February 2022

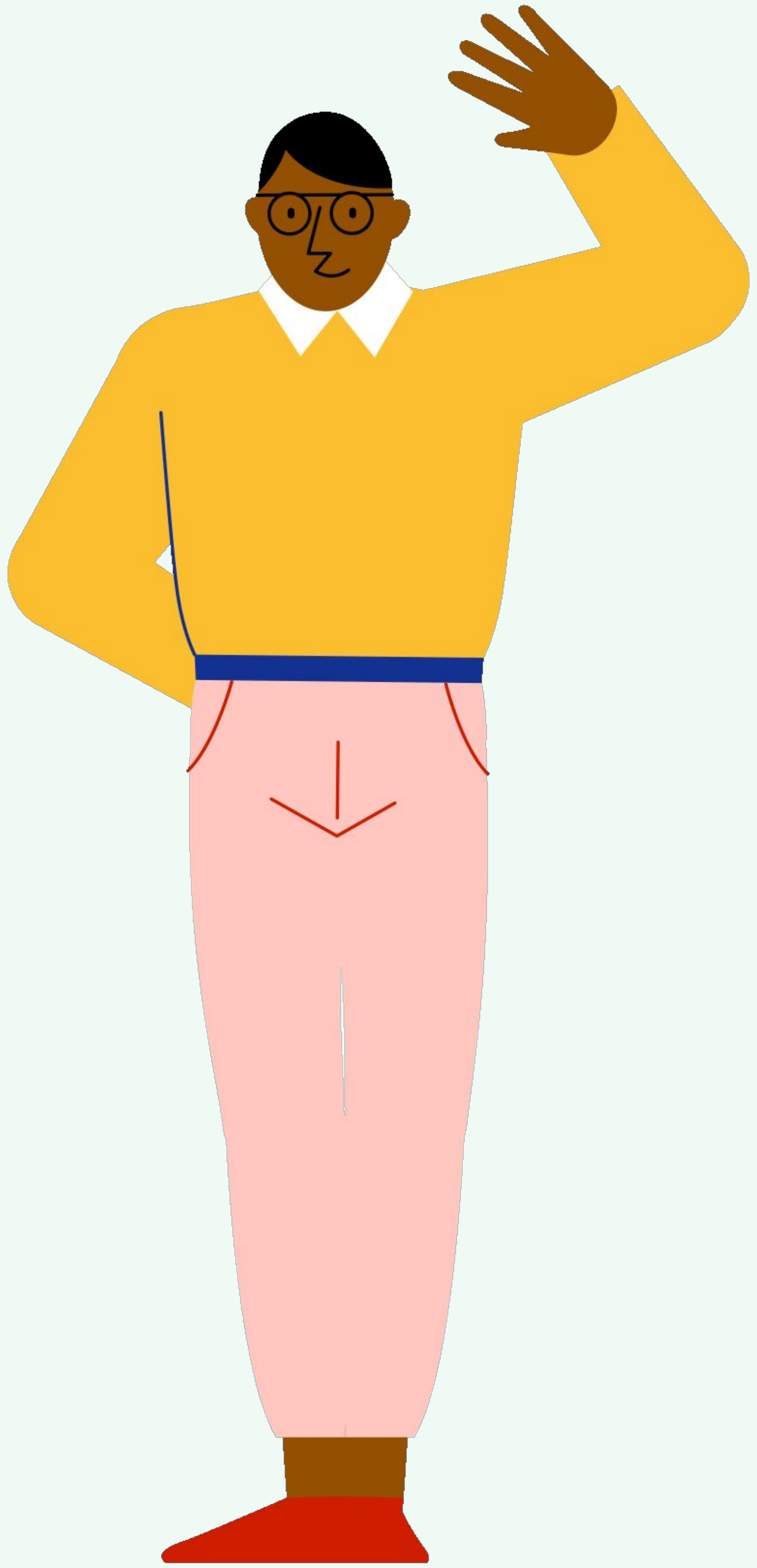
Content deduplication

Vector vs Keyword

WHO AM I?

