TECHNISCHE UNIVERSITÄT DRESDEN

DRESDEN concept

An introduction to Large Language Models

Anna Poetsch

Research Group „Biomedical Genomics“, Biotechnology Center TU Dresden, NCT Dresden, and CSBD

@apoetsch

Organisation

25.6.2024: Dimensionality reduction

2.7.2024: Machine Learning

9.7.2024: Machine Learning exercise
Deep Learning, Large Language Models

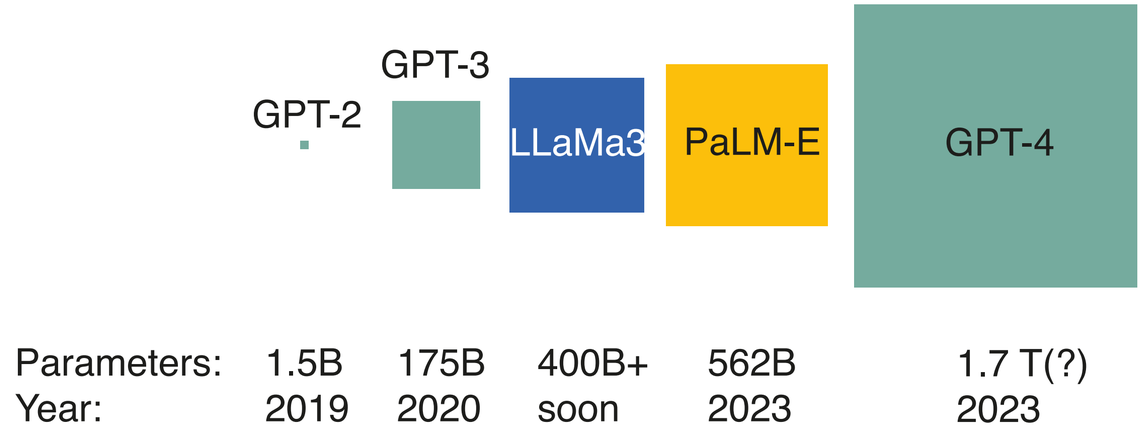
16.7.2024: Deep Learning exercise
Summary, recap, Q&A

Outline

- What are Large Language Models?
- What do they learn?
- Token embeddings
- A short glimpse into a transformer foundation model
- Hallucination and bias
- Large Language Models with language-like data

Pre-training and fine-tuning

Foundation models train language on large corpora of data



BIG

expensive

rigid

Fine-tuning

- assistant
- image generation
- translation
- speech recognition
- ...

“DallE2, please give me an illustration of damaged DNA in comic style”:

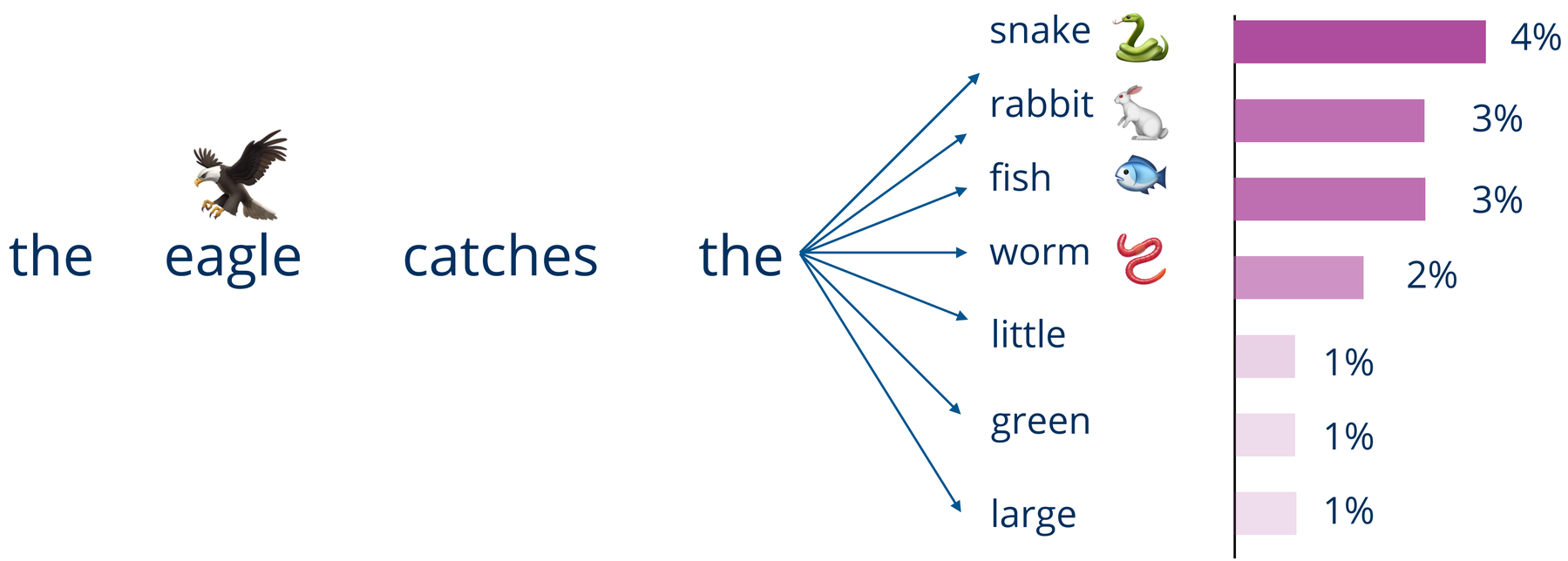
efficient

flexible

cheap

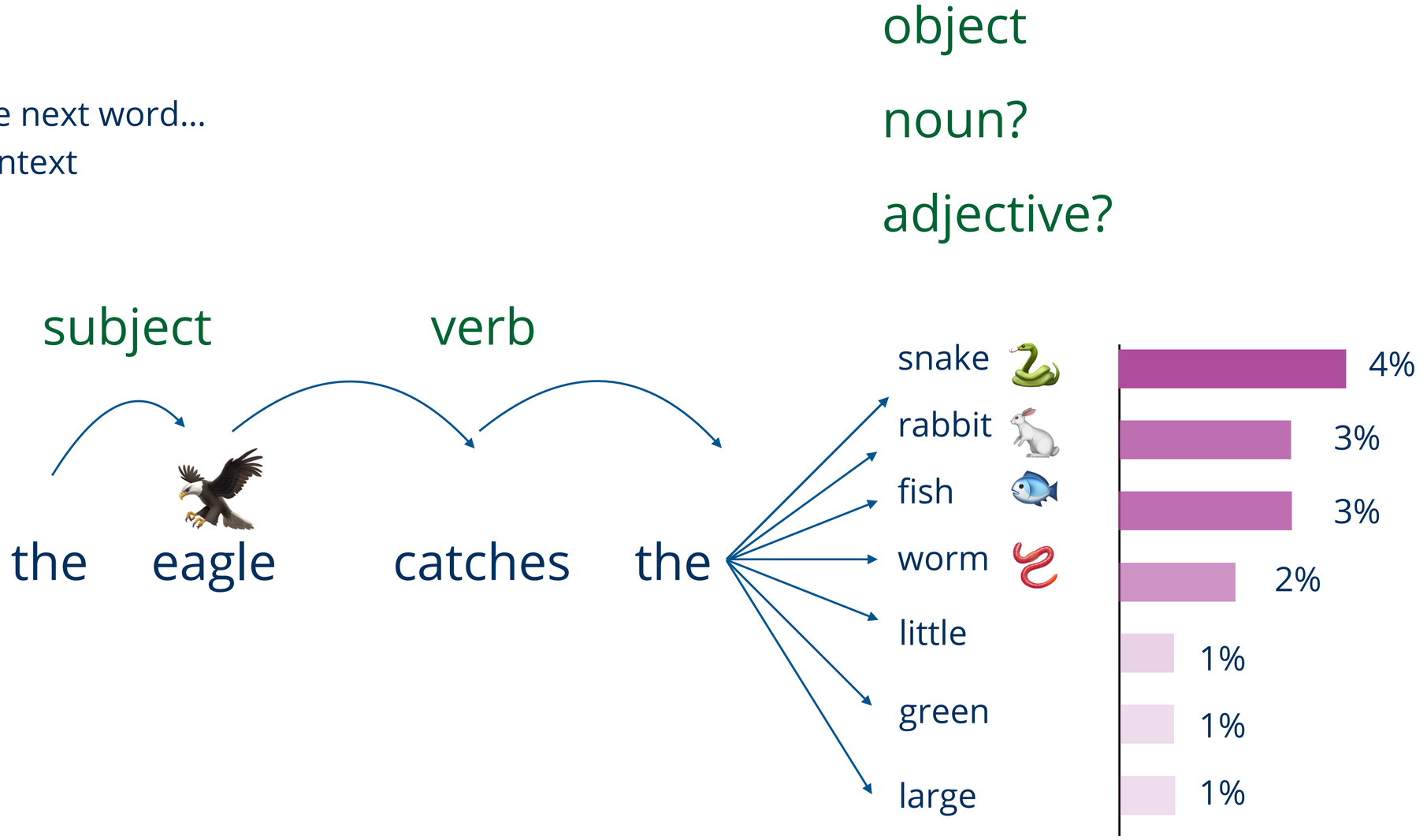
Large Language Models

Predicting the next word...



Large Language Models

