CM.com

Inhouse Day

CM.com

29 April 2024

Ennia
Senior Data Scientist

Hugo
Junior Data Scientist

Renée
Junior Data Scientist

Thomas
Junior Data Scientist

Agenda

01.

About CM.com

02.

Case Introduction

03.

Working on the case

04.

Wrap-up

05.

Drinks!

About CM.com

It started in 1999

CM.com facts.

1999
Founded in Breda

Listed company

750+
Employees

19
Offices globally

Billions
interactions
with users

24/7 Monitoring
& Support

CM.com

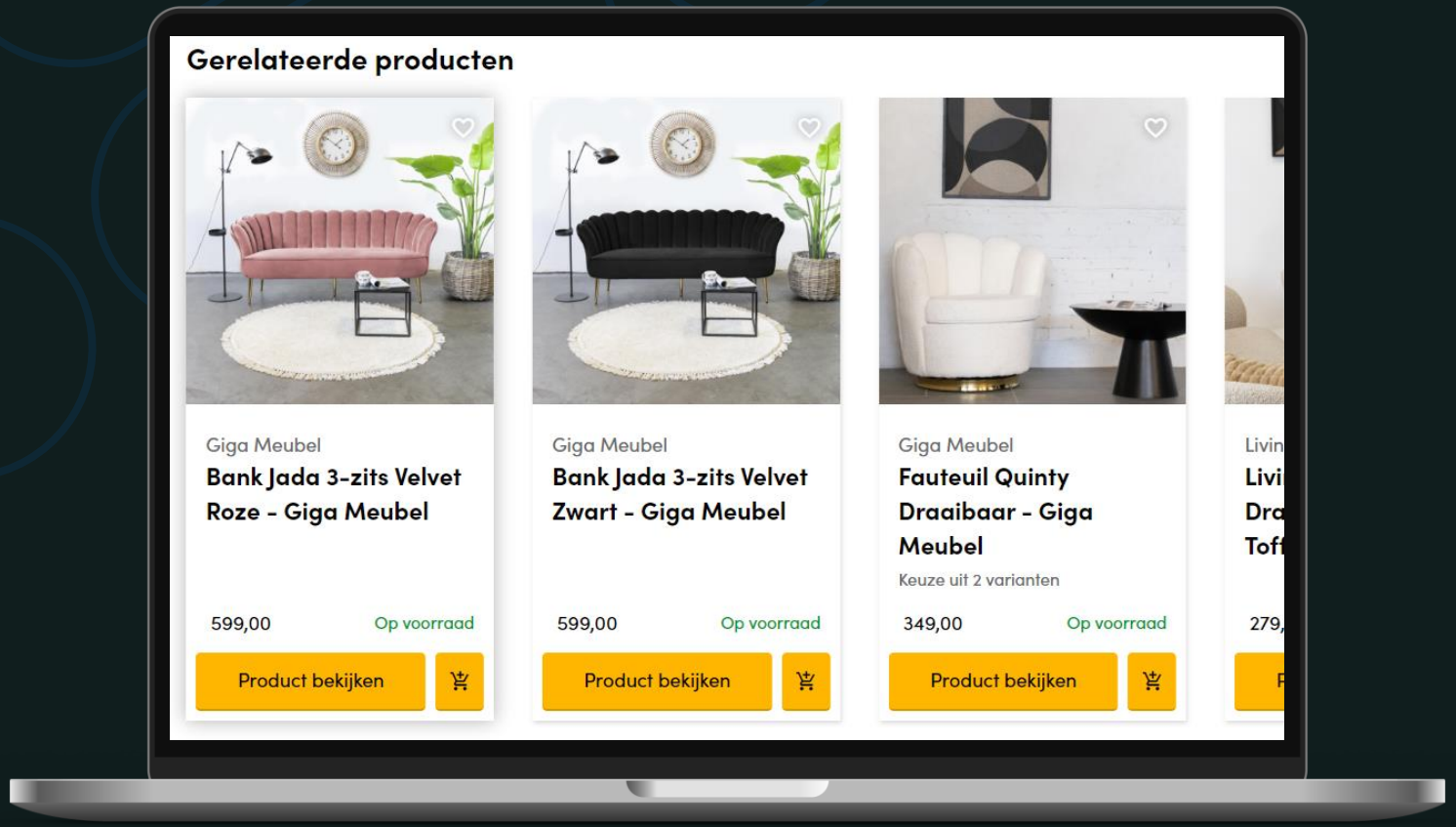
Do what you like,
Do what you're good at,
And contribute

