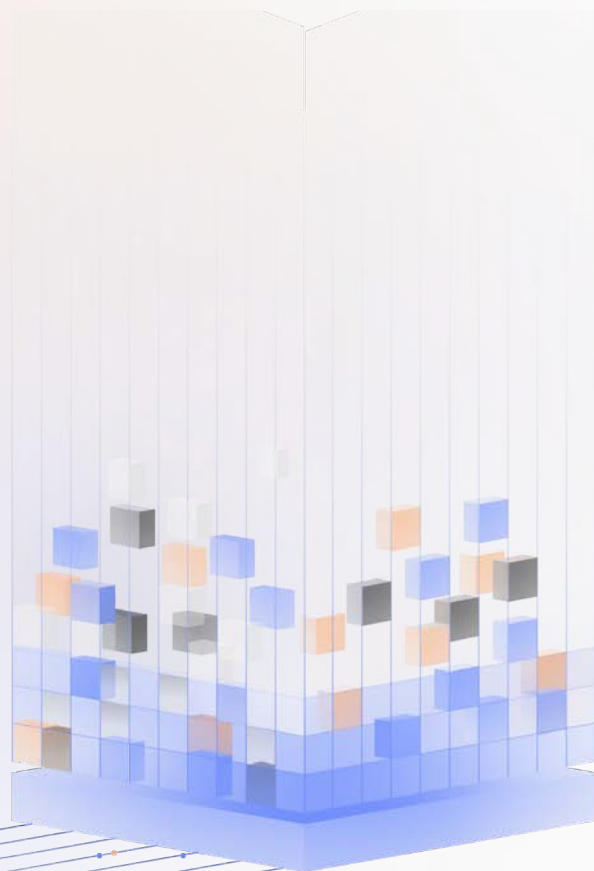


## Vector Ops

How to run vector  
embedding-powered  
apps in production



Overview

# Building vector-powered apps

**Why** are vectors useful?

**What** can you build with them?

**How** do you do that?

WHY

# What did we lose with language?



$[(37, 232, 113), (113, 17...$

1M pixels

WHY

# What did we lose with language?



Field of grass

3 words

WHY

# What did we lose with language?



**Field of grass**

WHY

# What did we lose with language?



Field of grass,  
**on a summer day**

WHY

# What did we lose with language?



Field of grass,  
on a summer day,  
**a few clouds in the  
blue sky**

WHY

# What did we lose with language?

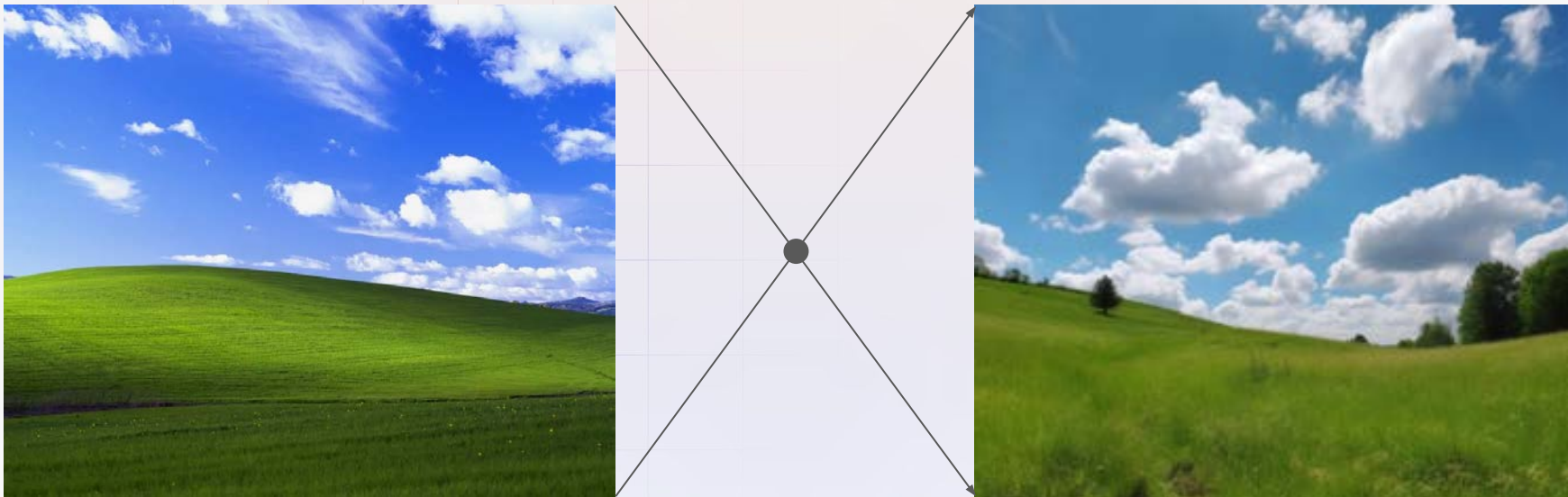


Field of grass,  
on a summer day,  
a few clouds in the  
blue sky, **slightly hilly**



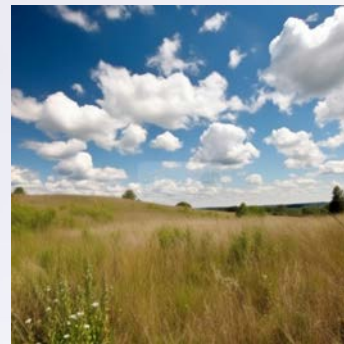
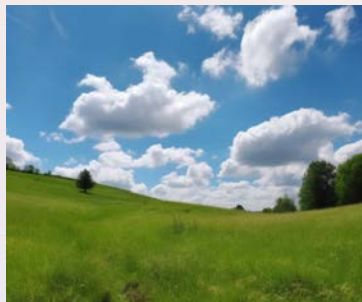
WHY

