AI with Your Own Data

Nuri Halperin | https://linkedin.com/in/nurih

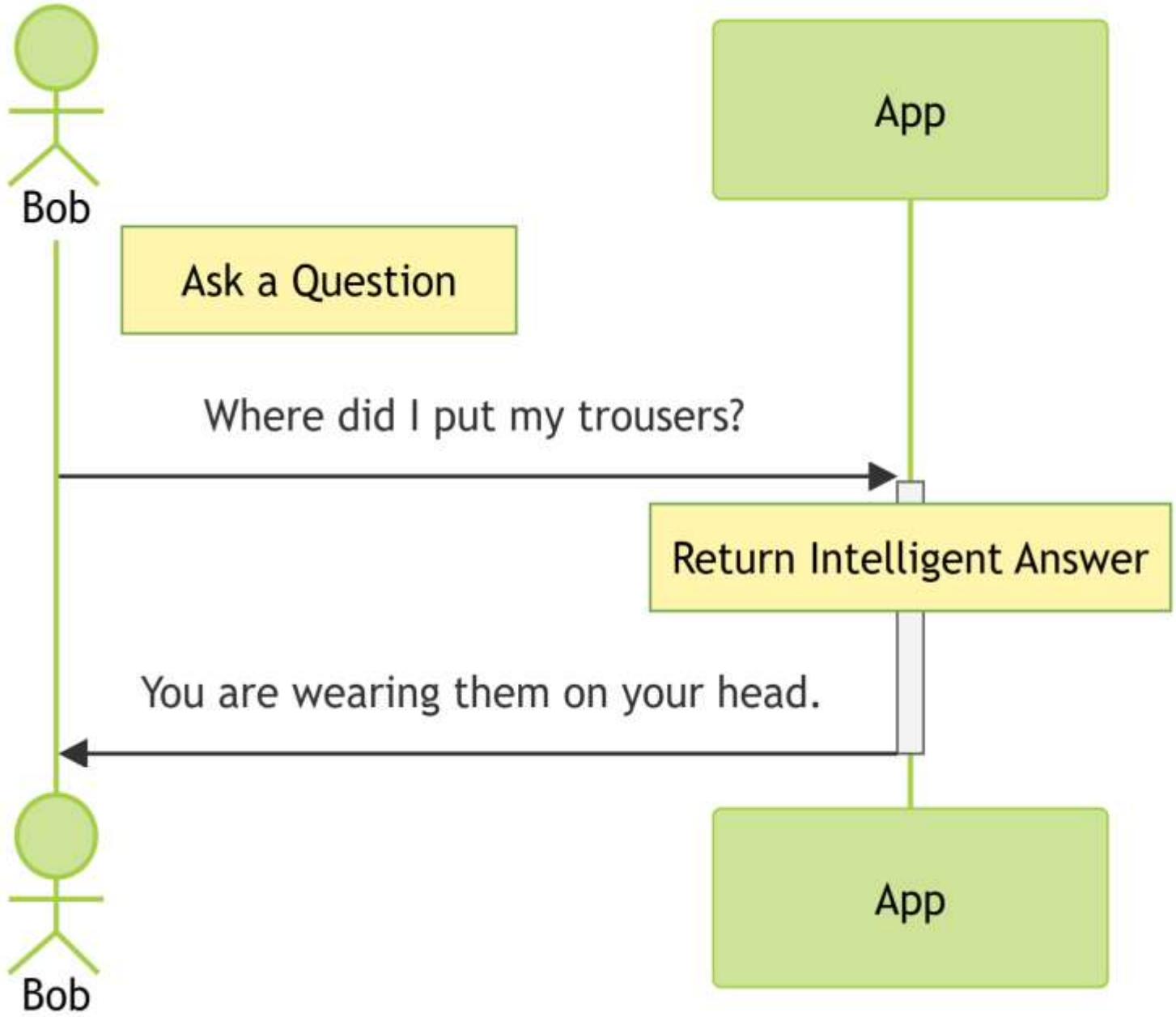
Agenda

- There is no AI, only math
- It's all in the process

What Are We After?