CM.com

CM.com loves interns

CM.com & Econometricians

Client Case: Product Recommendations



Client Case: Workout Reactivation Program

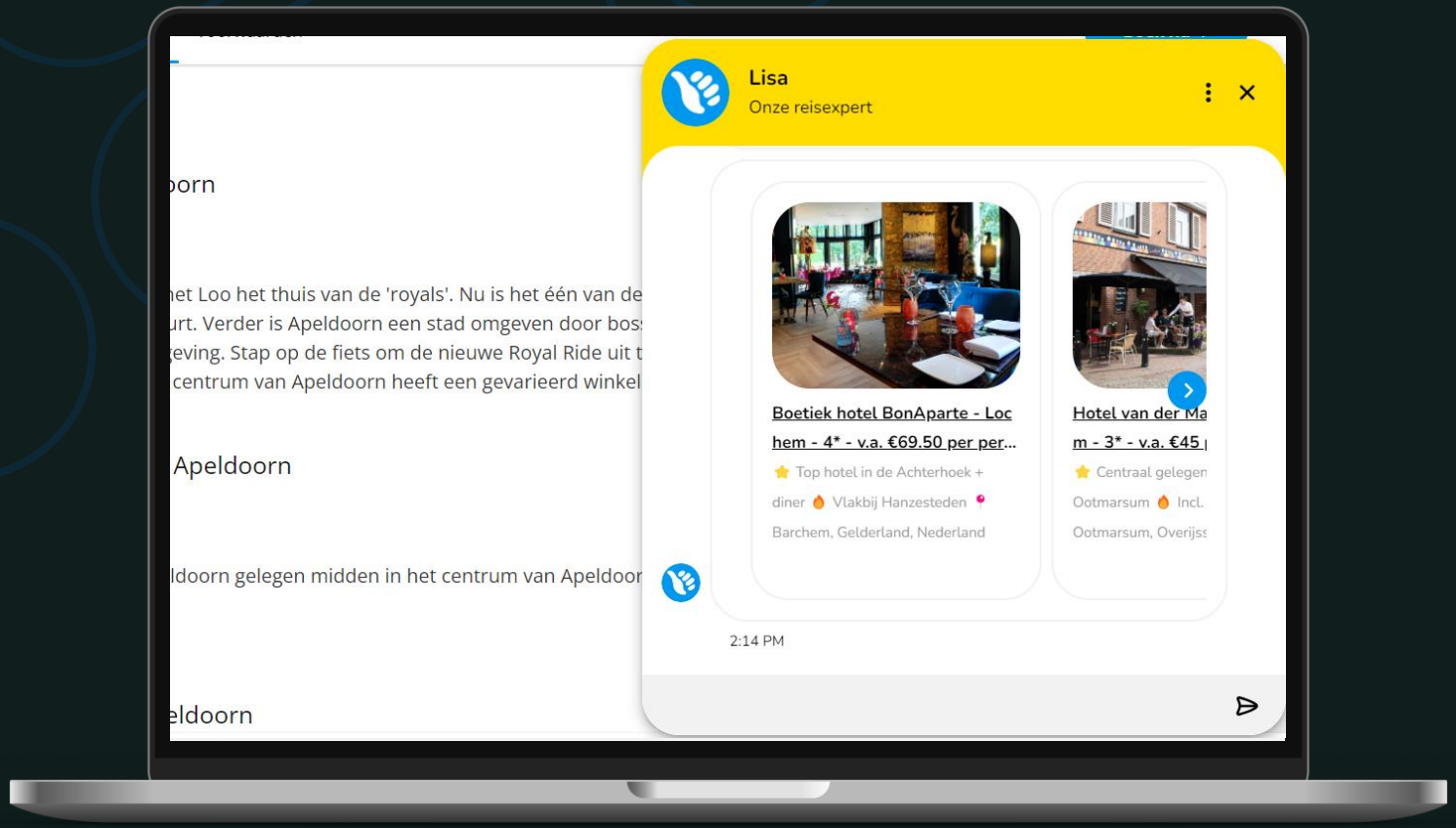
BASIC-FIT

-5%

Customer churn

Client Case: Taste Discovery

bookit.
Powered by data, driven by people



Possible roles for Econometricians

Data Scientist

is responsible for the correct predictions/outcomes

- Research & development
- Programming
- Modelling

Data Translator

is responsible for value being added for clients

- Client contact
- Data insights & experimenting
- Business understanding

Day in the life of a ...