ZBYSZKO PAPIERSKI

Search Tech Lead @Brainly



February 2022

Content deduplication

Vector vs Keyword

February 2022

Educational, user-generated content deduplication @Brainly

Vector vs Keyword

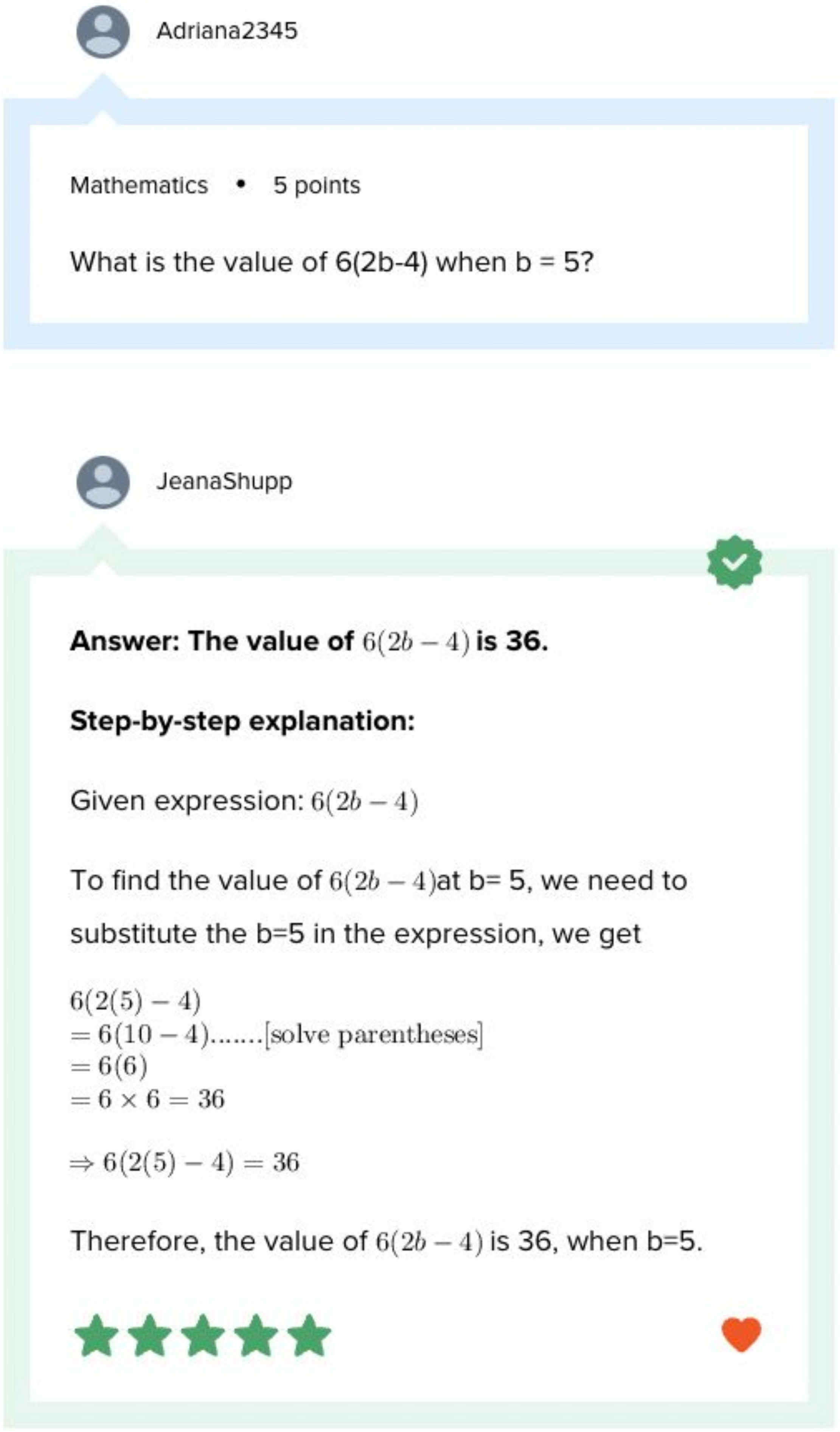
February 2022

Educational, user-generated Content deduplication @Brainly

Vector vs Keyword

ASK QUESTIONS
GET HELP
GO BEYOND

Our community of experts consists of students, schoolteachers, PhDs, and other geniuses just waiting to tackle your toughest questions.



Adriana2345

Mathematics • 5 points

What is the value of 6(2b-4) when b = 5?

JeanaShupp

Answer: The value of $6(2b-4)$ is 36.