Predicting the next word...
...requires context



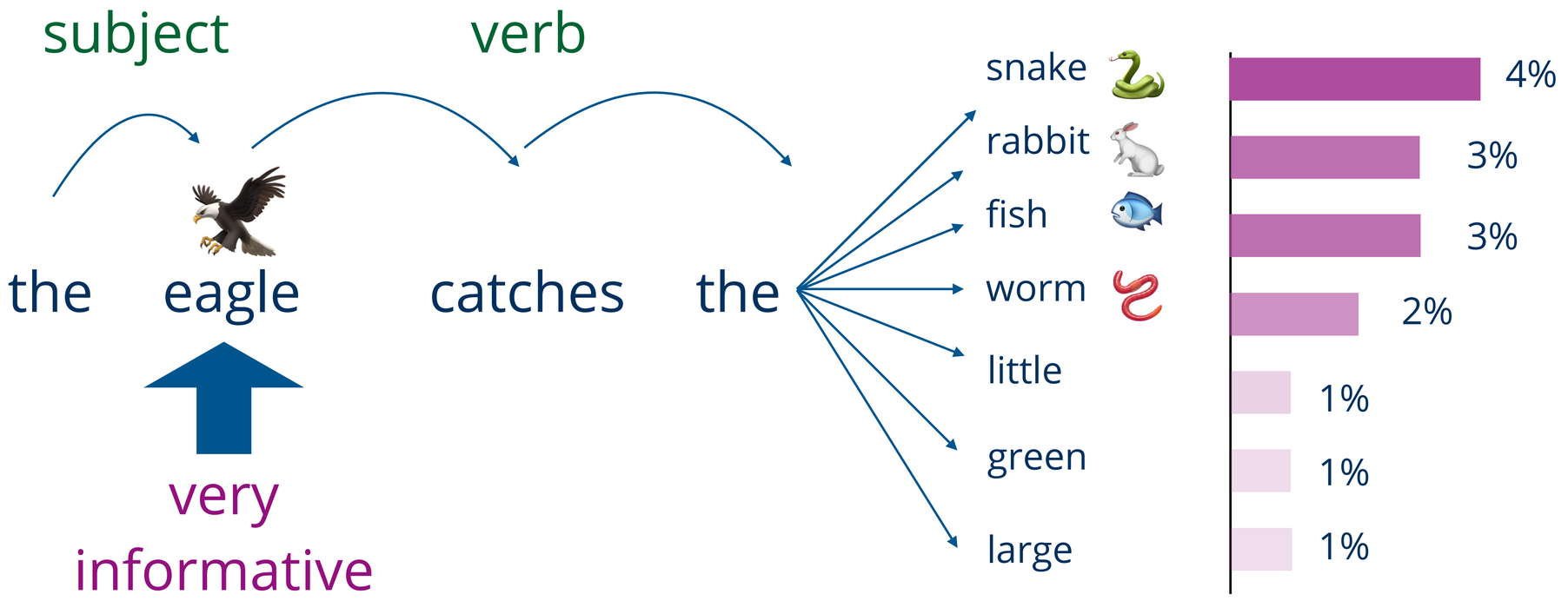
Large Language Models

Predicting the next word...
...requires context

object

noun?