Data Scientist

Client implementations

- Implementing new solutions at clients
- Maintaining existing solutions
- Solving incidents

Product development

- Design for flexible code
- Writing & testing code
- Documenting & releasing

Data Translator

Client implementations

- Scoping new solutions with clients
- Data understanding & insights to optimize value
- Guide client in adopting AI within organization

Product development

- Bridge between product and markets
- Challenge outcome of product developments

Data Scientist / Translator

Meetings

- Team meetings
- Client meetings

Other

- Thesis supervision
- Recruitment events

Company-wide

- Quarterly meetings
- Drinks & Events!

CM.com

Case Introduction

What is the case about?

The **goal** of this case is to build a **Movie Recommender!**

What data will you use?

You will use IMDb data that contains information about movies, like:

- Title
- Description
- Language
- Popularity
- Runtime
- Genre
- Release date
- Revenue
- ...

Title	Description	Release Date	Revenue
Avatar	In the 22nd century...	12/10/2009	2.79 billion
Star Wars: The Force Awakens	Thirty years after defeating the Galactic Empire...	12/15/2015	2.06 billion
Forrest Gump	A man with a low IQ has accomplished...	7/6/1994	677 million

There is a lot of data present, you don't need to use it all!

What will you do in the case?

- **Make recommendations using the movie descriptions!**
- **Some code already written for you, but some you have to yourself:**
  - Descriptive analysis
  - Data preprocessing
  - Improving the recommendations
- **You will split into teams of 3 or 4 people.**
- **You will have access to a Colab Python Notebook.**
- **Afterwards, you will need to pitch your findings to a jury**

What do you need to pitch?

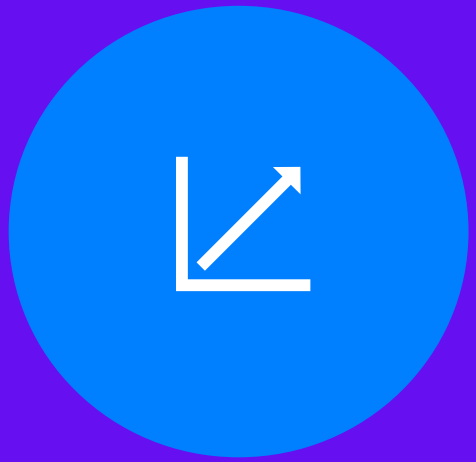
In your pitch (3 minutes maximum) you should:

- Show what descriptive analytics and data preprocessing steps you took
- Share your insights on the recommendations:
  - Share you **chosen movie** and its **recommended movies**
  - Tell which **improvements** you made to the recommender
  - Discuss whether the recommendations **make sense**, and if they are logical given the improvements you made
- Convince us of your ideas for the future, think for example about:
  - **Suggestions** that could further improve the recommender
  - Suggest **other recommender methodologies** that could be used

Background knowledge: NLP basics

Three basic steps in NLP

1. Text Preprocessing



2. Calculating Embeddings

3. Calculating Similarity

Preprocessing text to only get relevant information

A epic movie, it’s is about the Superheroes., and the avengers with Thanos but not Inf. War movie, popular

epic movie superheroes avengers thanos not infinity war movie popular

Embedding to convert text into a numerical format

epic movie
superheroes
avengers thanos not
infinity war movie
popular

Calculating cosine similarity to get simililarity score

epic movie superheroes avengers thanos not infinity war movie popular

$$\begin{bmatrix} 1.009 \\ 5.826 \\ 7.518 \\ 3.027 \\ \vdots \\ 9.241 \\ 4.745 \end{bmatrix}$$

After the devastating events of Avengers: Infinity War (2018), the universe is in ruins. With the help of remaining allies, the Avengers...

$$\begin{bmatrix} 1.119 \\ 4.826 \\ 7.523 \\ 3.027 \\ \vdots \\ 2.241 \\ 5.045 \end{bmatrix}$$

