Research

“How to do cool stuff and get paid for it”

Today's Speakers

Complex data

David Procházka

Learned Metric Index

Jaroslav Olľha

AlphaFind

Terézia Slanináková

PhD study

Miriama Jánošová

What is research?

What is research?

Working on cutting edge things

Designing experiments

Staying abroad

Programming

Going to conferences

Realizing your ideas

Teaching*

Presenting

Doing cool stuff

Being paid

Making friends

Writing publications

Creating diagrams

Being creative

Supervising students*

Meeting clever people

What is research?

Doing cool stuff

How to make cool stuff?

Either find *something* you think is cool,
or find *someone* who makes cool things.

CODA Research Group

We find patterns in data and mine information from complexity.

- We use **Python, Rust, PyTorch, Docker, Kubernetes, JupyterHub, ...**
- We work with **images, proteins, human motion, ...**
- We cooperate with partners from **Switzerland, Denmark, Germany, ...**
- We organize invited talks from **Kiwi.com, JAMF, SAP, ...**

Complex Data

big

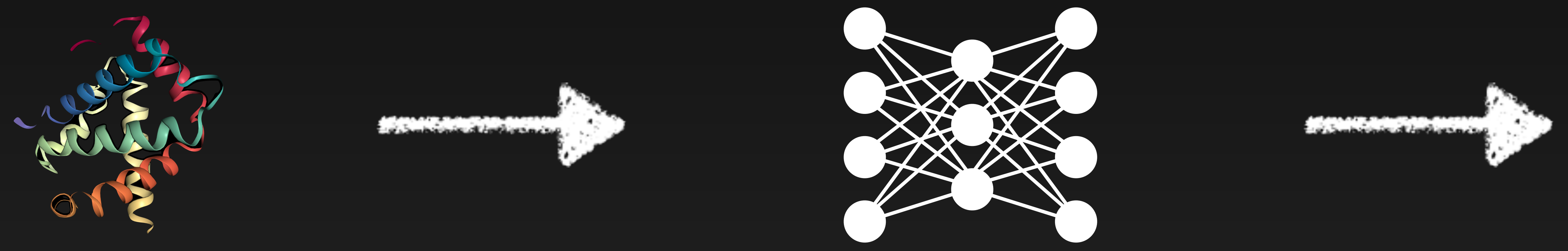
Today’s data is complex and that is a problem...

abundant

How does **Spotify** recommend **similar songs**?

How does **Netflix** determine what should you **watch next**?

Everything can be a vector...



(0.9259, -0.4775, ..., 0.7019, -0.5630)

Complex object

Embedding model

High-dimensional dense vector embedding

Dimensionality of Embeddings

Data Source	Dimensionality
DINOv2 (image)	384 – 1,536
CLIP (image + text description)	512 – 1,024
Llama 3 (text)	**4,096 – 16,384**

(0.9259, -0.4775, ..., 0.7019, -0.5630)

(7.1041, -3.8554, ..., -12.5602, 0.9923)

(-2.5482, 0.2563, ..., 3.002, 8.8223)

(3.8520, -39.459, ..., 15.3019, -1.0592)

(1.3222, -0.8269, ..., 7.3929, 4.6901)

...