Step-by-step explanation:

Given expression: $6(2b-4)$

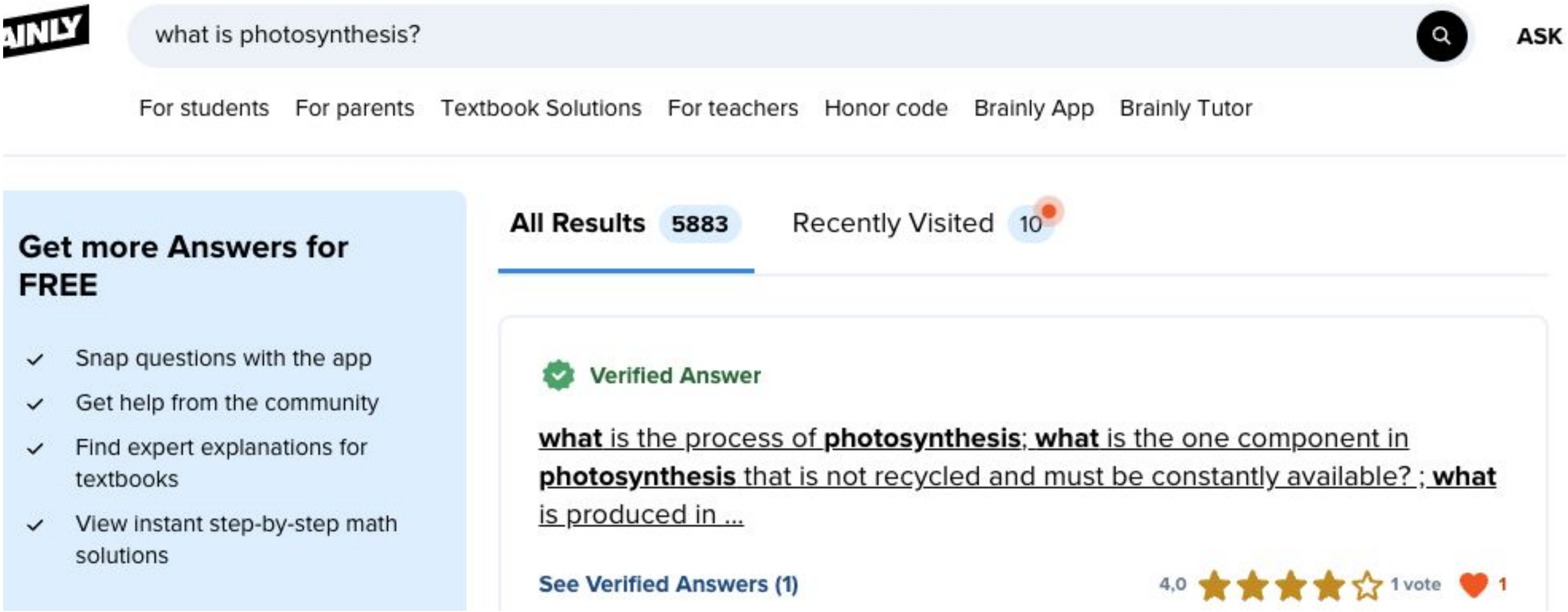
To find the value of $6(2b-4)$at b= 5, we need to substitute the b=5 in the expression, we get

$$\begin{aligned}&6(2(5)-4)\\&=6(10-4).......[\text{solve parentheses}]\\&=6(6)\\&=6\times6=36\end{aligned}$$

$$\Rightarrow 6(2(5)-4)=36$$

Therefore, the value of $6(2b-4)$ is 36, when b=5.

★★★★★ ♥



Get more Answers for FREE

- ✓ Snap questions with the app
- ✓ Get help from the community
- ✓ Find expert explanations for textbooks
- ✓ View instant step-by-step math solutions

Verified Answer

what is the process of **photosynthesis**; **what** is the one component in **photosynthesis** that is not recycled and must be constantly available? ; **what** is produced in ...