1. User askes question
2. Machine answers:
    - Using MY data
    - Using MY language
    - Responds like an expert
    - Intelligent, infers, less-than-perfect matching

What We Want

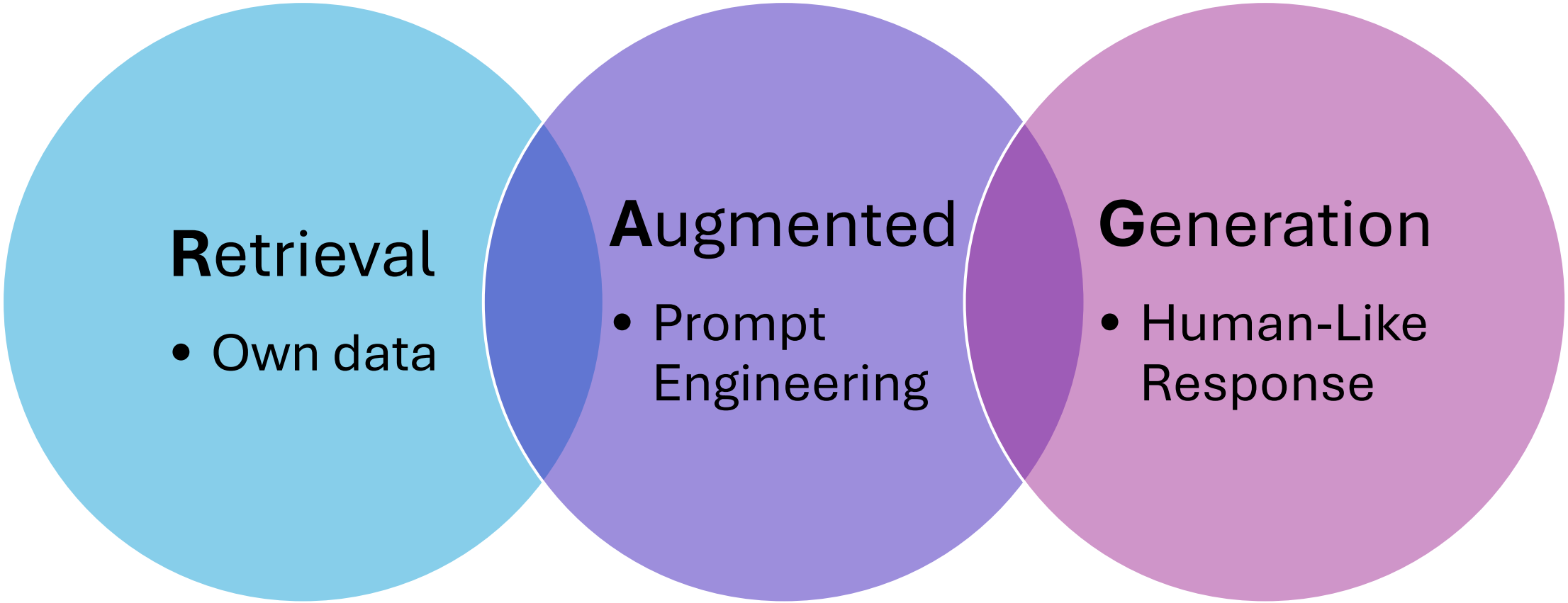


Why Not Search Engine?