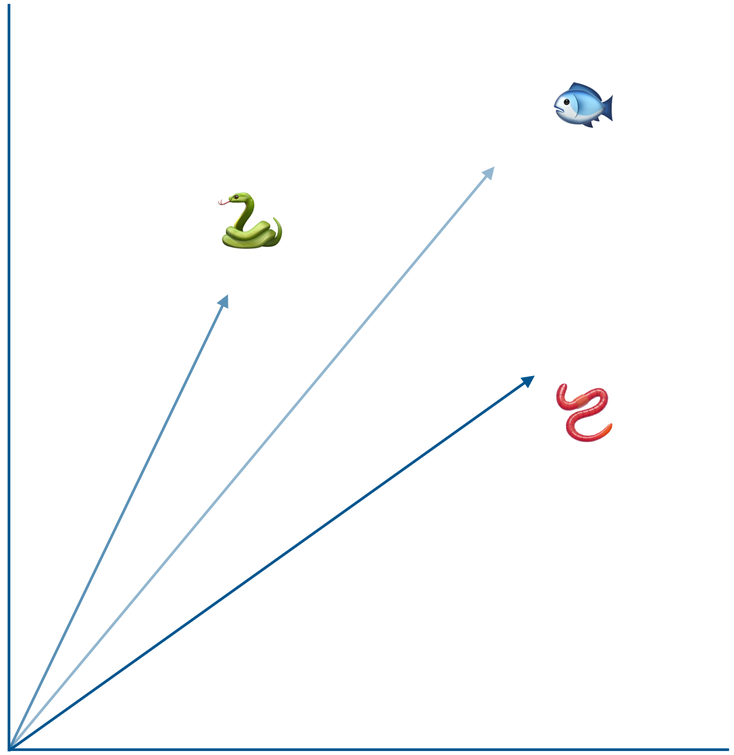
adjective?