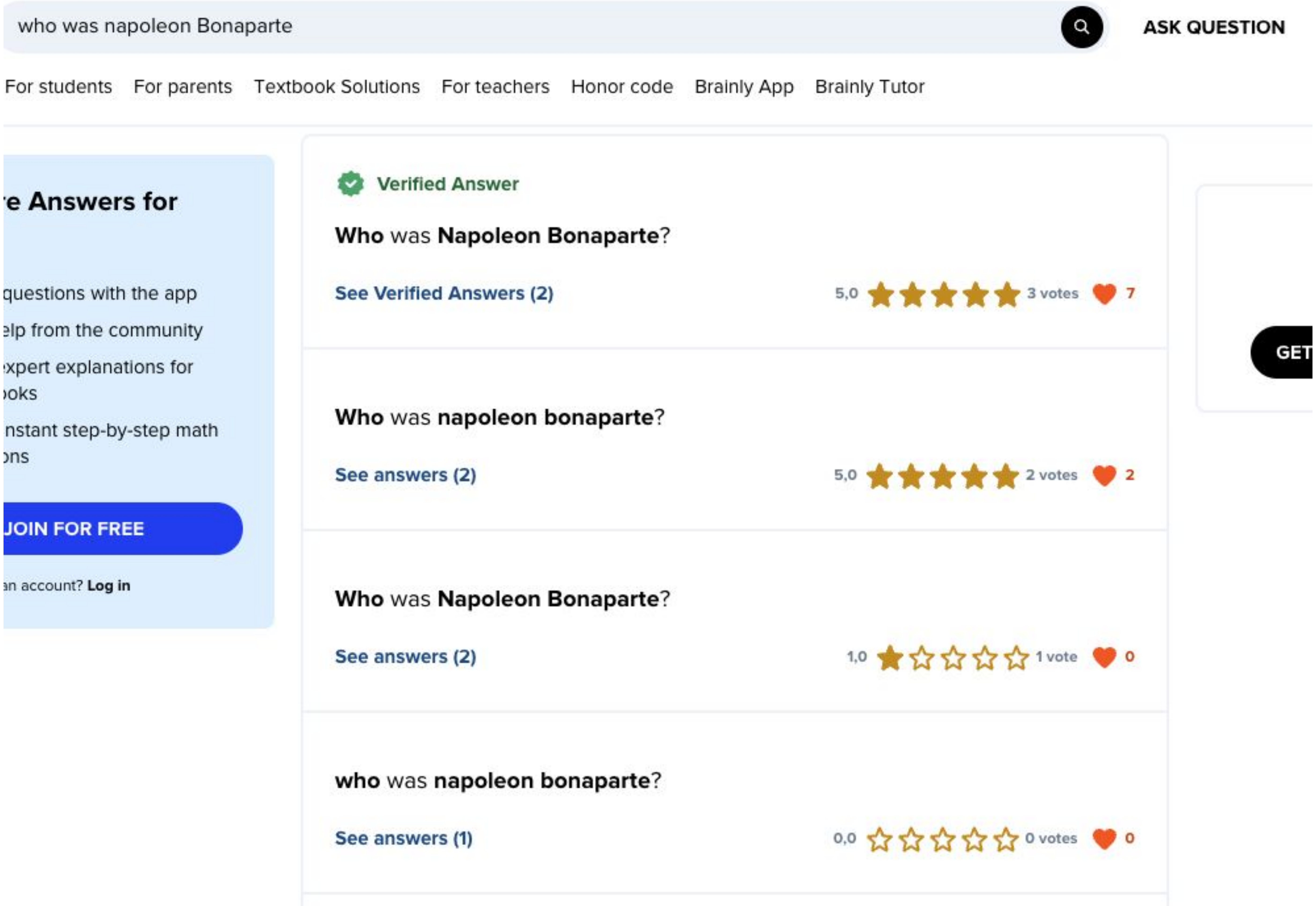
See Verified Answers (1)

4,0 ★★★★☆ 1 vote ♥ 1

February 2022

Educational, user-generated Content deduplication @Brainly

Vector vs Keyword



✓ Verified Answer

Who was **Napoleon Bonaparte**?

See Verified Answers (2) 5,0 ★★★★★ 3 votes ♥ 7

Who was **napoleon bonaparte**?

See answers (2) 5,0 ★★★★★ 2 votes ♥ 2

Who was **Napoleon Bonaparte**?