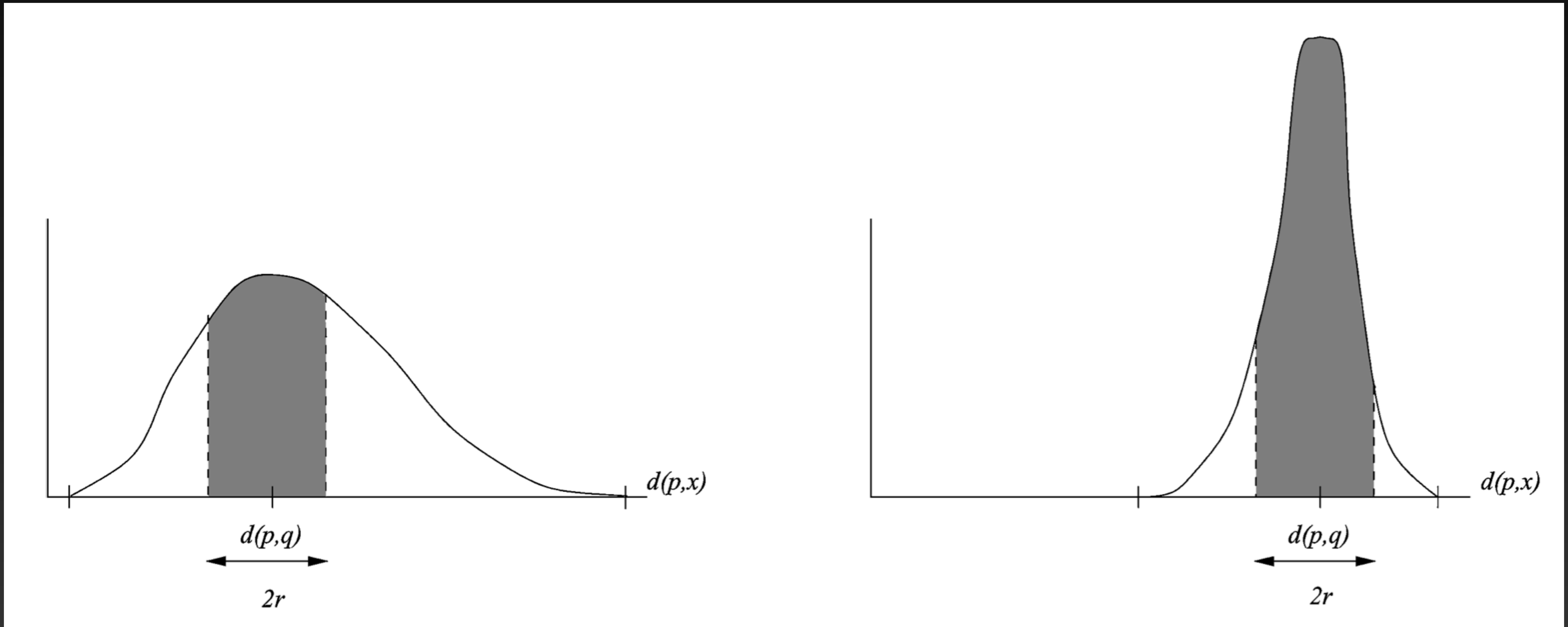
1,000,000,000+
vectors

1,000+
dimensions

4+ TB
of memory

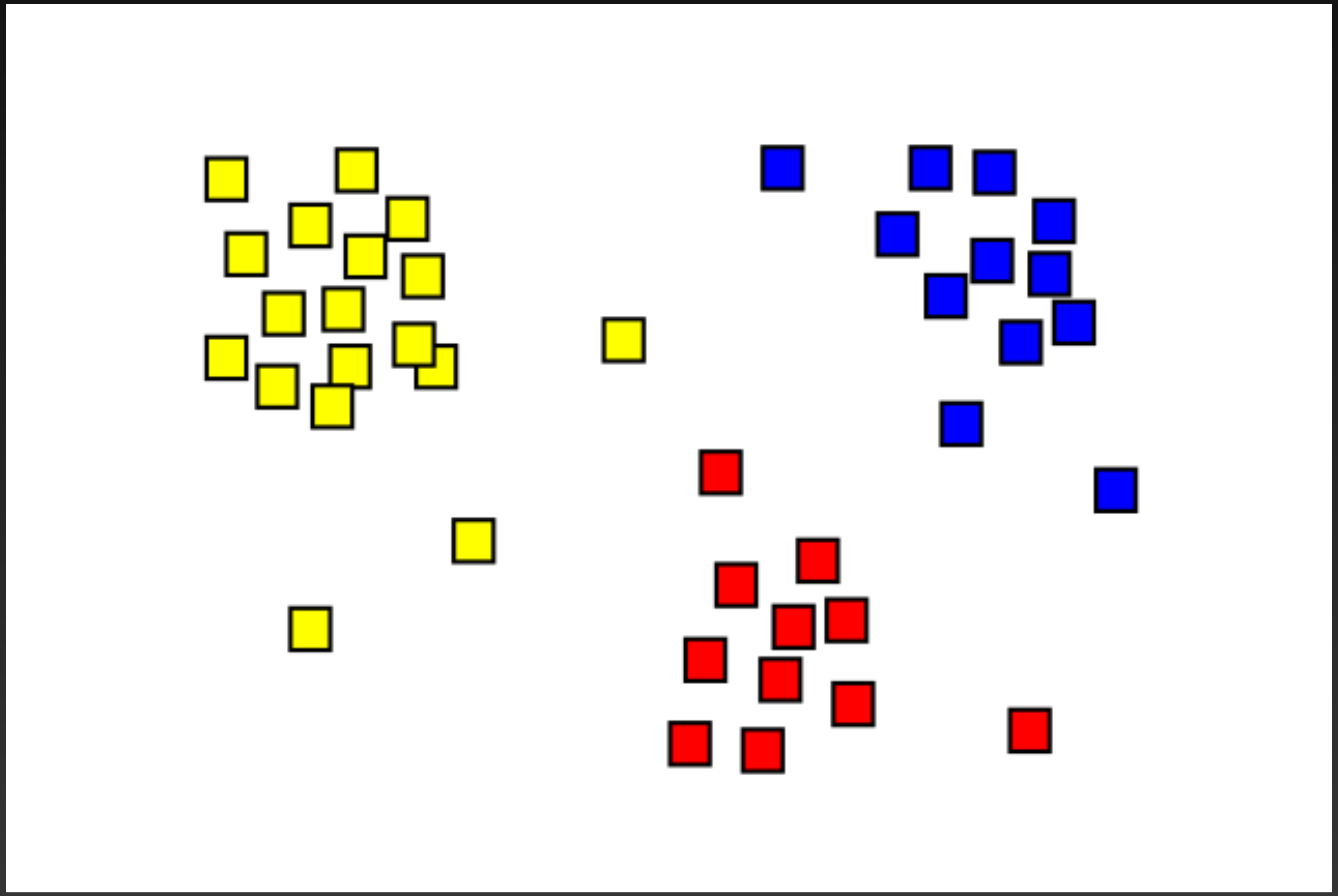
Any two vectors have similar distance

Low-dimensional dataset VS **High-dimensional** dataset

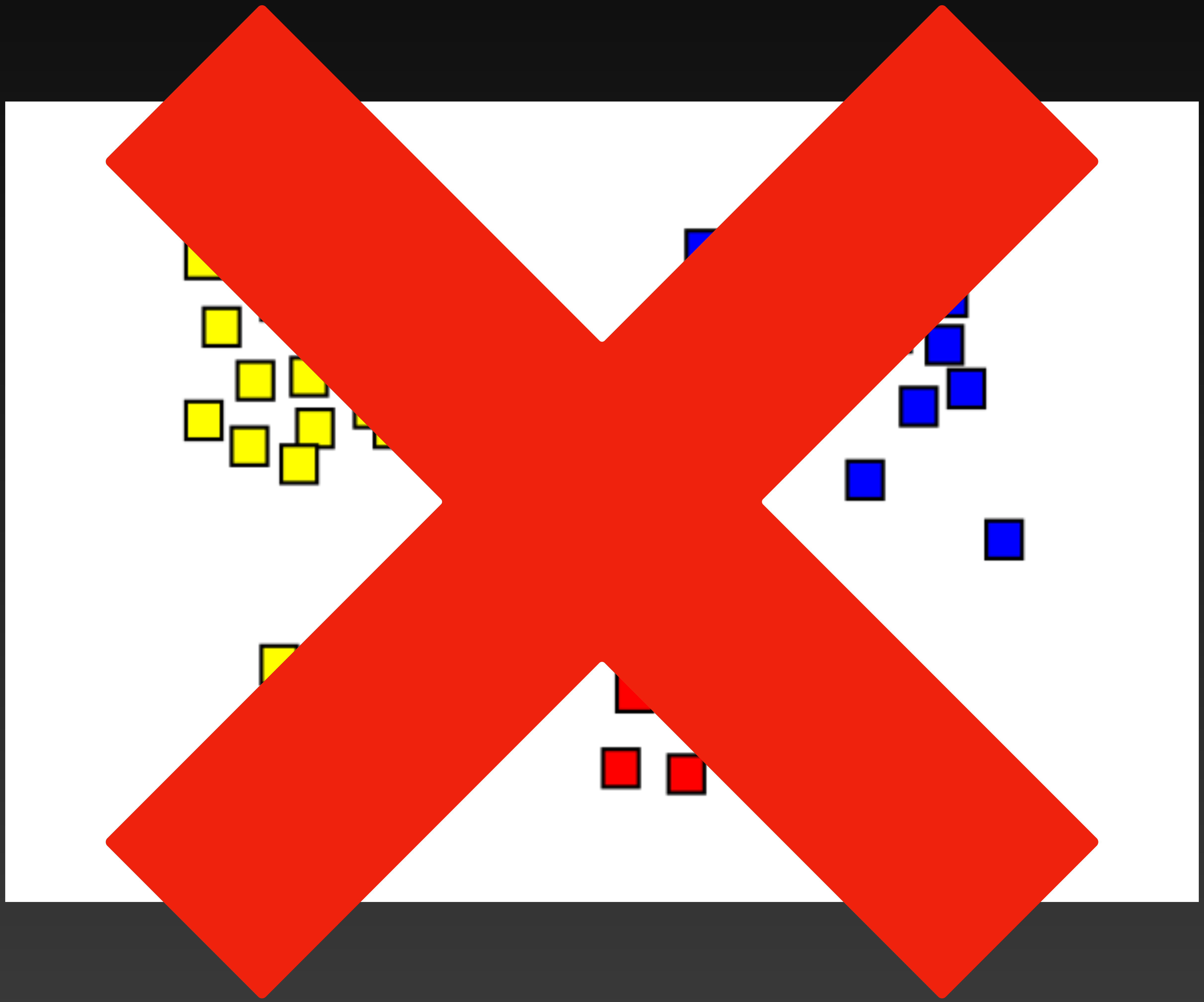


Source: Chávez et al., Searching in metric spaces, 2001

Clusters?



NO!



Curse of Dimensionality

1. Problems get **exponentially** harder
2. Any two **vectors have similar distance**
3. **All vectors** are near **orthogonal**
4. **No** notion of **locality**
5. **No clusters**

(0.9259, -0.4775, ..., 0.7019, -0.5630)

(7.1041, -3.8554, ..., -12.5602, 0.9923)

(-2.5482, 0.2563, ..., 3.002, 8.8223)

(3.8520, -39.459, ..., 15.3019, -1.0592)

(1.3222, -0.8269, ..., 7.3929, 4.6901)

...