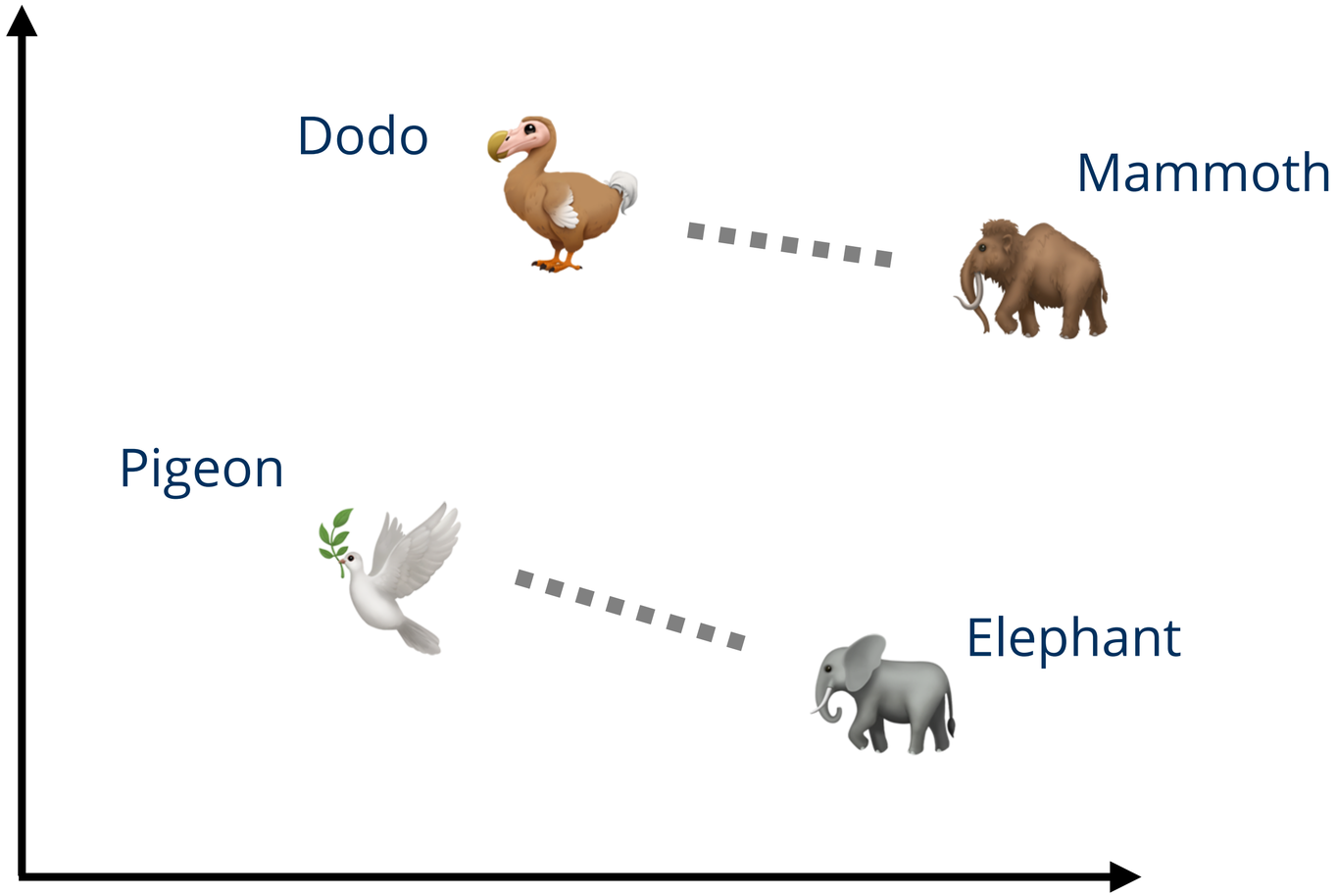
The embedding of words/ tokens

Tokens are assigned a vector, which places them into a multi-dimensional space

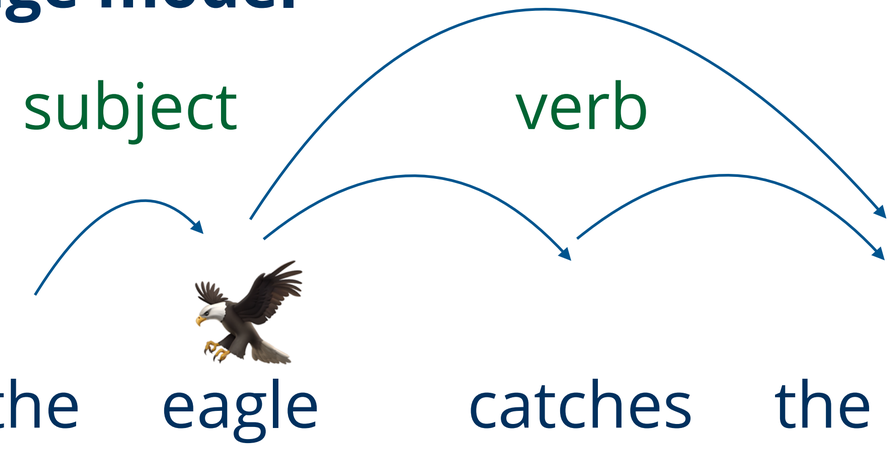
the	eagle	catches	the	???
5.4	3.2	0.2	5.4	
2.3	1.4	0.3	2.3	
1.2	0.4	5.6	1.2	
.	.	.	.	
.	.	.	.	
.	.	.	.	
6.4	1.2	3.2	6.4	
0.4	8.6	4.1	0.4	