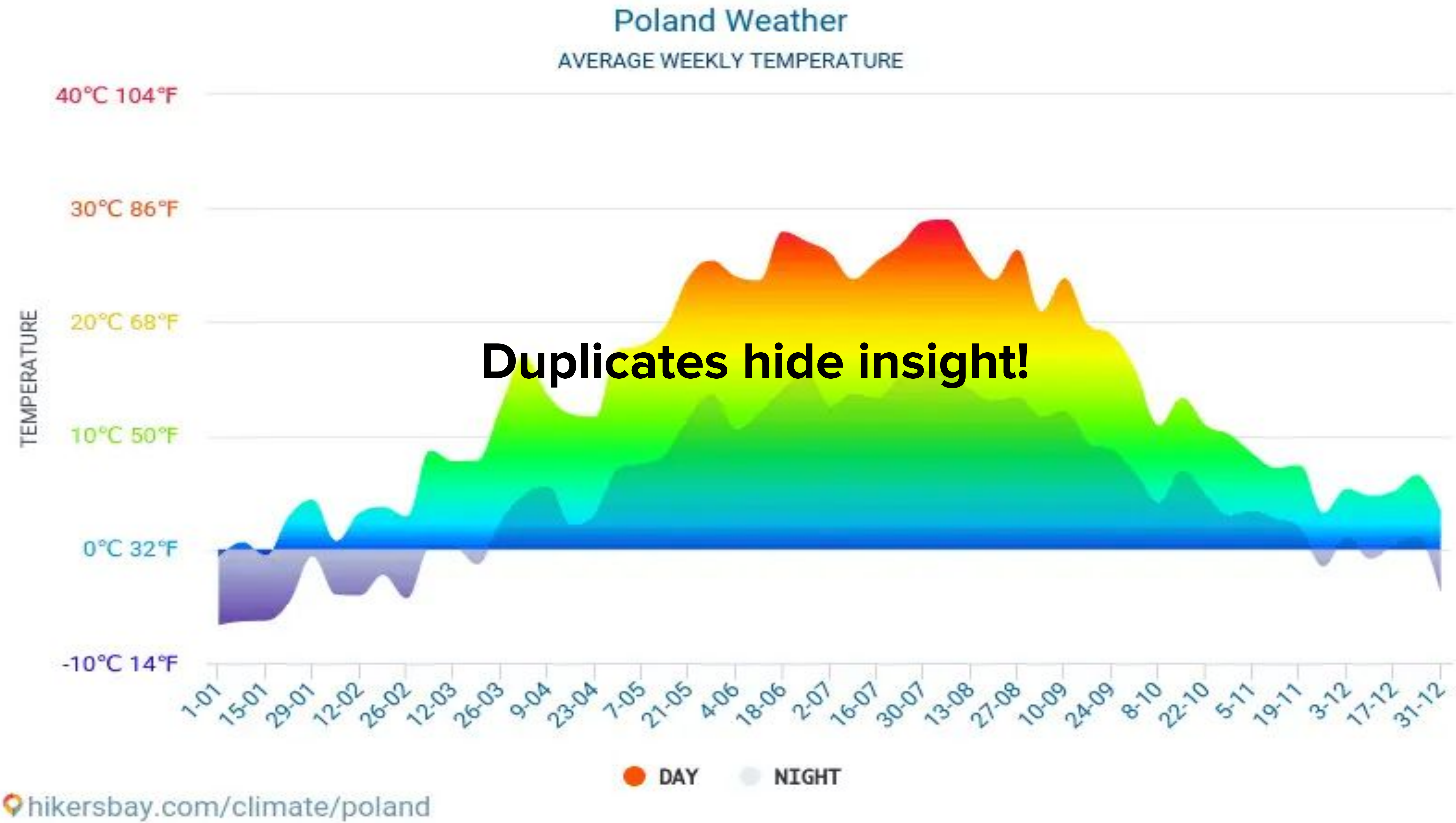
See answers (2) 1,0 ★☆☆☆☆ 1 vote ♥ 0

who was **napoleon bonaparte**?

See answers (1) 0,0 ☆☆☆☆☆ 0 votes ♥ 0

Relevant!

Taken from: https://www.flickr.com/photos/vizzzual-dot-com/2655969483



Duplicates hide insight!