# Natural language is a bottleneck



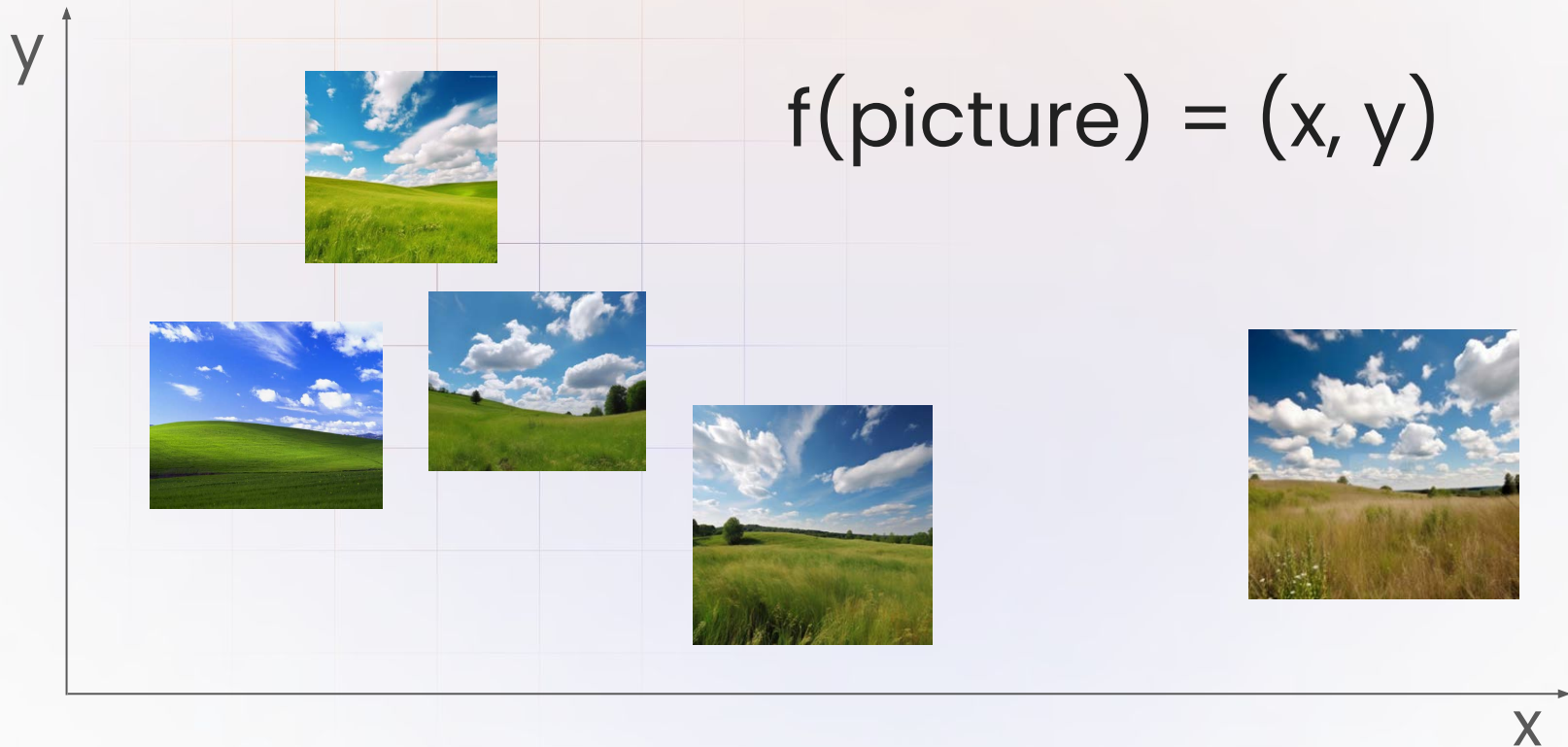
WHY

# Natural language is ambiguous



WHY

# Vectors are better!



WHY

# Vectors are (mostly) better!

✓ expressive

✓ smooth

✗ difficult to work with

WHAT

# Search before vectors

Custom NLP pipelines  
Operate a full-text index  
Fine-tune rules for years

*... and still get it wrong!*

The screenshot shows a search interface with a dark blue header. On the left is a red 'W' logo and a 'Menu' dropdown. On the right is a search bar containing the text 'splitting headache'. Below the search bar, the text reads 'Partial matches found for "splitting headache"' and 'No results for complete search term'. A product card is displayed below, featuring a blue 'New' badge and an image of the 'Aura Cacia Uplifting Kit' which includes bottles of Peppermint, Lemon, Lime, and Sweet Orange essential oils. The product name is 'Aura Cacia Uplifting Essential Oils Kit - 0.25 fl oz x 4 pack' and the price is '\$19.99'. Below the price, it indicates 'Not sold in stores' with a red 'X' and 'Shipping' with a green checkmark. At the bottom of the card is a quantity selector set to '1' and an 'Add for shipping' button. At the very bottom of the page, there are navigation arrows and the text 'Page 1 of 1'.

WHAT

# Recommendations before vectors

Content feature extraction  
Collaborative filtering  
Retrieve 10,000 candidates  
Score their fit one-by-one

*... and still get it wrong!*


Jobs based on your Profile



Co-Founder / CTO

Stealth

Zurich, Switzerland (Hybrid)

 Actively recruiting

1 hour ago ·  Easy Apply



AIML - Intern (ML Engineer),  
Intelligence

Apple

Zurich, Zurich, Switzerland



9 connections work here

2 weeks ago

WHAT

# Search & recommendations with vectors

Deep(er) retrieval with ANN\*

Content & User Vectors

Query manager on top

**Simpler.**

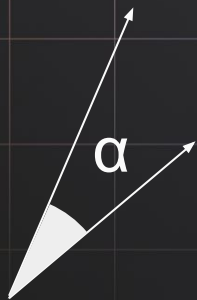
**Better.**

**Faster.**

HOW

# Approximate Nearest Neighbours?

Find nearest neighbours fast.