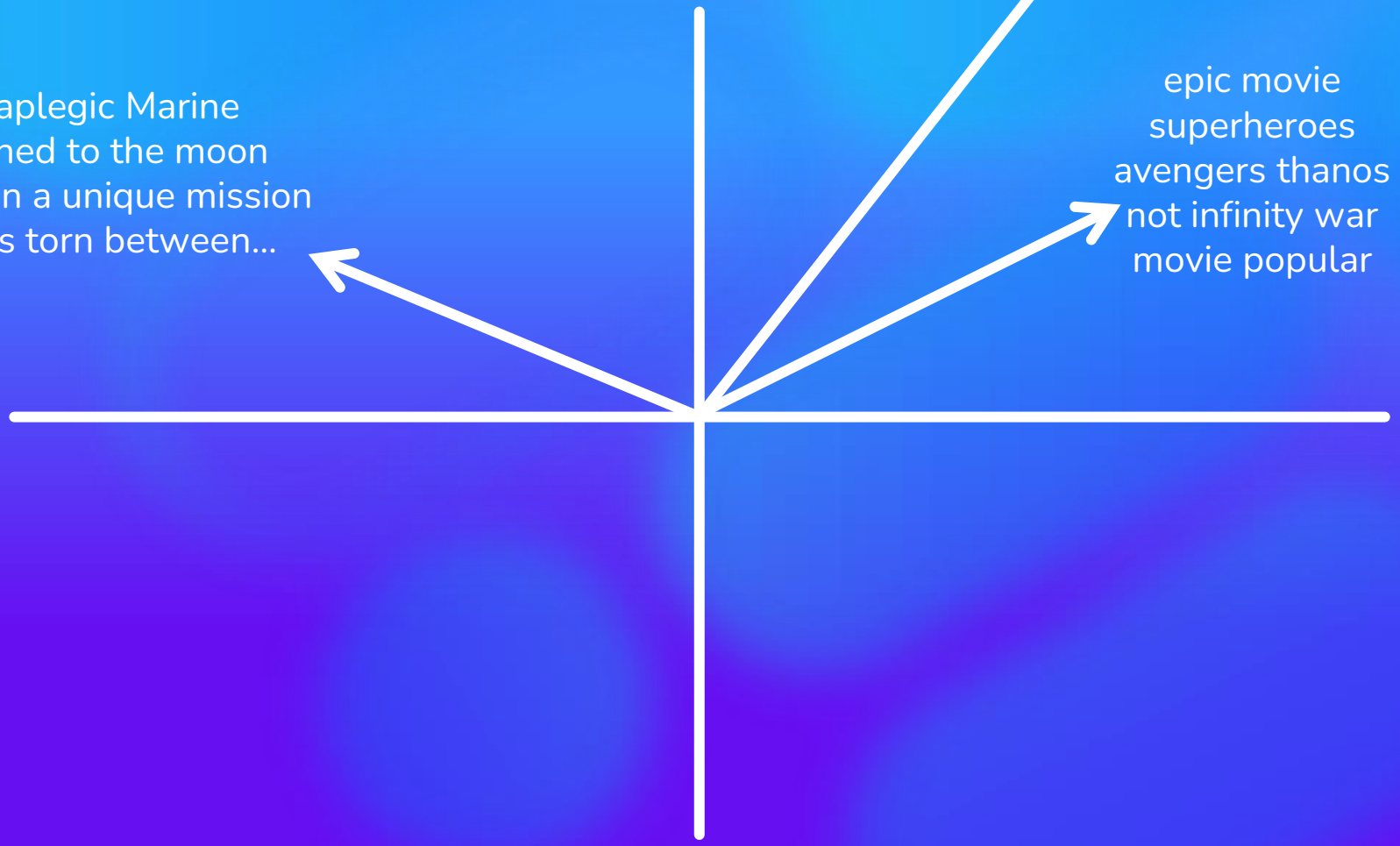
Cosine Similarity: 0.78

Embeddings can also be illustrated graphically

After the devastating events of Avengers: Infinity War (2018), the universe is in ruins. With the help of remaining allies, the Avengers...

epic movie superheroes avengers thanos not infinity war movie popular

A paraplegic Marine dispatched to the moon Pandora on a unique mission becomes torn between...



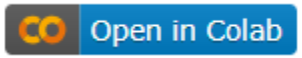
Background knowledge: NLP

We will generate our NLP recommendations based on the **movie description** column. We will do this by determining which descriptions are similar to the base movie description

You need to perform 3 steps that are commonly used in NLP to arrive at recommendations:

1. **Text Preprocessing**
    - You need to 'clean' the movie descriptions such that we only keep relevant text.
    - Think of removing characters/words that do not add to determining similarity between descriptions
2. **Calculating Embeddings of the Descriptions**
    - Embedding a text means to convert it into a **numerical representation**
    - This numerical representations is a vector which represents the description
    - The model we use has 384 dimensions, so our movie description will be transformed into a vector with 384 numbers!
3. **Calculating Similarity between Embeddings**
    - After we have calculated the embeddings of all the movie descriptions, we need to see which embeddings are similar
    - A commonly used metric for this is the **cosine similarity**.
    - A cosine similarity of 1 implies the vectors are identical.
    - A cosine similarity of 0 implies the vectors are very dissimilar.