hikersbay.com/climate/poland

February 2022

Educational, user-generated content deduplication @Brainly

Vector vs Keyword

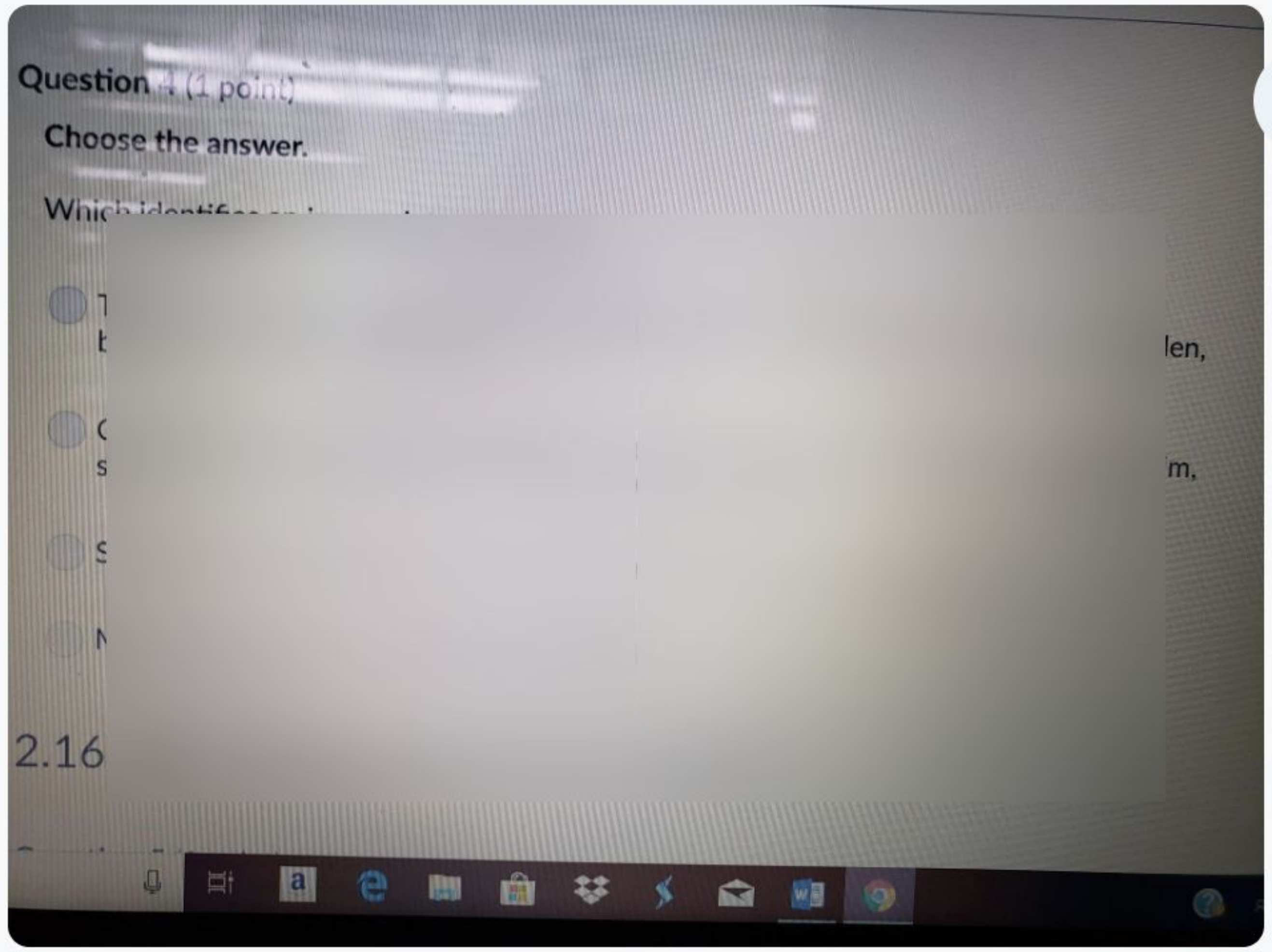
MEET THE USER

I need help... i don't really understand this...

In 2000, the circulation of a local newspaper was 3,250. In 2001, its circulation was 3,640. In 2002, the circulation was 4,100. a. Find the percent of increase in the newspaper's circulation from 2000 to 2001 and from 2001 to 2002. b. Which period had a higher percent of increase, 2000 to 2001, or 2001 to 2002?

Mr. Utterson had been some minutes at his post, when he was aware of an odd light footstep drawing near. In the course of his nightly patrols, he had long grown accustomed to the quaint effect with which the footfalls of a single person, while he is still a great way off, suddenly spring out distinct from the vast hum and clatter of the city. Yet his attention had never before been so sharply and decisively arrested; and it was with a strong, superstitious prevision of success that he withdrew into the entry of the court. What is the mood of the excerpt?

Which identifies an incomplete sentence



Question 4 (1 point)

Choose the answer.

len,

m,

2.16

February 2022

Educational, user-generated content deduplication @Brainly

Vector vs Keyword

Educational content is hard!

Diverse

Where did Napoleon die?

vs

How did Napoleon die?

The What

$$(x-1)(x+3)=15$$

vs

$$x^3+15=1$$

February 2022

Educational, user-generated content deduplication @Brainly

Vector vs Keyword

February 2022

Educational, user-generated content deduplication @Brainly

Vector vs Keyword

Approach I

1. Select a target content to deduplicate (curated, or externally sources)
2. Calculate cosine similarity against dataset to be deduplicated
3. Select >0.9 matches
4. ???
5. Profit!