1,000,000,000+
vectors

1,000+
dimensions

4+ TB
of memory

Learned Metric Index

Next-generation indexing for high-dimensional data

Learned Indexing for 1D data

Kraska et al. 2018

tree → model

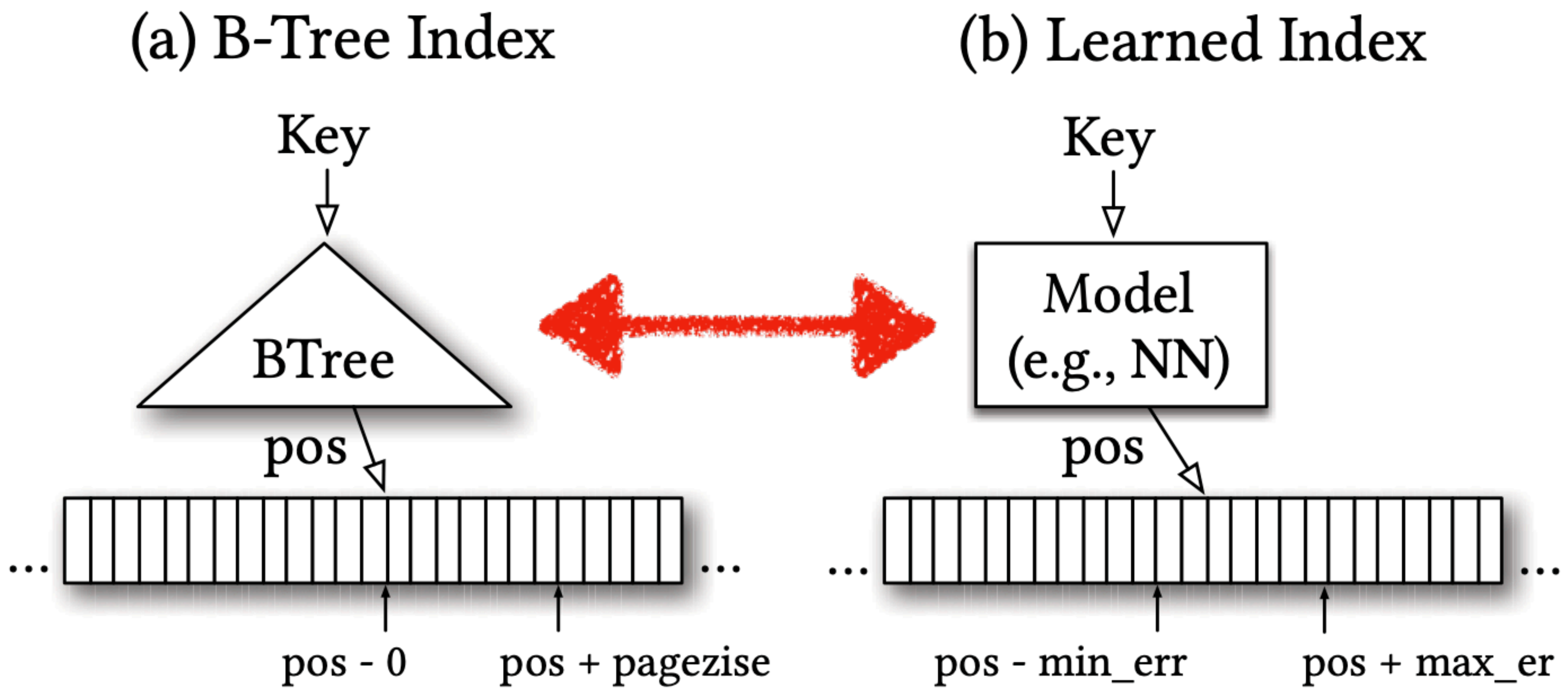


Figure 1: Why B-Trees are models

Learned Indexing for 1D data

Kraska et al. 2018

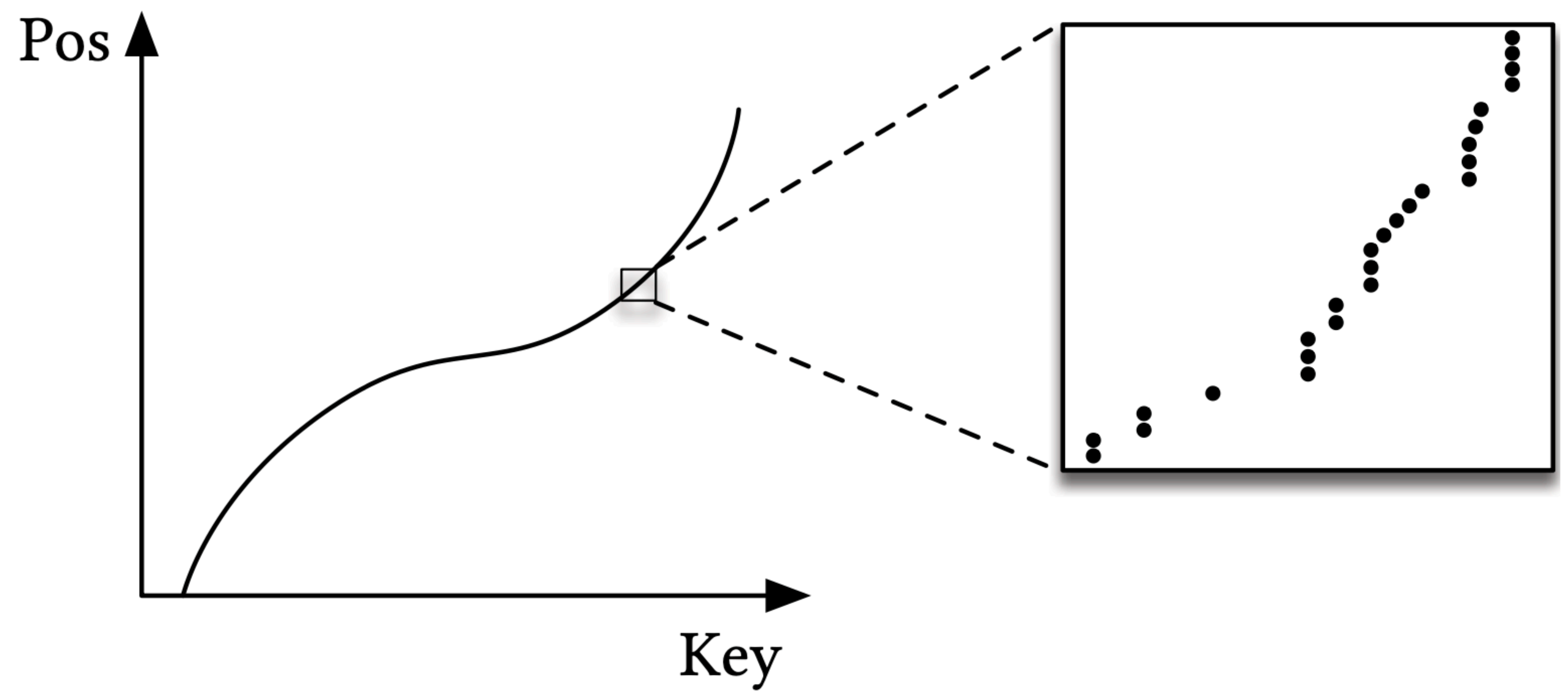


Figure 2: Indexes as CDFs

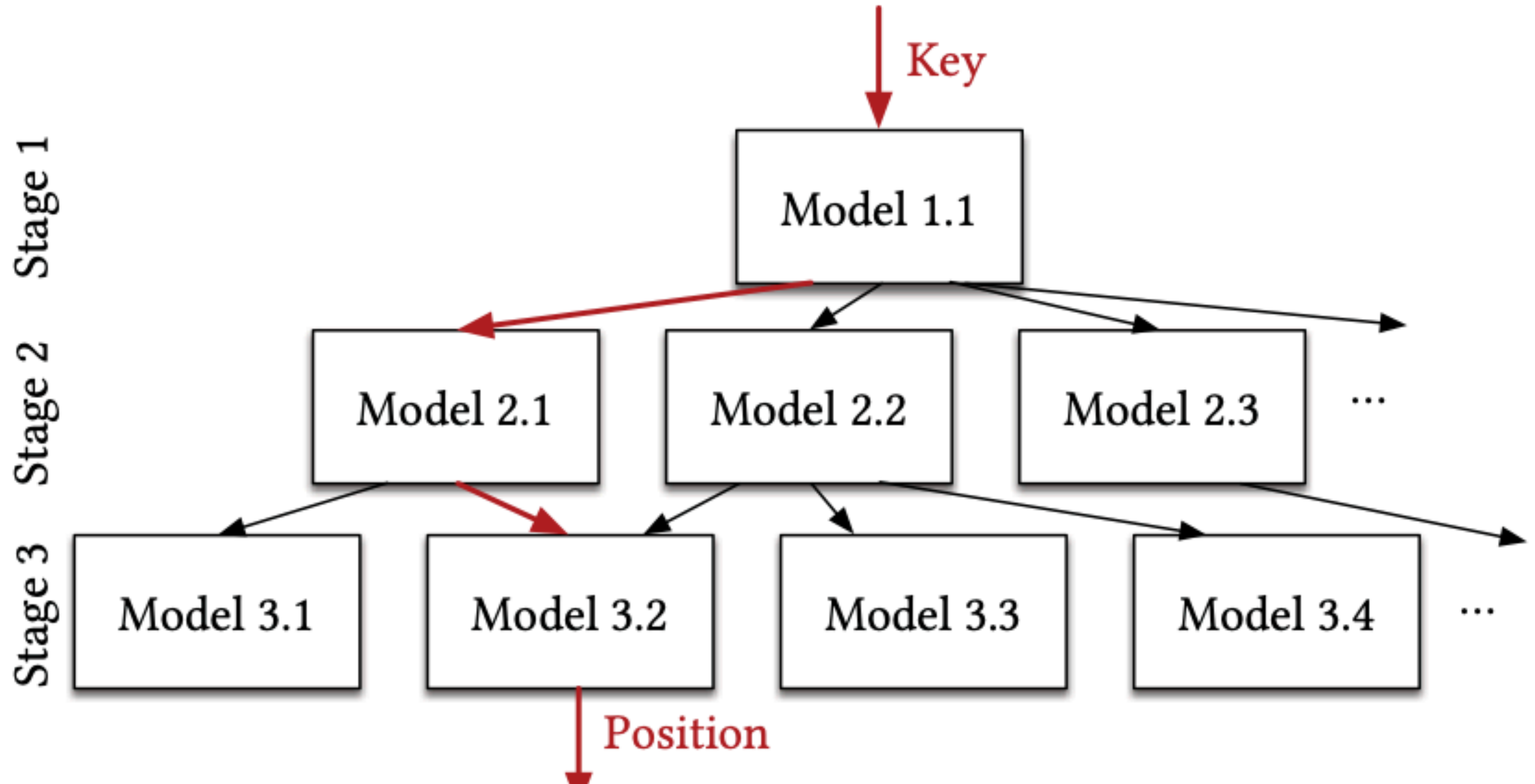


Figure 3: Staged models

Why? Because we could reduce O(log n) to O(1).

Learned Indexing

There is more to it

One-dimensional data 1D	Learn cumulative distribution function.
Multi-dimensional data 2 to 20D	Transform into 1D case.
High-dimensional data 20D+	Learn existing clustering or iteratively improve one.