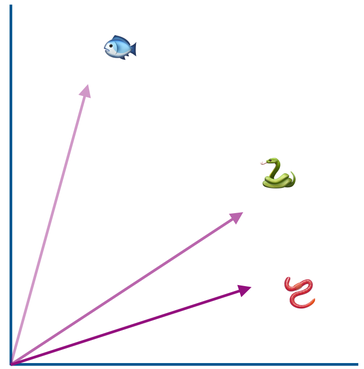


Recap: PCA in the embedding of language



Training a large language model





Input embedding

5.4	3.2	0.2	5.4
2.3	1.4	0.3	2.3
1.2	0.4	5.6	1.2
.	.	.	.
.	.	.	.
.	.	.	.
6.4	1.2	3.2	6.4
0.4	8.6	4.1	0.4

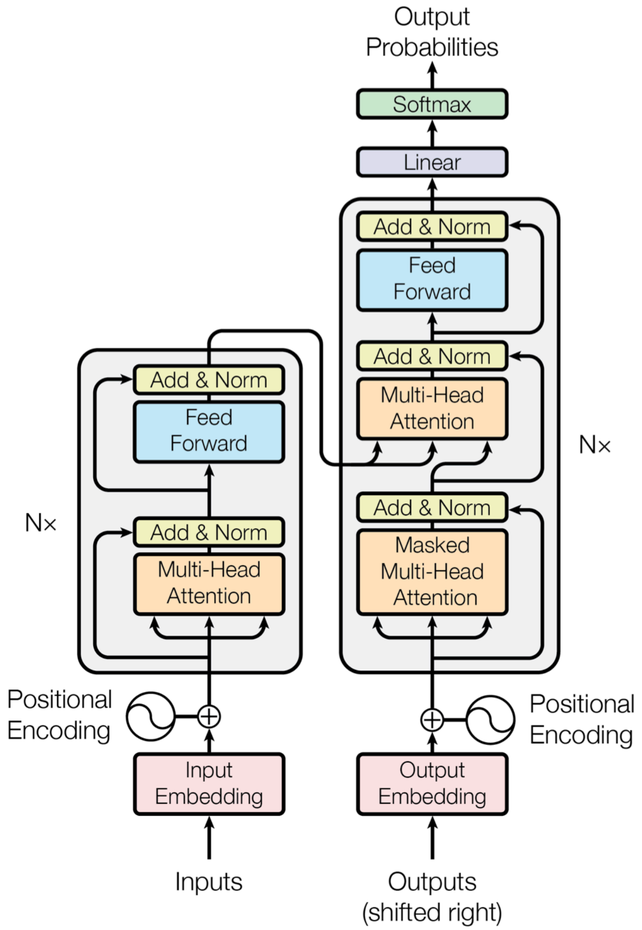
context

learned through attention blocks

Trained embedding

3.2	6.3	0.6	3.7
4.6	0.4	1.7	5.3
5.2	8.2	4.8	7.2
.	.	.	
.	.	.	
.	.	.	
5.9	3.2	3.9	2.1
0.3	0.4	6.2	0.9

“Attention is all you need”



$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

https://arxiv.org/abs/1706.03762

https://medium.com/@tech-gumptions/transformer-architecture-simplified-3fb501d461c8#:~:text=The transformer architecture consists of,translation or a text continuation.

“Attention is all you need”