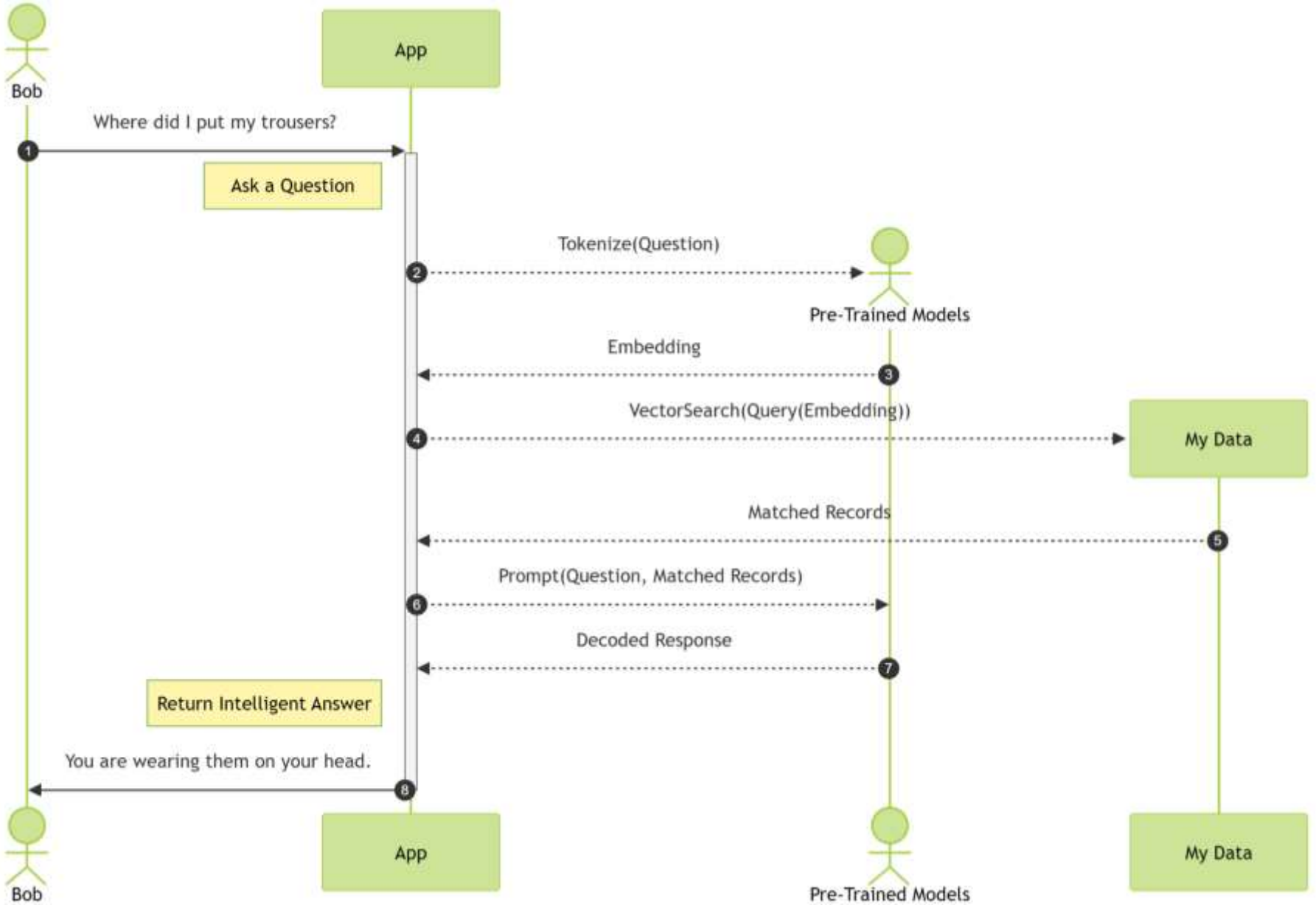
- Keyword based
- Synonyms
- Stemming

1. ☑ User askes question
2. Machine answers:
   - ☑ Using MY data
   - ☒ Using MY language
   - ☒ Responds like an expert
   - ☒ Intelligent, infers, less-than-perfect matching

What is RAG?



What We Are In For...



What do Models do?

- Compute "meaning" by "nearness" of vectors – similarity.
- "Completion"