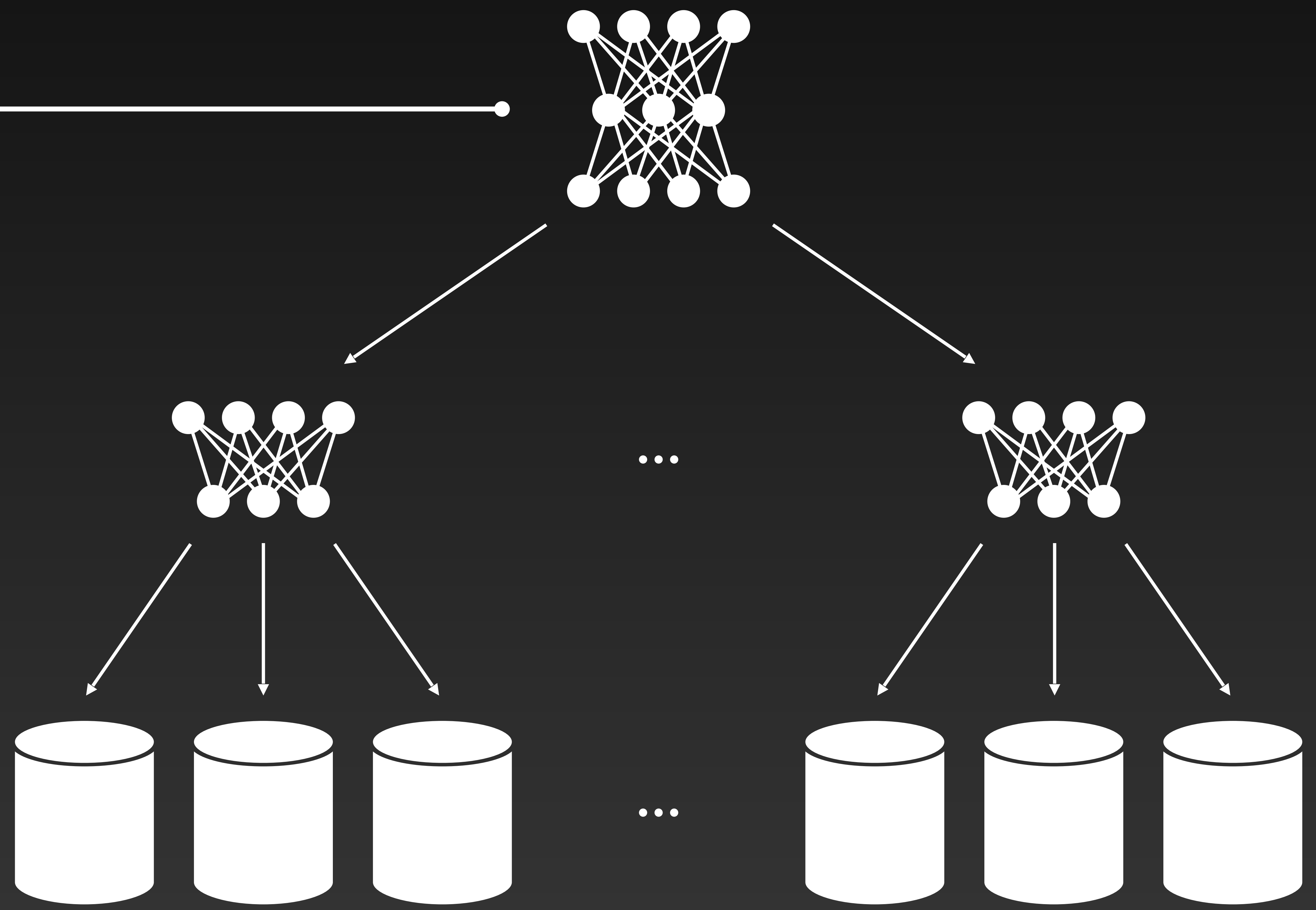
Dimensionality of Embeddings

Data Source	Dimensionality
DINOv2 (image)	384 – 1,536
CLIP (image + text description)	512 – 1,024
Llama 3 (text)	4,096 – 16,384

Learned Indexes for High-Dimensional Data

Learned Metric Index, NeuralLSH, BLISS, BATL, FLEX, ...

A machine learning **model** solving a supervised classification problem: “Where are the most similar objects?”



A **bucket**, basic organizational unit of a collection of vectors.

AlphaFind

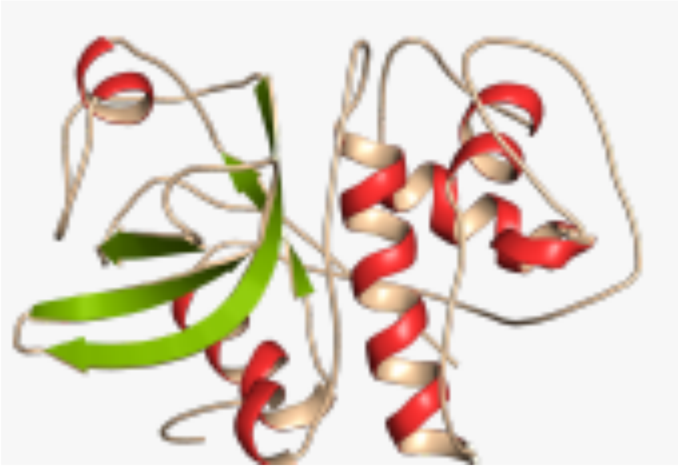
Redefining what it means to efficiently search within 214M proteins