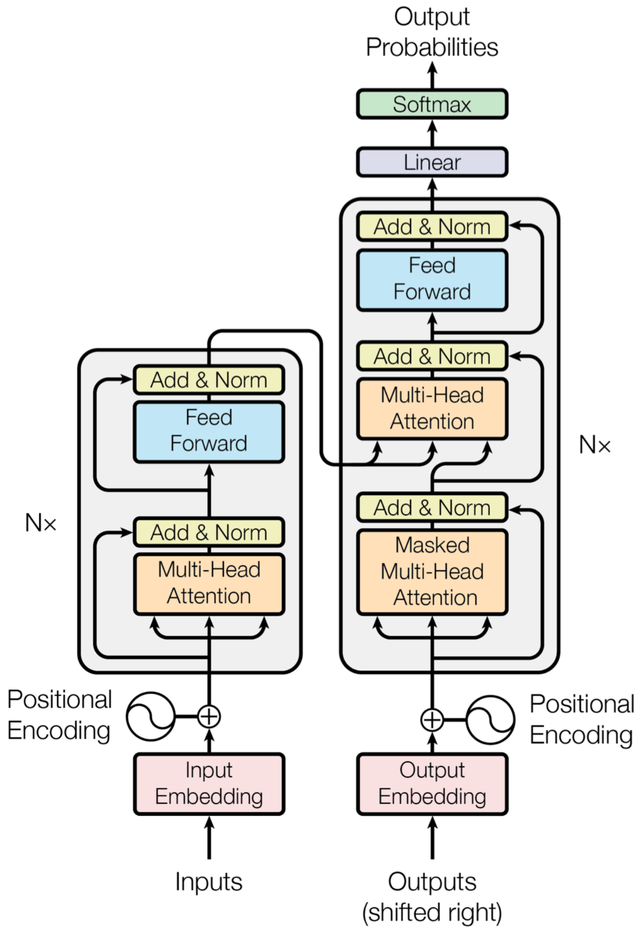
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Query (Q): Represents the word we are currently focusing on. Each word has its corresponding Query vector.

Key (K): Represents all words we want to pull information from to help determine the relevance of each word to the Query.

Value (V): Contains the information of words that we’ll extract when there’s a match between Query and Key.

“Attention is all you need”



GPT-3							Total parameters: 175,181,291,520 27,938 matrices
Embedding	12,288 d embed	*	50,257 n vocab				= 617,558,016
attention – Key	128 d query	*	12,288 d embed	* 96 n heads	* 96 n layers		= 14,495,514,624
attention – Query	128 d query	*	12,288 d embed	* 96 n heads	* 96 n layers		= 14,495,514,624
attention – Value	128 d query	*	12,288 d embed	* 96 n heads	* 96 n layers		= 14,495,514,624
Output	12,288 d embed	*	128 d value	* 96 n heads	* 96 n layers		= 14,495,514,624
Up-projection	49,152 n neurons	*	12,288 d embed	* 96 n layers			= 57,982,058,496
Down-projection	12,288 d embed	*	49,152 n neurons	* 96 n layers			= 57,982,058,496
Unembedding	50,257 n vocab	*	12,288 d embed				= 617,558,016

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

https://arxiv.org/abs/1706.03762

Hallucination, “making things up”

Anna Poetsch is a Computational Biologist of the TU Dresden.
She has received a grant from the European Research Council (ERC) and is one of the winners of the “Cluster of Excellence: Inflammation at Interfaces”. The Cluster of Excellence led by the TU Dresden and the University of Leipzig has been funded with more than 40 million euros by the German Federal Ministry of Education and Research (BMBF) for a period of ten years.

Out of context conversations

ChatGPT

I apologize if it seems like I am asking the same questions over and over again. I am a large language model trained by OpenAI and do not have personal experiences or memories. As a result, I am unable to remember past conversations or specific questions that I have asked. I generate responses based on the prompts given to me and do not have the ability to browse the web or access other information.

Bias

A model is only as good as the data used to train it.

Trained it with racism and sexism, it will return just that.

Compensating for bias is hard.

Google Gemini: Adi Robertson / The Verge

Other text-like data

Proteins

DNA

Sound

Music

.

.

.

Summary, take-home messages, and further information

- Large Language models are models trained self-supervised on next- or masked-token prediction
- They learn a sense of grammar and syntax, as well as language context
- They can be fine-tuned for a myriad of tasks (e.g. assistants and image generators)
- They are at risk of hallucination and amplifying bias
- Any data that resembles “language” can be used for training a model like this

Attention is all you need: https://arxiv.org/abs/1706.03762

How Transformers Work: A Detailed Exploration of Transformer Architecture: https://www.datacamp.com/tutorial/how-transformers-work

But what is a GPT? Visual intro to transformers: https://www.youtube.com/watch?v=wjZofJX0v4M

Attention in transformers, visually explained: https://www.youtube.com/watch?v=eMlx5fFNoYc