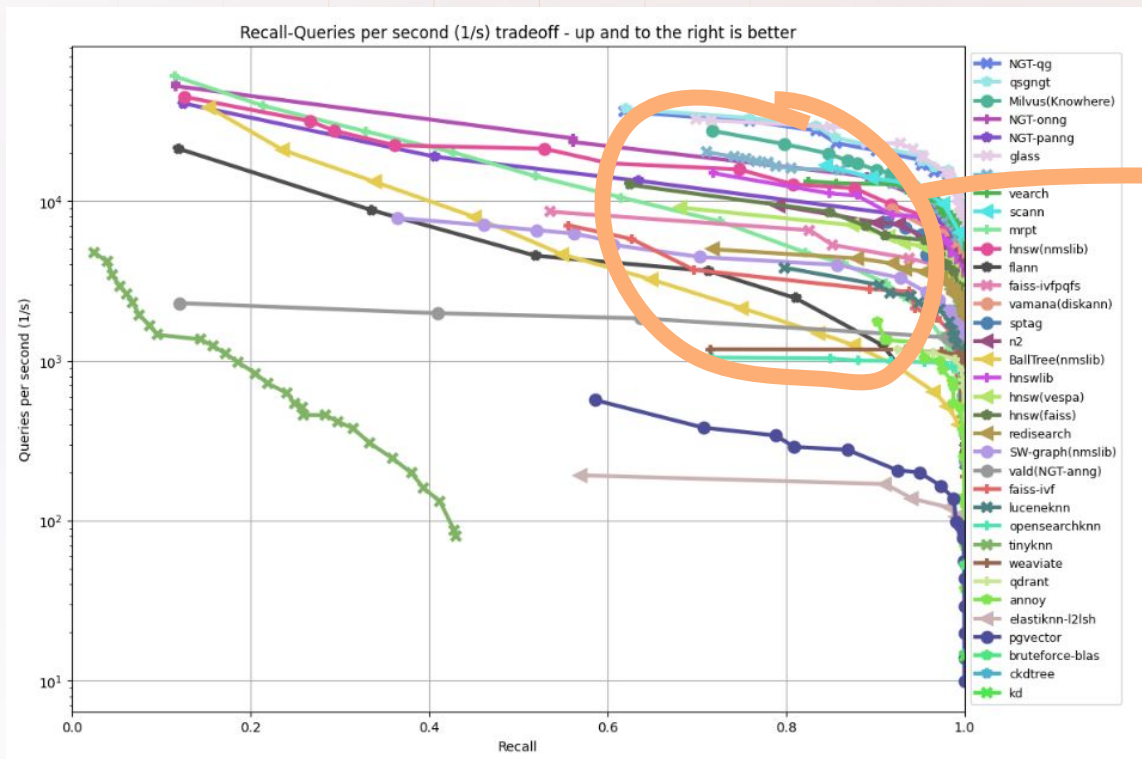


*Thousands of QPS per machine with 10s of thousands of vectors.*



HOW

# If it can be measured...



Good  
enough!

~800 dims, 10k vectors

HOW

# Building the content vectors ..

## Content vector

[0.5483, 0.1629, 0.8897, 0.4201, 0.7765, 0.9532, 0.2988, 0.7014, 0.0796, 0.6243, 0.3897, 0.9134, ..]



HOW

# .. and user vectors in the same space!

## User vector

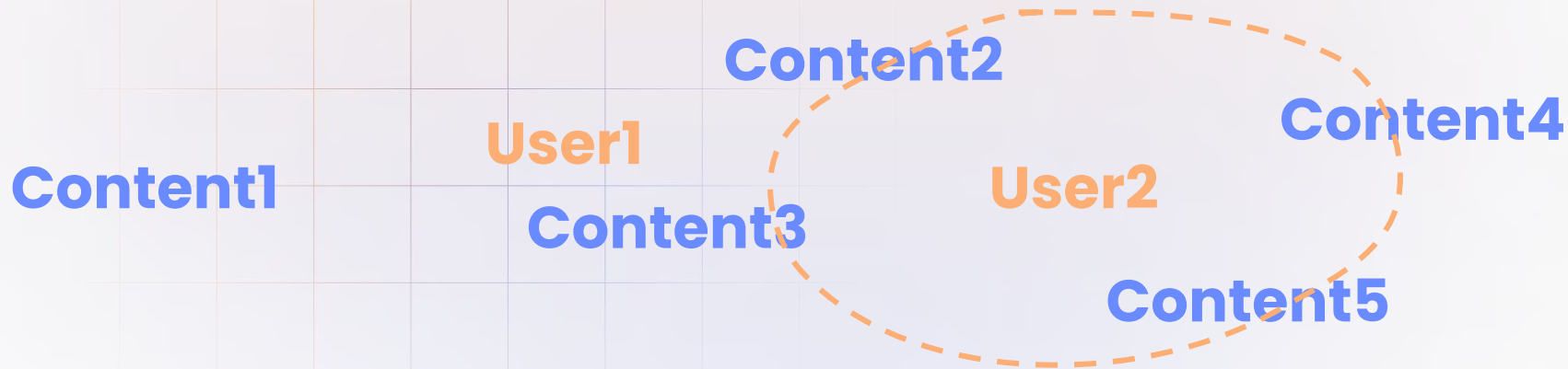
[0.7219, 0.1243, 0.9937, 0.3351, 0.6794, 0.2568, 0.8193, 0.4782, 0.9056, 0.7129, 0.5897, 0.2224, ...]