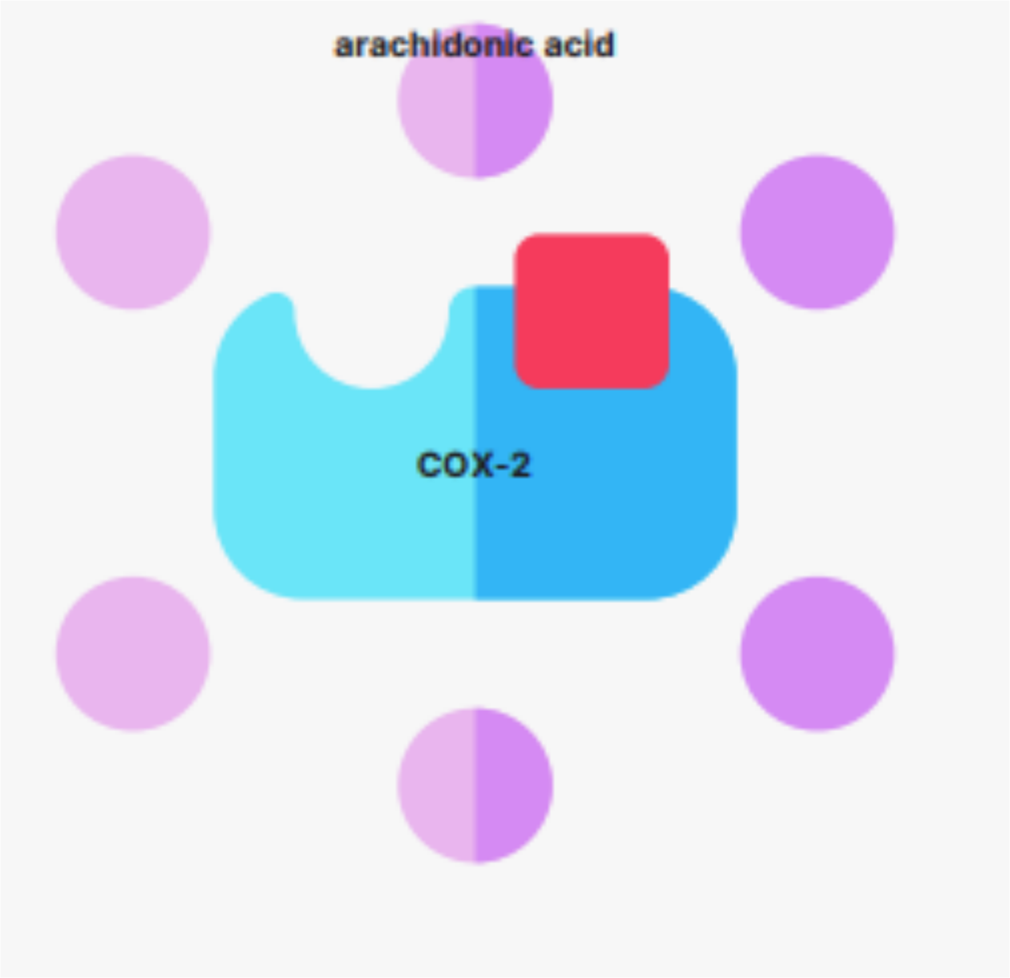
AlphaFind: Similarity search applied to 214-million proteins

T. Slaninakova, 27.9.2024

The why

- Proteins == chains of amino acids folded in 3D space
- The shape determines the function
- Use case for similarity: drug design





The „can-I-do-sim-search-on-it“ checklist

1. There is a ground truth we can rely on
   - We know what similar and not similar look like
   - Ideally, we can quantify it
2. We can represent the data as vectors

The „can-I-do-sim-search-on-it“ checklist

1. There is a ground truth we can rely on
   - We know what similar and not similar look like
   - Ideally, we can quantify it
   - Necessary
2. We can represent the data as vectors
   - Optional, but saves us a lot of headache

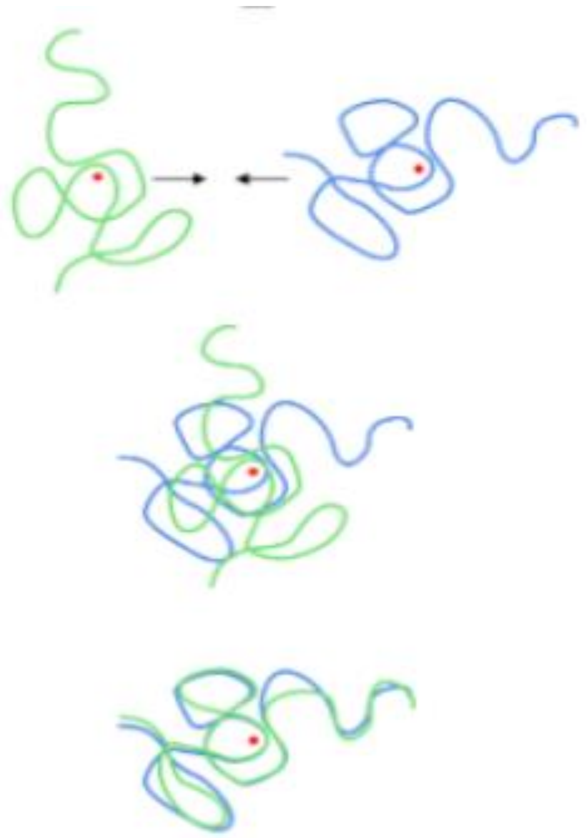
The „can-I-do-sim-search-on-it“ checklist

1. There is a ground truth we can rely on
   - We know what similar and not similar look like
   - Ideally, we can quantify it
   - Necessary
2. We can represent the data as vectors
   - Optional, but saves us a lot of headache

Proteins:

☑ % of protein alignment

☑ Graph neural networks, polynomials, …



Ok, now what?

- Task: Given an input protein, find *k* most similar proteins in 214M AlphaFold database