Let's start with the case!

- You will have **60 minutes** for the case
- 3 things for you to **improve**:
  - Descriptive Analytics
  - Text Preprocessing
  - Improving Recommendations
- Pitch your findings to us in **3 minutes**
- **Scan** the QR code to open case repository
- To start the case, **open** "NLP Case 2024.ipynb" and click on 

- Make a **copy** of the notebook on your drive
- **Slides** are also available in repository

me-qr.com/MdaPgkLJ

CM.com

Wrap-Up

Case in Real Life!

Case in Real Life!

Data Science Topics

- Text Embedding
- Similarity Search

Case in Real Life!

Data Science Topics

- Text Embedding
- Similarity Search

Generative AI Engine

Case in Real Life!

Concept of Generative AI Engine

- ChatGPT
- Implemented in a client environment

Large Language Model (LLM)

Large Language Model (LLM)

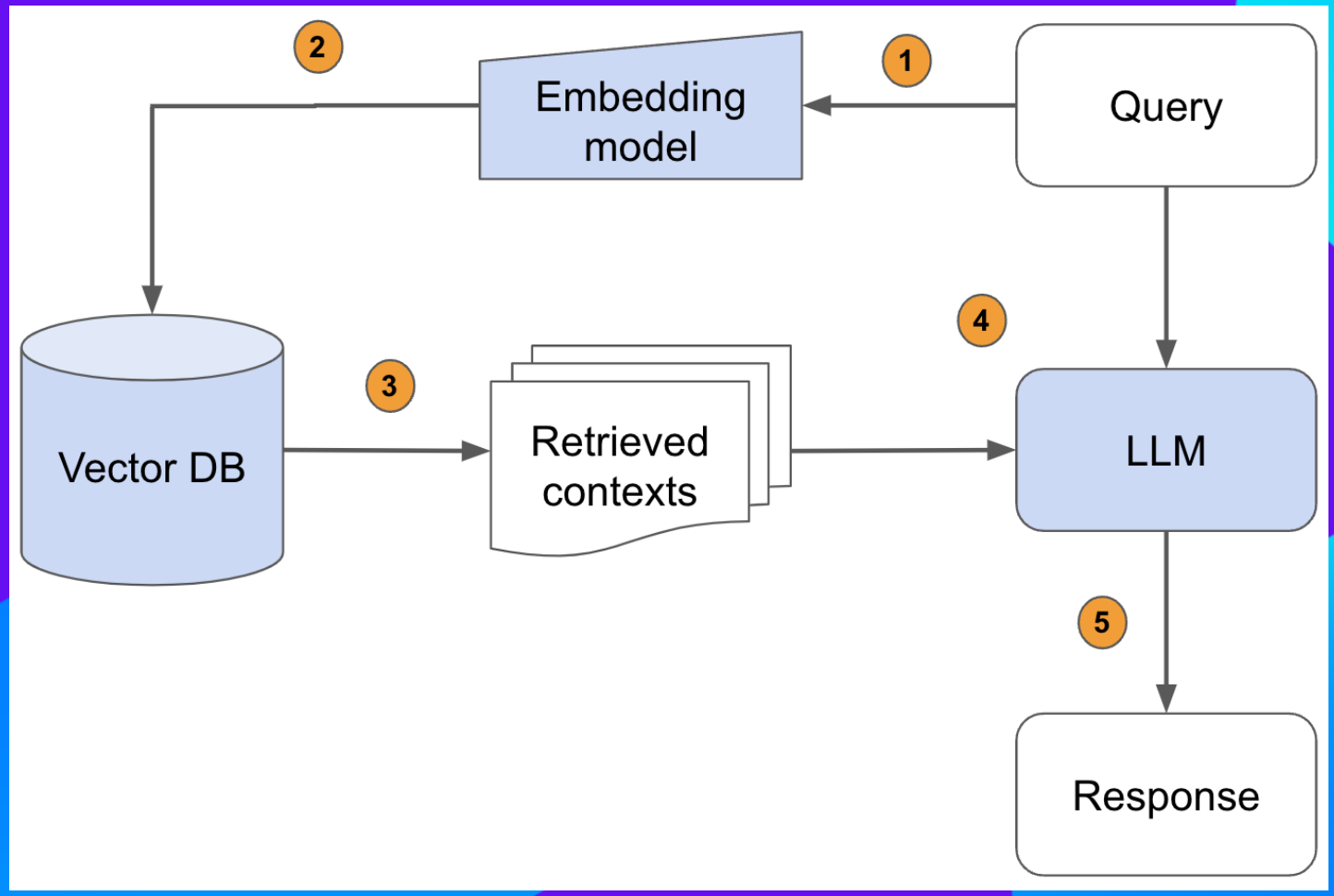
Hi! Today, I had a very cool NLP case at CM, but do they also like to have some drinks and bites after such an interesting event?