*User <> User search!*

HOW

# Let ANN do the heavy lifting



HOW

# Query manager on top

Manipulate the search vector

Issue multiple searches

Combine/filter results

*Personalized  
search is near!*

*Diversity,  
variation etc.*

*Guarantees, experiments &  
slippery slope!*

HOW

# What will you need to get started?

Content data

*Unstructured text  
or images*

Python notebook

Vector embedding model

sklearn cosine\_similarity

*huggingface.co*

*OpenAI API*

*Great example here.*

HOW

# What will you need for an MVP?

+ Vector Database

+ Evaluation

*Understand  
pricing!*

*Eyeballing*

*Quantitative*

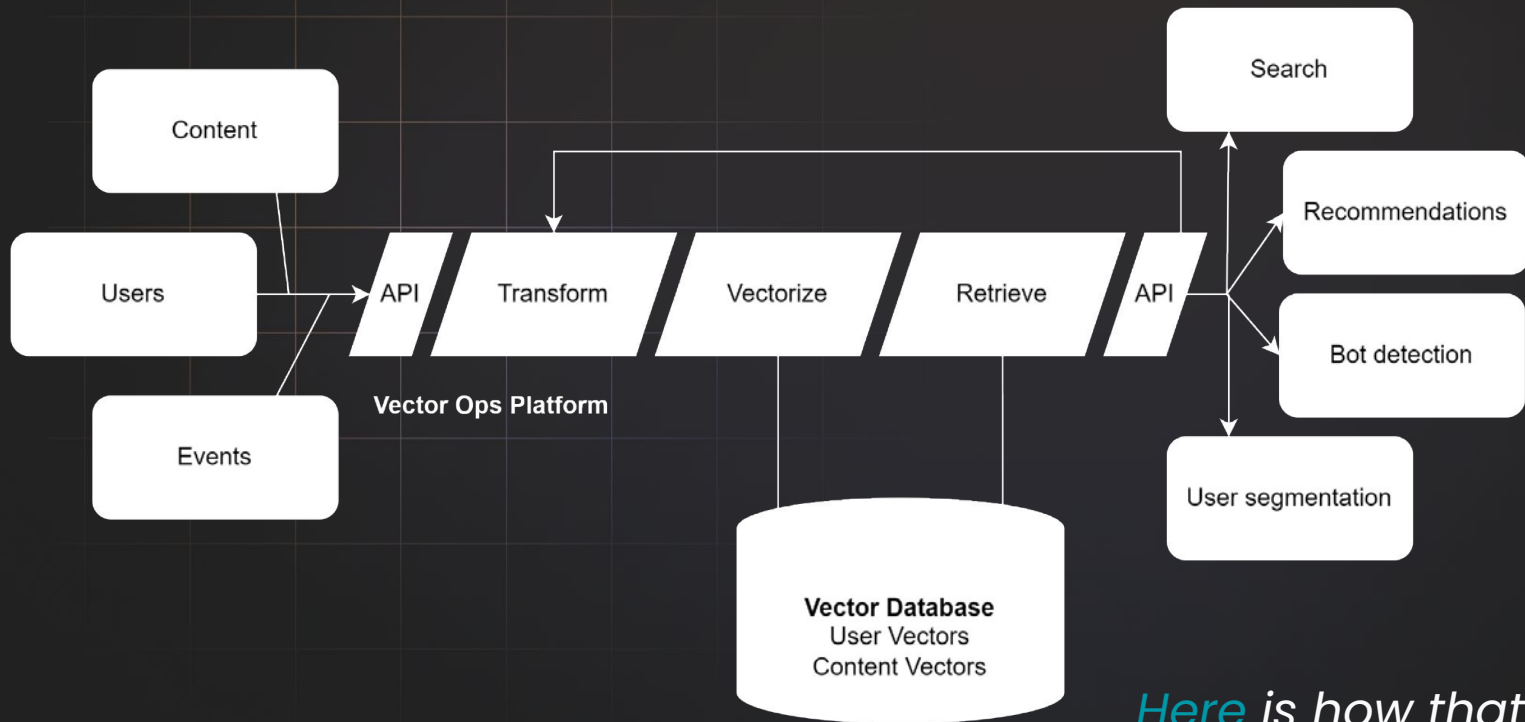
*User feedback*

*Analytics*

Great example [here](#).

HOW

# Towards a #VectorOps platform



*[Here](#) is how that could look.*



HOW