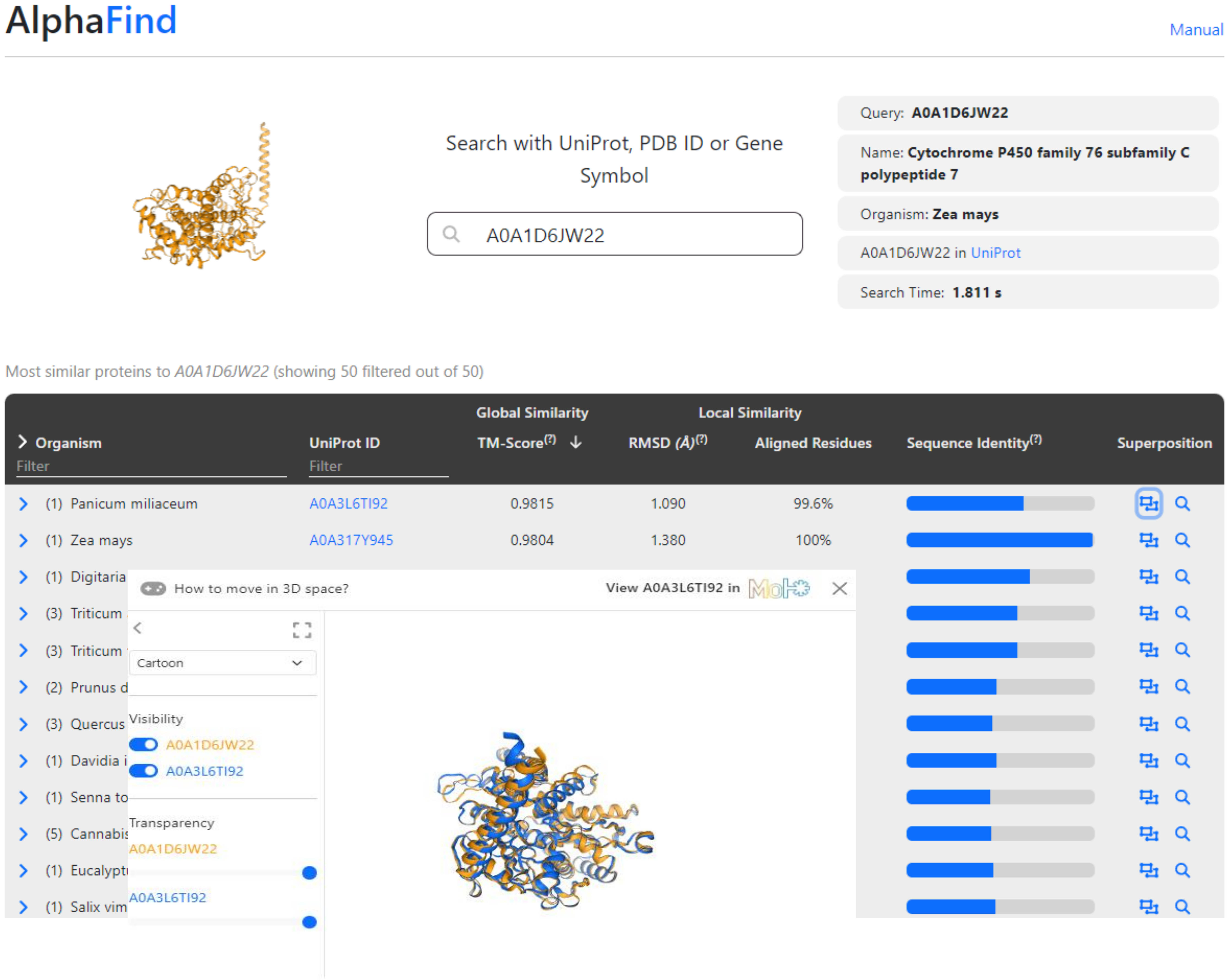
- Approach:
  1. Offline phase:
     1. transform the data into vectors
     2. pre-cluster based on mutual distance
     3. create an index to help with navigation to the clusters

  ← Optimize as much you want, you have (in theory) all the time in the world

  2. Online phase:
     1. locate protein ID in the database
     2. predict its location with the index
     3. do quick (but not as accurate) pre-filtering
     4. do slow (but accurate) post-filtering
     - The challenge: carefully strike the balance between accuracy and speed

  ← You better be fast here

Result



AlphaFind

Manual

Search with UniProt, PDB ID or Gene Symbol

A0A1D6JW22

Query: **A0A1D6JW22**

Name: **Cytochrome P450 family 76 subfamily C polypeptide 7**

Organism: **Zea mays**

A0A1D6JW22 in UniProt

Search Time: **1.811 s**

Most similar proteins to *A0A1D6JW22* (showing 50 filtered out of 50)

		Global Similarity	Local Similarity			
Organism Filter	UniProt ID Filter	TM-Score ↓	RMSD (Å)	Aligned Residues	Sequence Identity	Superposition
(1) Panicum miliaceum	A0A3L6TI92	0.9815	1.090	99.6%		
(1) Zea mays	A0A317Y945	0.9804	1.380	100%		
(1) Digitaria						
(3) Triticum						
(3) Triticum						
(2) Prunus d						
(3) Quercus						
(1) Davidia i						
(1) Senna to						
(5) Cannabis						
(1) Eucalypt						
(1) Salix vim						

- backend of an app AlphaFind running on alphafind.fi.muni.cz

- associated publication in Nucleic Acid Research journal

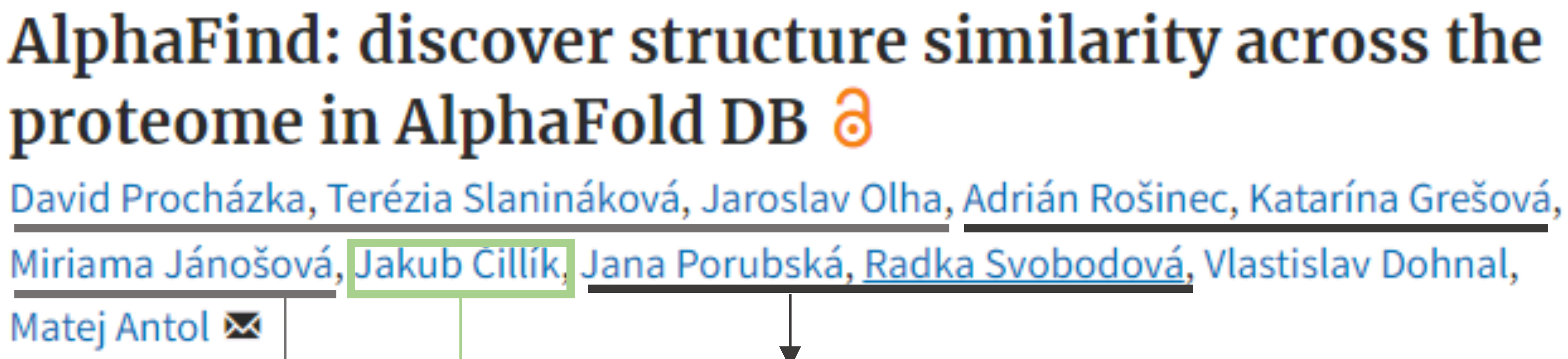
bioRxiv @biorxivpreprint · Feb 18

AlphaFind: Discover structure similarity across the entire known proteome
biorxiv.org/cgi/content/sh... #bioRxiv

biorxiv.org
AlphaFind: Discover structure similarity across the er
AlphaFind is a web-based search engine that provides fast structure-based retrieval in the entire...

Summary

1. Research does not have to be theoretical
2. Research is not reserved for professors:



AlphaFind: discover structure similarity across the proteome in AlphaFold DB

David Procházka, Terézia Slanináková, Jaroslav Olha, Adrián Rošinec, Katarína Grešová, Miriama Jánošová, Jakub Cillík, Jana Porubská, Radka Svobodová, Vlastislav Dohnal, Matej Antol ✉

Your presenters today

Bio friends from CEITEC

Bc. student who created the entire front-end as his bachelor's thesis

Next plans:

- search in protein complexes / RNA / DNA
- Search in ESM Atlas (700M proteins)
- discussing the integration into big protein databases

We're looking for collaborators :)

Also, check us out *today* in Sitola (A502) at Researcher's night (Noc vědců)

PhD

What is it?

Why should I care?

How much does it cost?

...

What about the 💸💸💸?