Large Language Model (LLM)

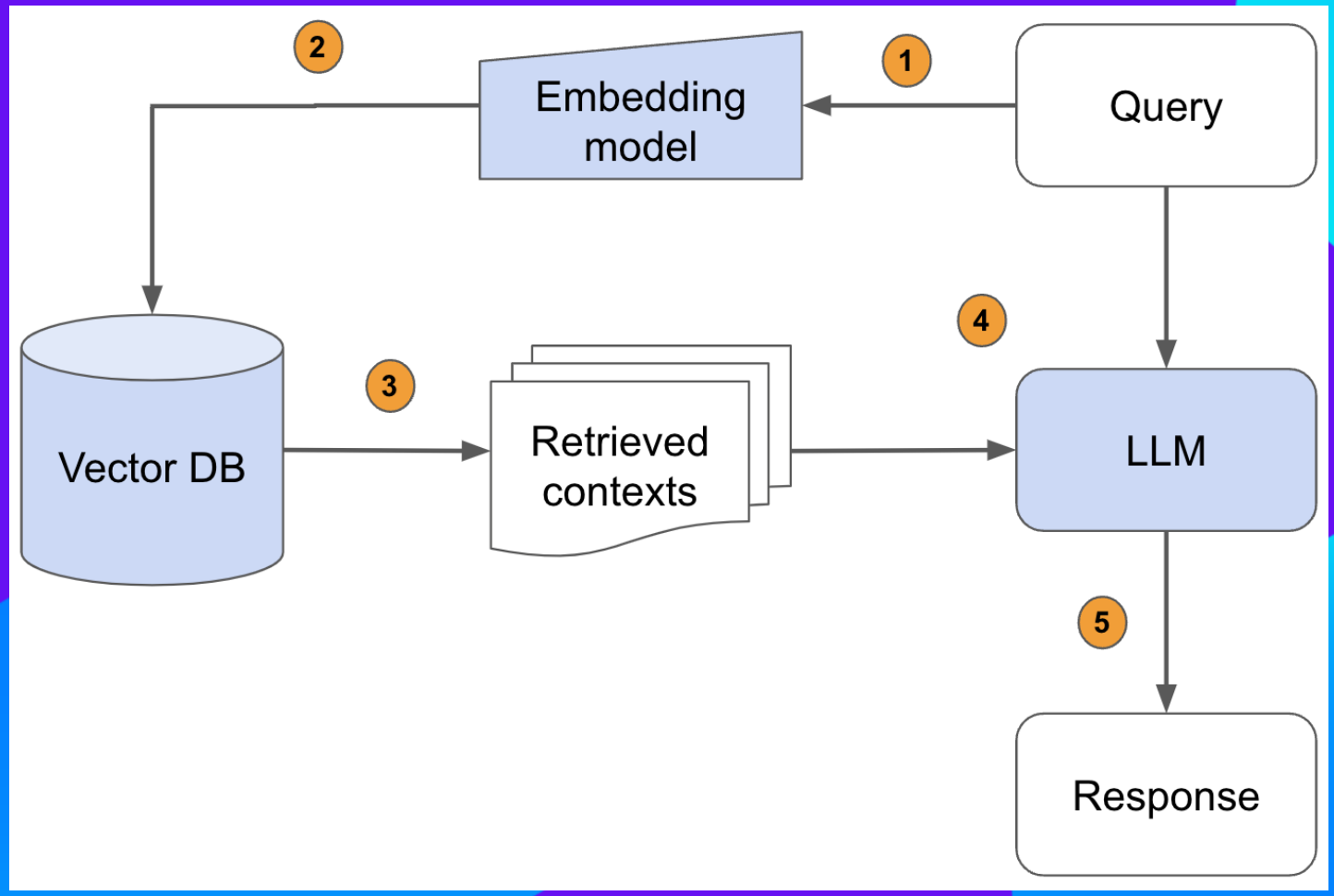
Hi! Today, I had a very cool NLP case at CM, but do they also like to have some drinks and bites after such an interesting event?