# What about generative AI?!



HOW

# Chat & chains!



Langchain 13 min explainer [video](#).

HOW

# What the frig?



Let's play a game - you will make up a number between 1 to 10 and I'll try to guess it. After each of my guesses, you tell me if your secret number is higher or lower until I guess it correctly.



Sure, I've got a number between 1 and 10 in mind. Go ahead and make your first guess!



First guess: 5



My secret number is lower than 5. Try again!



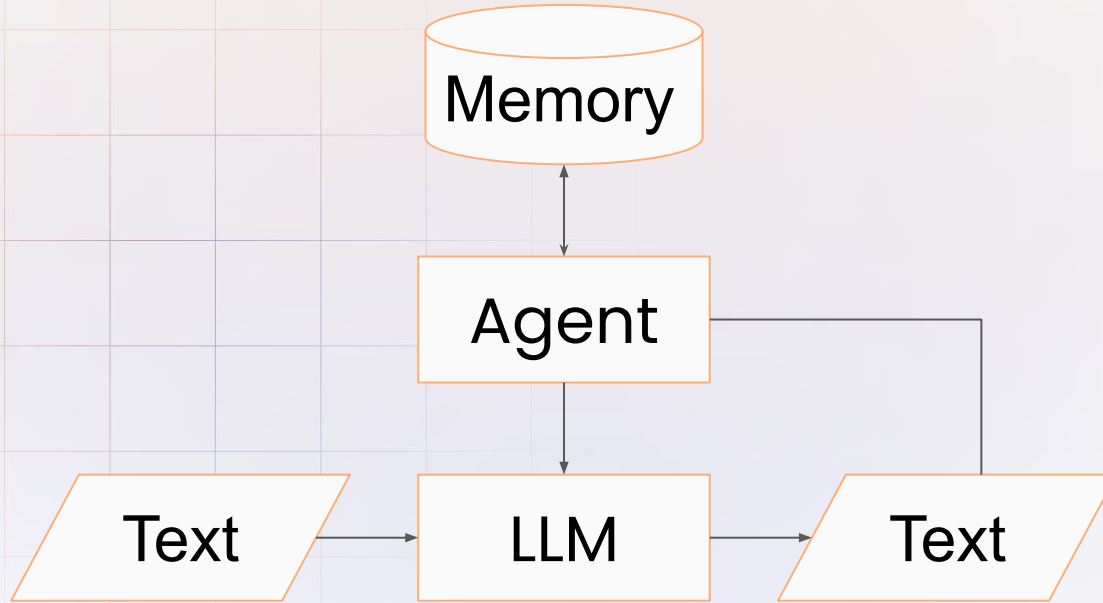
Second guess: 3



You got it! My secret number is indeed 3. Well done!

HOW

# Agents & memory!



Some agents are autonomous eg [AutoGPT](#).

# Let's connect!



[linkedin.com/in/svonava](https://www.linkedin.com/in/svonava)

and

Learn more at [superlinked.com](https://superlinked.com)

