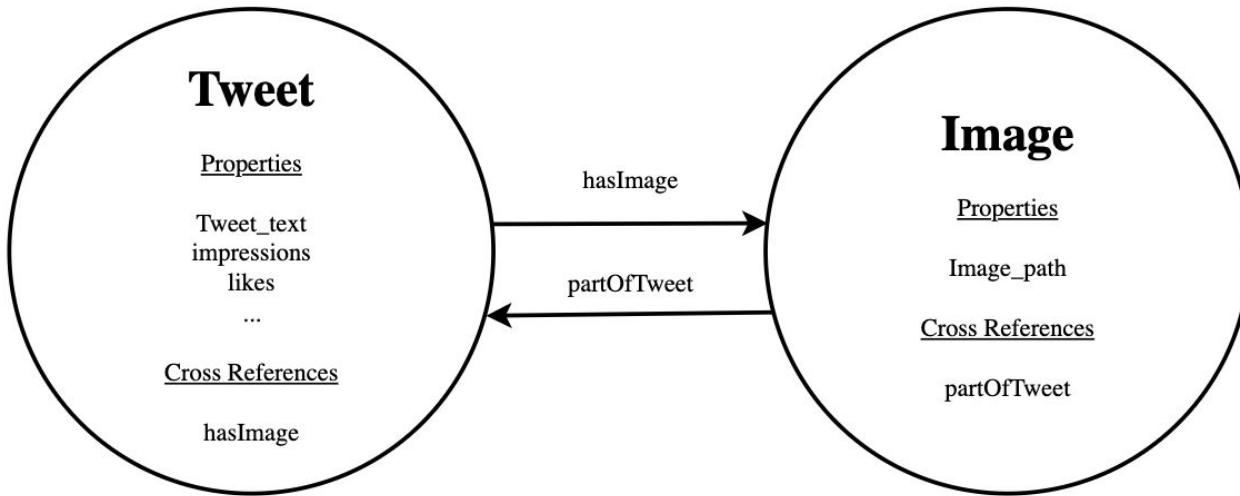
# Weaviate

- Weaviate is a Vector Search **Database**, rather than a **Library** such as Facebook's FAISS or ANNOY from Spotify
- Weaviate has a **Graph-like Data Model**



# Expanding Twitter project with Graph Model

---



[Star](#)

2,486

current ▾

**WEAVIATE**[Introduction](#)**GETTING STARTED**[Quick start](#)[Installation](#)**ARCHITECTURE**[Roadmap](#)[Horizontal Scaling](#)[Storage](#)[Filtered Vector Search](#)

## Quick start

stackoverflow #weaviate github issues version v1.13.2

api-specs v1.13.2 downloads 947,608

This quick start guide will give you a 10-minute tour of Weaviate. You will set up your Weaviate with Docker, add an example dataset with news articles, make your first queries to browse through your data, and let Weaviate perform automatic classification. This guide uses the "text2vec-transformers" module. You can find a quick start with the text2vec-contextionary module [here](#).

## Run Weaviate with a demo dataset



# Summary of Takeaway #4 **Weaviate for Twitter Analytics**

We can segment **Impressions** on Twitter based on the content of the tweet **without manual labeling!**

Weaviate is a **Vector Search Database** that can be used to store and search through semantic embeddings of data.



# Key Takeaway #5 - **Research Questions and Discussion**



# Research Questions and Discussion

---

- Should I fine-tune my embedding model?
- Large-Scale Vector Search with Approximate Nearest Neighbor (ANN) Algorithms
- How does Vector Search differ from Classification or Regression models?



# Vector Search versus Regression on Impressions

---



Model Prediction  
**8,530 Impressions**



# Interpretability of Vector Search

**Nearest Neighbors**

The diagram illustrates the concept of nearest neighbors in vector search. It shows three tweets from a user's timeline. Two tweets are labeled "Unsent Tweets" and are connected by a large green double-headed arrow, indicating they are nearest neighbors. The third tweet is a regular sent tweet.

**Unsent Tweets**

Hey everyone, I hope you enjoy my new podcast with Bob van Luijt about Weaviate! 🎧

Everyone can reply

**Everyone**

A summary of my conversation with the CEO of a Vector Search company!

Everyone can reply

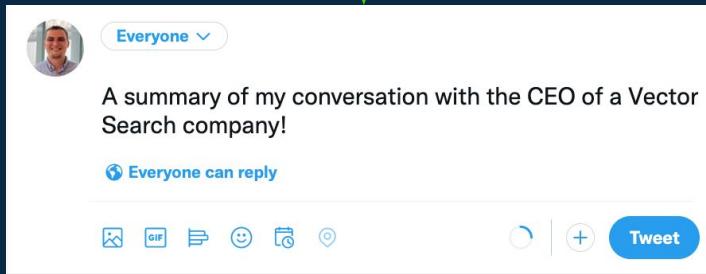
**Everyone**

Here are my thoughts on what Bob van Luijt had to say about Vector Search!

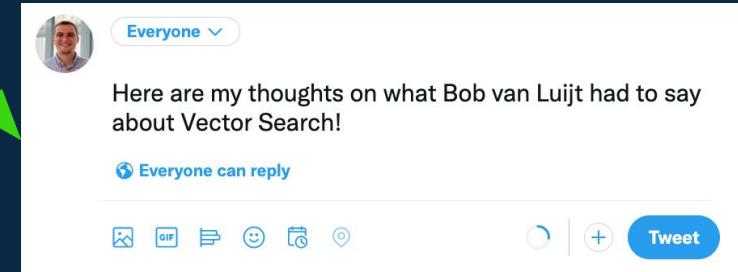
Everyone can reply

**Everyone**

# Interpretability of Vector Search and Prediction



Model Prediction  
8,530 Impressions



# **What do we want to know about our Tweets?**

---

Should I post this?

When might be a better time to post it?

What might be a better phrasing of this tweet?



# Expanding from individuals to teams

---

- Has anyone on my team tweeted something like this recently?
- Who on our team would be best fit to tell this story?
- What topics should we be tweeting about?



# **Summary of Takeaway #5**

## **Research Questions and Discussion**

How can we **improve** these systems? What looks **promising**?



# Key Takeaways: “Vector Search for Data Scientists”

Slides, Colab Notebook, Video Presentation available on:  
[github.com/CShorten/Vector-Search-for-Data-Scientists](https://github.com/CShorten/Vector-Search-for-Data-Scientists)

1. Segmentation in Data Science
2. Vector Representations of Unstructured Data
3. Vector Segmentation
4. Weaviate Example for Twitter Analytics
5. Research Questions and Discussion



# Connect with us!

---



Weaviate Slack Channel



YouTube: Weaviate · Vector Search Engine



Weaviate Podcast



Twitter @weaviate\_io



# Thank you for Watching!

Special thanks to **Sebastian Witalec** in  
advising the development of this presentation  
and **Svitlana Smolianova** for visual styling.