Approach I - results

Recall	Precision
0.420	**0.895**

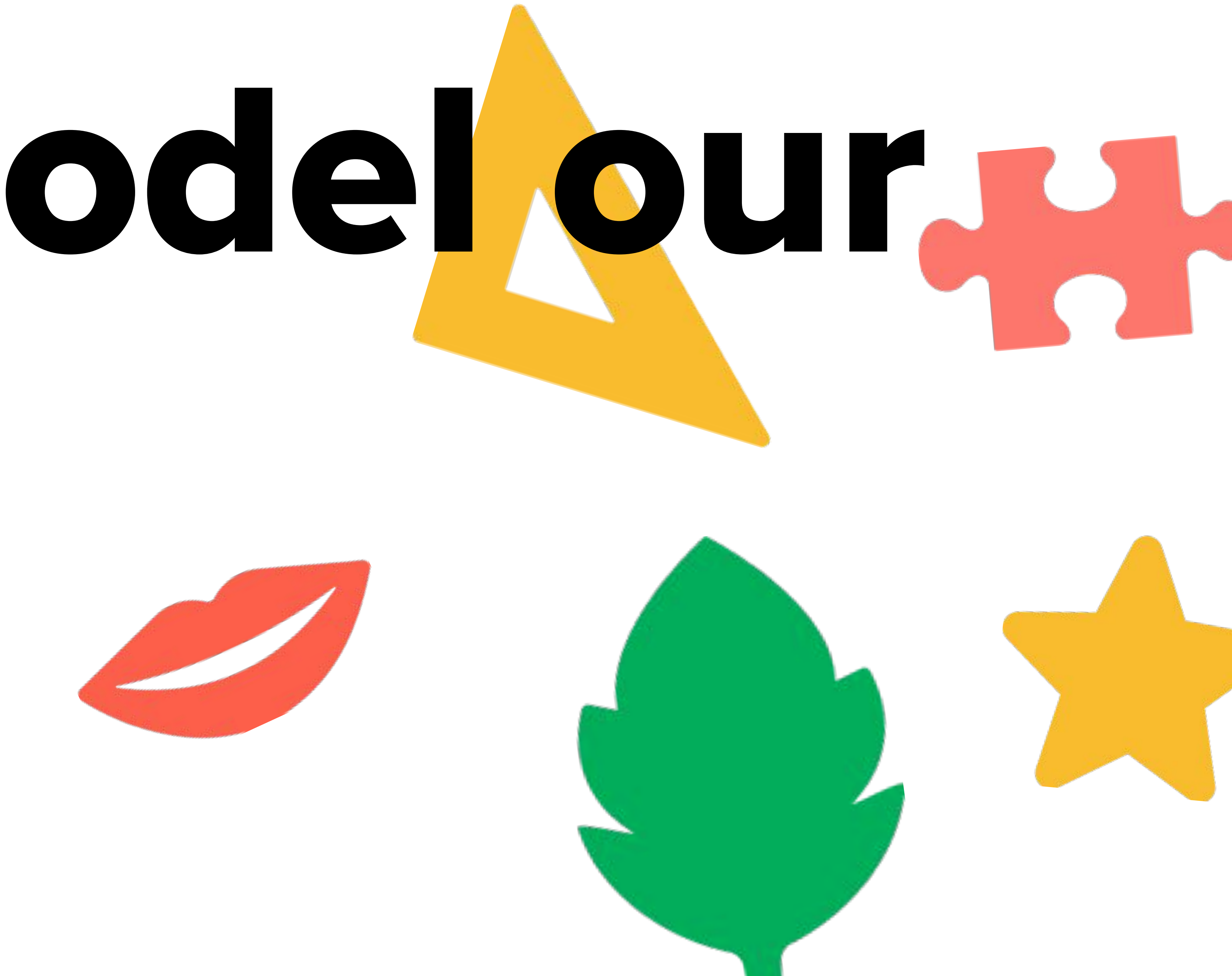
Approach I - outcome

Way too slow to deduplicate larger amounts of content

Computationally expensive

Not really impressive, result-wise

Make the model our own.



Use preexisting deduplication models as a base.

Many preexisting deduplication models.

Light text preprocessing.

Use FAISS index.

Approach “II”

1. Train model based on quora deduplication dataset and own Brainly content
2. Create FAISS index out of content to deduplicate
3. Select a target content to deduplicate (curated, or externally sources)
4. Search the FAISS index
5. Select >0.9 matches
6. ???
7. Profit!

Approach II - results

Recall

0.512

Precision

0.898

Approach II - outcome

Much faster, but results still not great

A lot of work just to try out something new

STOP
HAMMER TIME

WON'T SOMEBODY PLEASE
THINK OF THE CONTENT??
imgflip.com

V

KNOW THE USER

Approach Keyword

1. Create Opensearch index out of content to deduplicate
2. Prepare a strict strategy (with score thresholds)
3. Select a target content to deduplicate (curated, or externally sources)
4. Search the Opensearch index
5. ???
6. Profit!