It's quite common at companies to socialize after such an event, but I do not know CM.

Retrieval Augmented Generation (RAG)



Retrieval Augmented Generation (RAG)

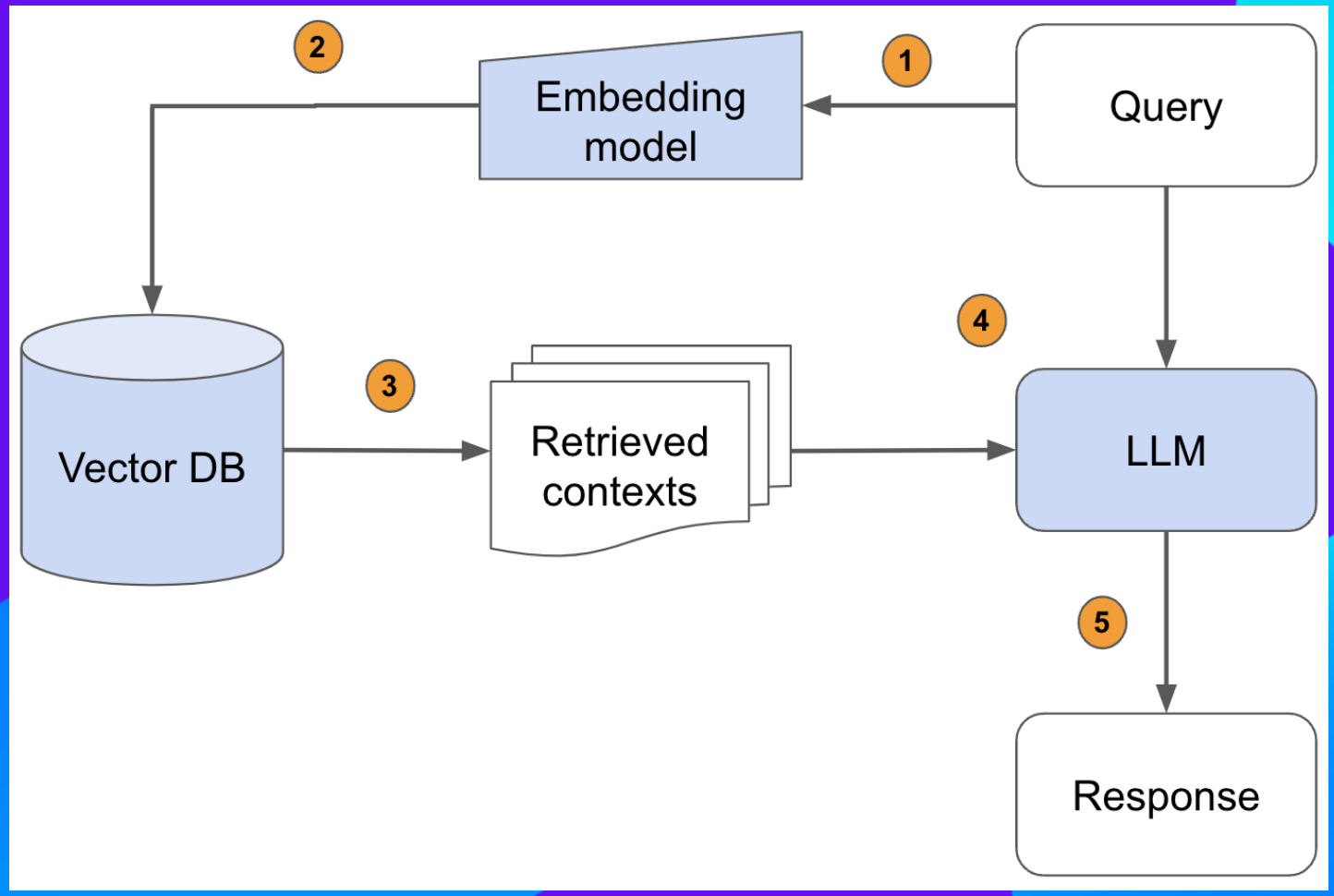


Hi! Today, I had a very cool NLP case at CM, but do they also like to have some drinks and bites after such an interesting event?

Retrieval Augmented Generation (RAG)

Hi! Today, I had a very cool NLP case at CM, but do they also like to have some drinks and bites after such an interesting event?