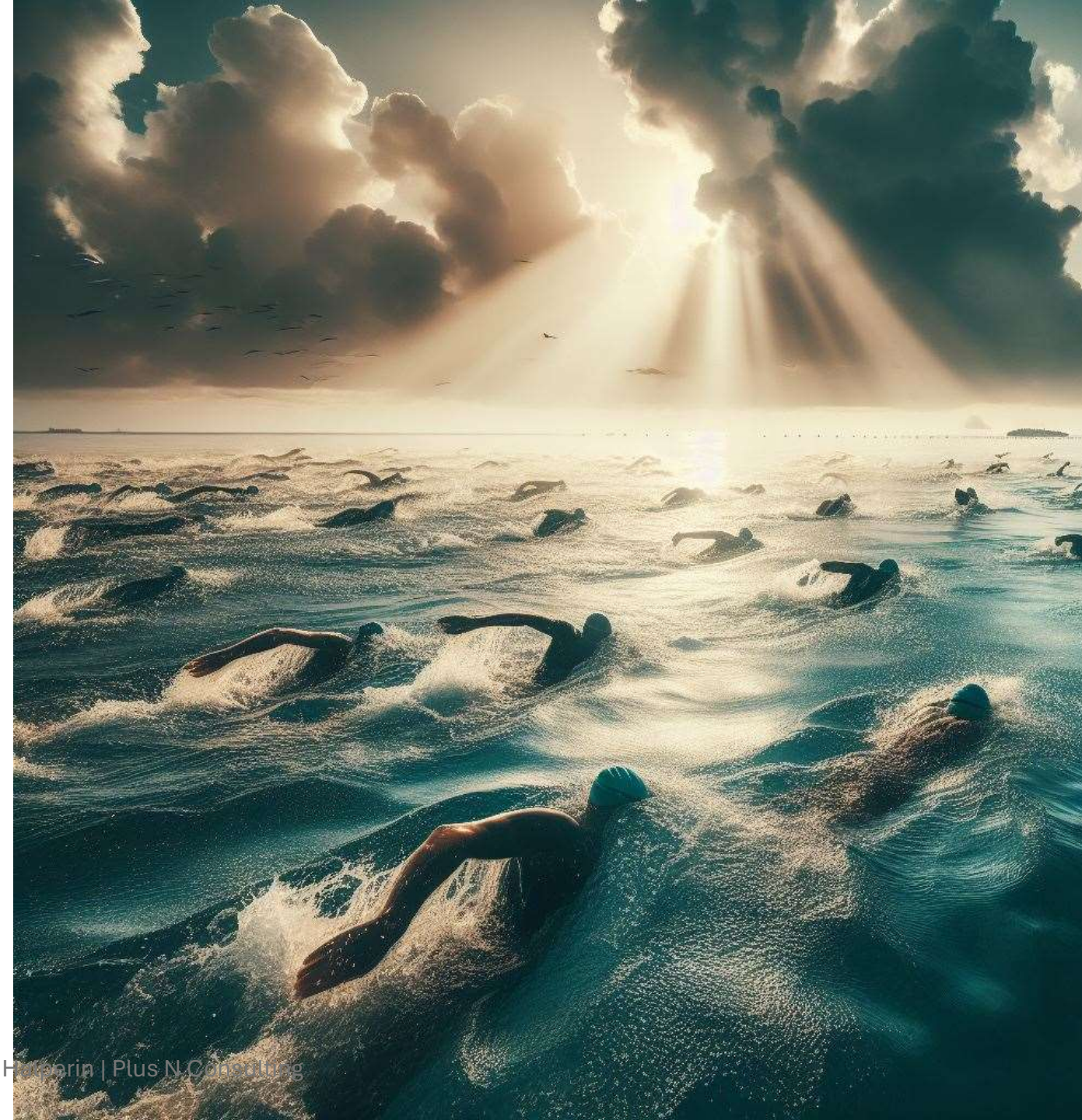
How are models built?

- Tokenize input text/images into numbers
- Create vectors of these numbers
- Encoded: string -> token sequence -> math::vector
- Decoded: math::vector -> token sequence -> string
- Train model, then save it

Training a Model

- Action of mutating interior vectors by changing formula coefficients until output is satisfactory
- Iterative, expensive to compute, random element

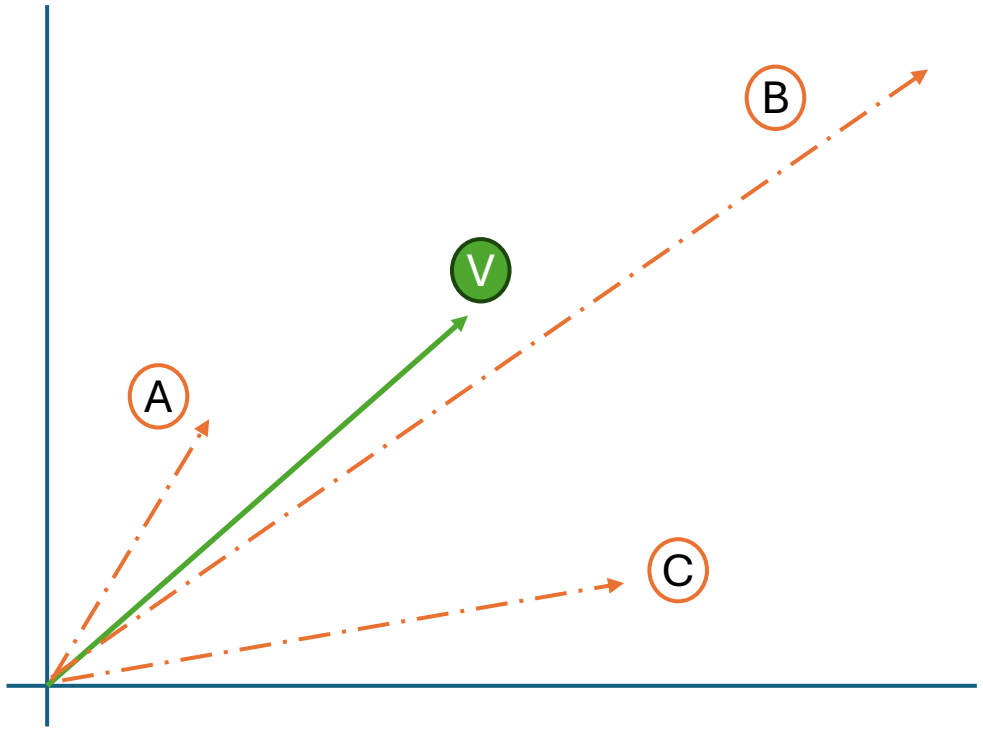
- We can use pre-trained models (Yippie!)