Approach Keyword - results

Recall	Precision
0.543	**0.902**

Approach Keyword - outcome

Results are somehow... better?

We can experiment with different strategies quickly

We're accidentally matching incomplete questions...

...which actually is quite helpful

Can we get better results with ML?
Probably, but at what cost?

ML costs

Single match run

10$

Single model training

100$

ML costs

Single match run

10$

model training

X * 100$

Opensearch costs

“Training” instances hourly

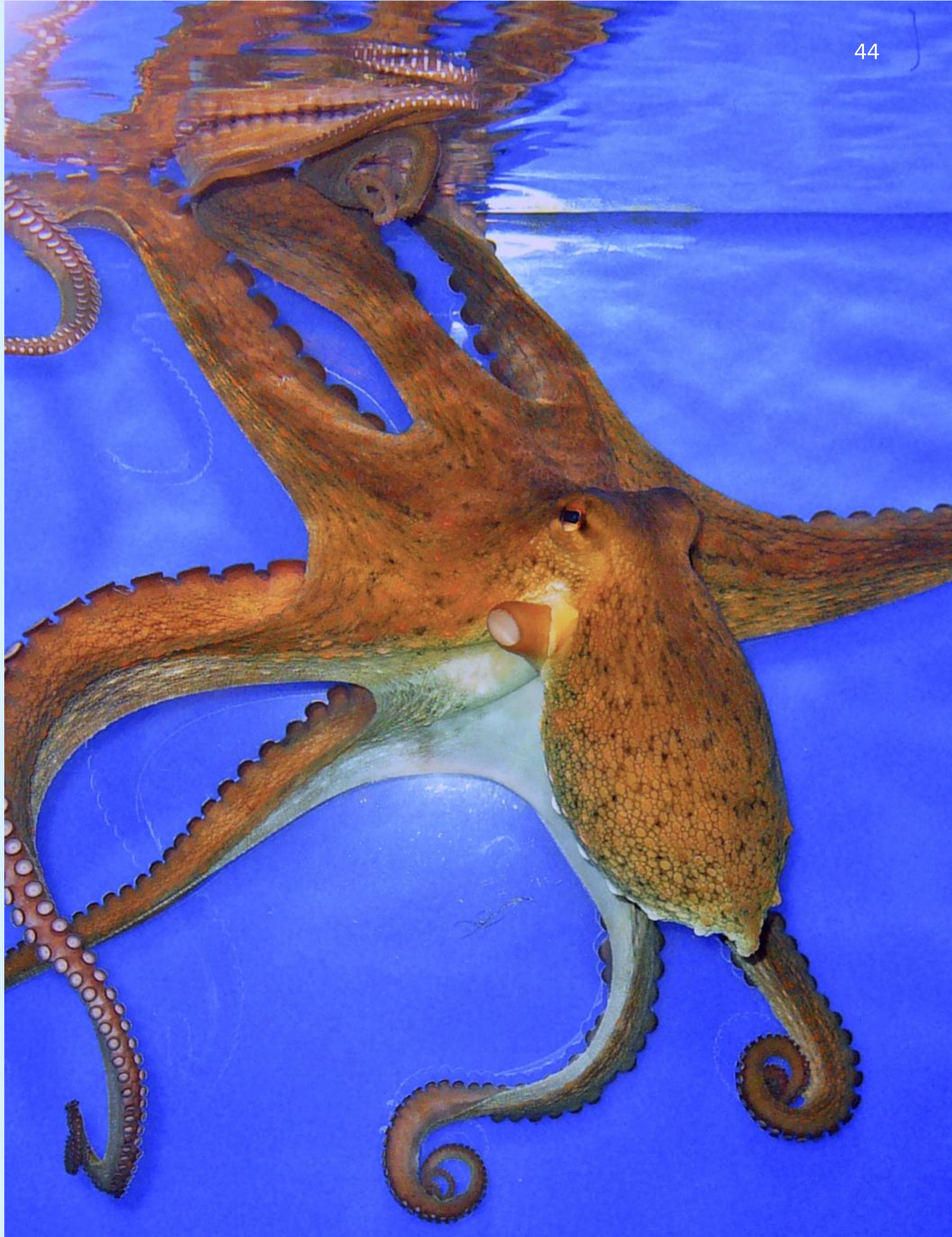
0.5$

Peak (final matching) cost

30$

MLOps vs DevOps

Taken from https://commons.wikimedia.org/wiki/File:Octopus_vulgaris_02.JPG



The Future

Take-aways

Not all user generated content is made equal

Pareto to rule them all

Everything has a cost

THANK YOU!