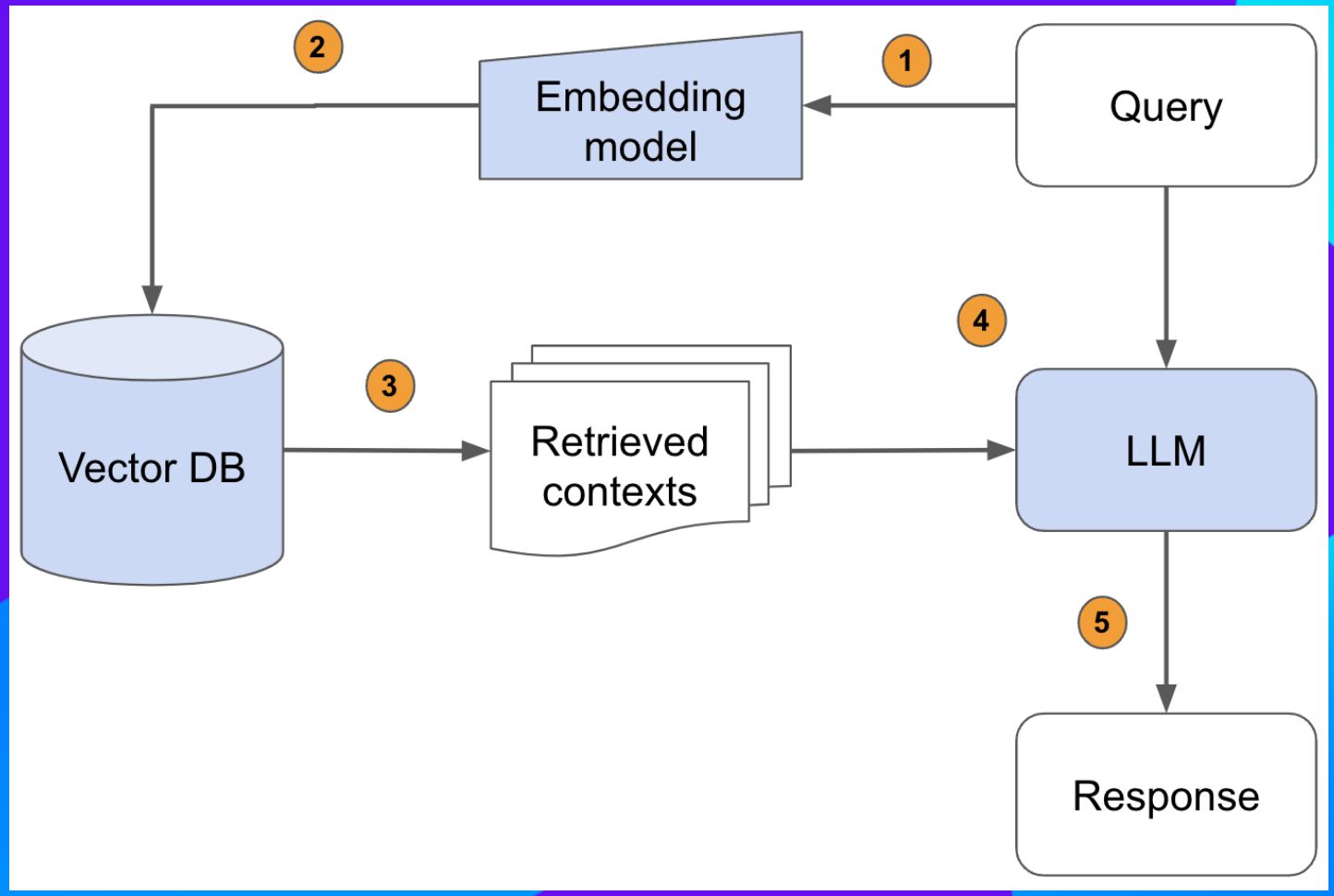
CM is a very nice company with many young people who enjoy engaging in social activities such as having drinks and snacks!



Retrieval Augmented Generation (RAG)

Hi! Today, I had a very cool NLP case at CM, but do they also like to have some drinks and bites after such an interesting event?

CM is a very nice company with many young people who enjoy engaging in social activities such as having drinks and snacks!



Yes! After events, CM employees often enjoy drinks and snacks together!

CM.com

Questions?

CM.com

And the winner is …

CM.com

Thank you.

If you have any questions or are interested in a career with us, don't hesitate to contact.

www.CM.com/careers

You can find our socials here:

Lidia de Heer

Connect on LinkedIn

Talent@cm.com