Vector Comparison

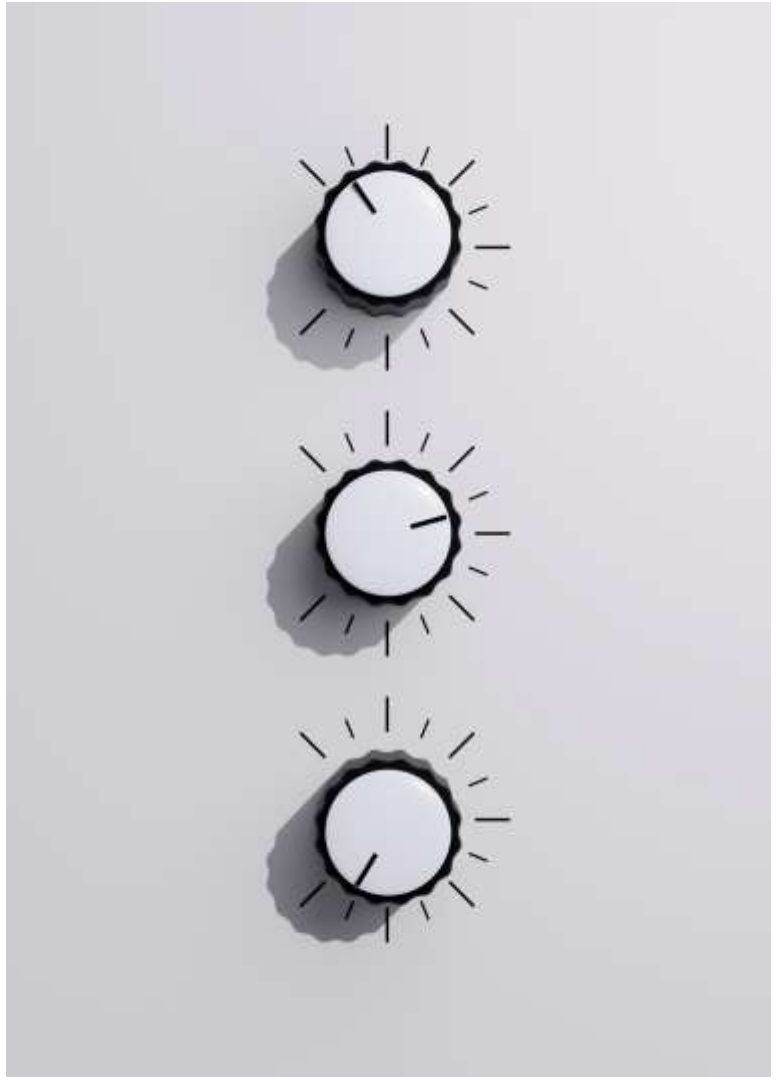
Demo uses Cosine

- Vectors can be compared in N-space
- Comparison Types
  - L2 – Euclidian Distance
  - Cosine
  - Dot Product



Vector Comparison Methods

L2 Euclidian	Distance
Cosine	Angle Direction
Dot Product	Magnitude Direction



"Vector Search"

TL;DR: Specialized DB index for vector similarity

Vector Search Implementation

KNN/ANN:

K-Nearest Neighbor

Approximate-Nearest Neighbor

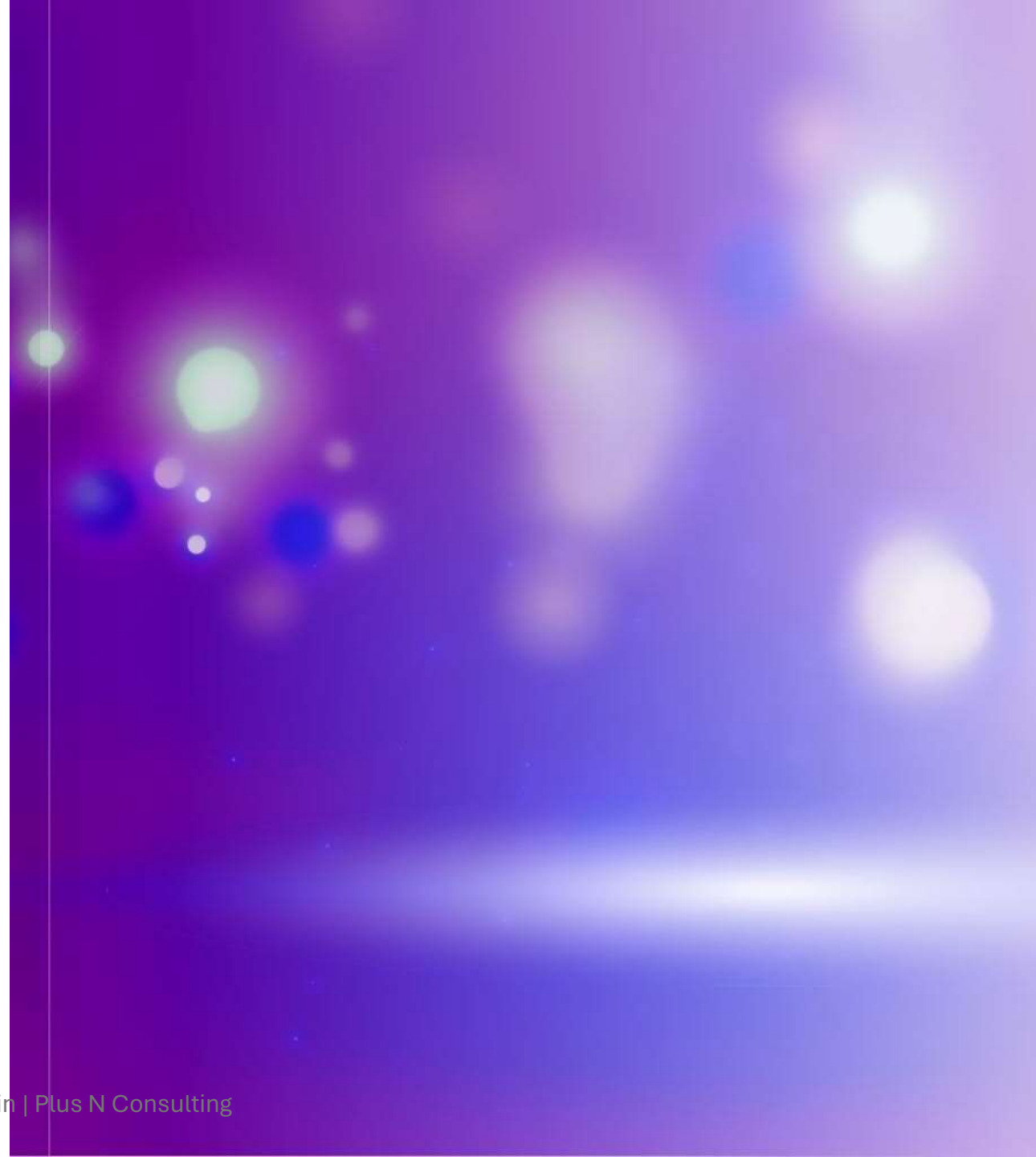
HNSW: Hierarchical Navigable Small Worlds

Implementation of ANN / KNN

Given a vector query, returns data associated to vectors that are "near" it

Pieces of the Puzzle

- **Token**: A number representing an input element
- **Tokenizer**: t(text) -> "token list" the "vocabulary of inputs"
- **Embedder**: e(tokens) -> "embedding"
- **Model:**
  - Encode: f(text) -> "vector"
  - Decode: f(vector) -> "text"
- **Training**: The act of making the fidelity of Decode(Encode(<<some-value>>)) as similar as possible.



Demo Time!

Squee!

Thank You!
Stay Connected

Nuri Halperin

+N Consulting, Inc.

nuri@plusnconsulting.com

https://linkedin.com/in/nurih

Consulting, Training, Inspiring