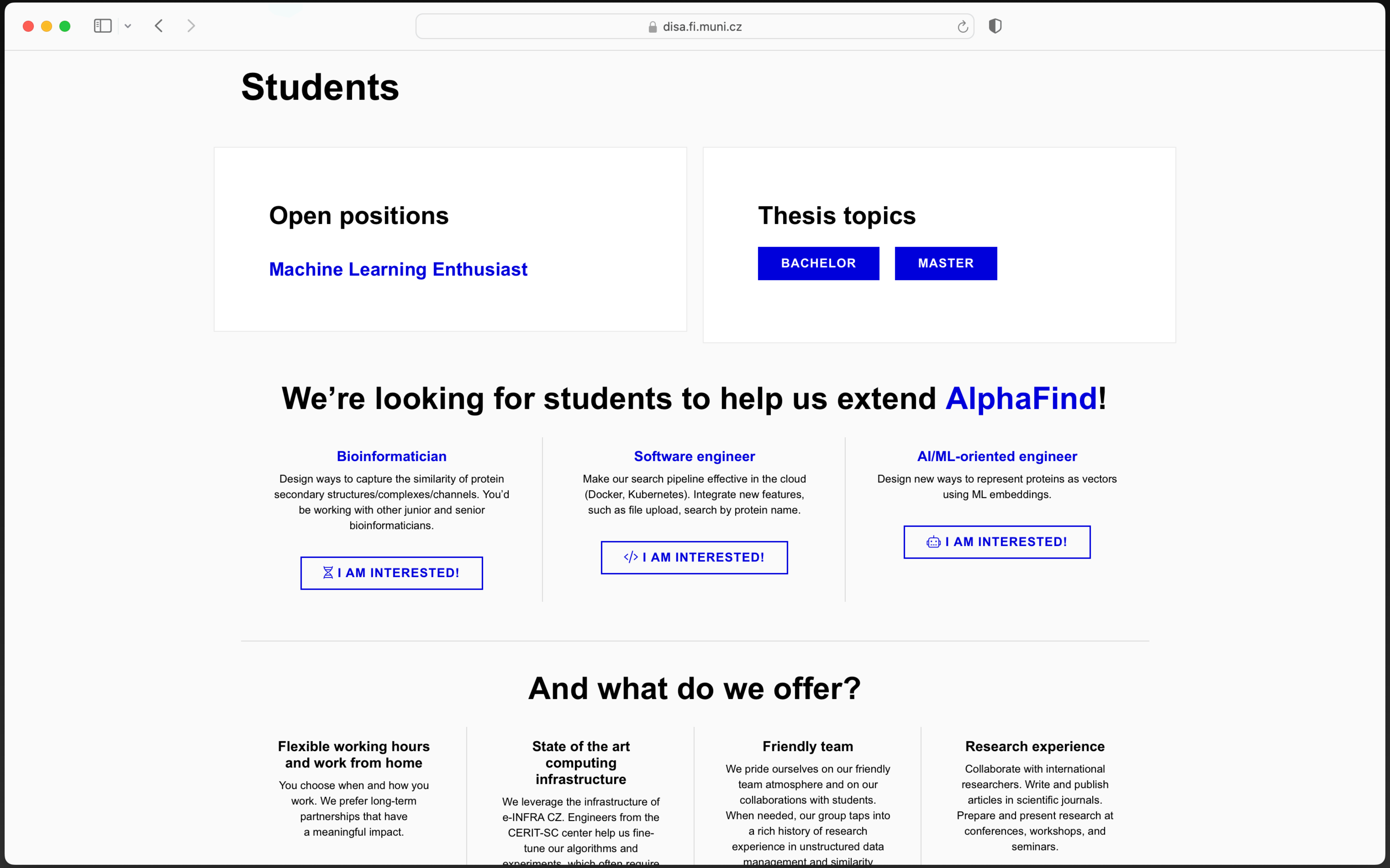
💸💸💸

16k + up to 14k + extra = ~50k net income

Involvement in projects, teaching, ...

What do we offer?

muni.cz/go/students



Students

Open positions

Machine Learning Enthusiast

Thesis topics

BACHELOR MASTER

We're looking for students to help us extend AlphaFind!

Bioinformatician

Design ways to capture the similarity of protein secondary structures/complexes/channels. You'd be working with other junior and senior bioinformaticians.

I AM INTERESTED!

Software engineer

Make our search pipeline effective in the cloud (Docker, Kubernetes). Integrate new features, such as file upload, search by protein name.

I AM INTERESTED!

AI/ML-oriented engineer

Design new ways to represent proteins as vectors using ML embeddings.

I AM INTERESTED!

And what do we offer?

Flexible working hours and work from home

You choose when and how you work. We prefer long-term partnerships that have a meaningful impact.

State of the art computing infrastructure

We leverage the infrastructure of e-INFRA CZ. Engineers from the CERIT-SC center help us fine-tune our algorithms and experiments, which often require

Friendly team

We pride ourselves on our friendly team atmosphere and on our collaborations with students. When needed, our group taps into a rich history of research experience in unstructured data management and similarity

Research experience

Collaborate with international researchers. Write and publish articles in scientific journals. Prepare and present research at conferences, workshops, and seminars.

Thesis Topics About Cool Stuff #1

Thesis tag: CODA research group

Machine Learning

- Indexing Complex Data With Transformers
- Continual Learning for Evolving Data
- Designing Model Architecture for Learned Indexing of Complex Data

Human Motion Data

- Quantization of Auto-Encoded Human Motion Features

Algorithms

- Designing Clustering Algorithm for Indexing Complex Data

Thesis Topics About Cool Stuff #2

Thesis tag: CODA research group

Curse of Dimensionality

- Understanding the Curse of Dimensionality: Implications for High-Dimensional Indexing

Indexing

- Nearest Neighbor Ordering Under Dimensionality Reduction Techniques
- Graph Navigation Approaches to ANN Indexing

Bioinformatics

- Search systems for biomolecular complexes (RNA+proteins)
- Extending metadata search with actual data for large molecular dynamics repositories

Open Position

Machine Learning Enthusiast

- **Develop novel machine learning and data mining approaches** to uncover patterns within large datasets.
- Collaborate with other researchers to **design and implement algorithms for fast indexing of complex data** such as human motion, proteins, images, etc.
- Analyze and **interpret results from experiments** using various visualization tools and statistical methods.

Get involved with the CODA Research Group

muni.cz/go/coda

1. Collaborate on **research** topics
   - Learned Metric Index, AlphaFind, …
2. Do your **PhD** in our group
3. Sign up for one of our **thesis** topics

Still not sure?

Come **ask us in person or contact us** (dohnal@fi.muni.cz) and we can figure it out together!