TECHNISCHE UNIVERSITÄT DRESDEN

DRESDEN concept

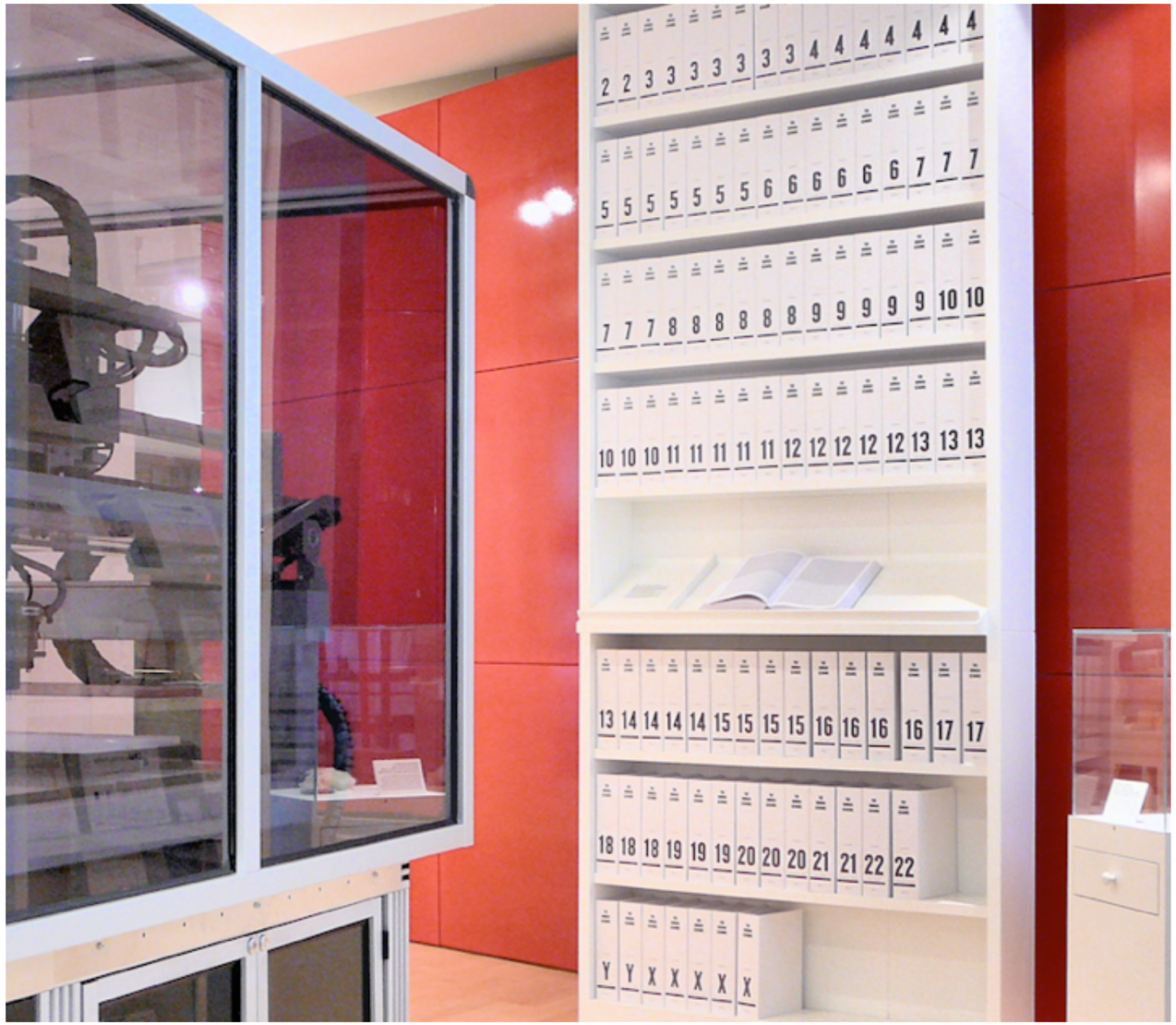
Using DNA language models to understand different layers of code in genomes

Anna Poetsch

Research Group „Biomedical Genomics“, Biotechnology Center TU Dresden, NCT Dresden, and CSBD

@apoetsch

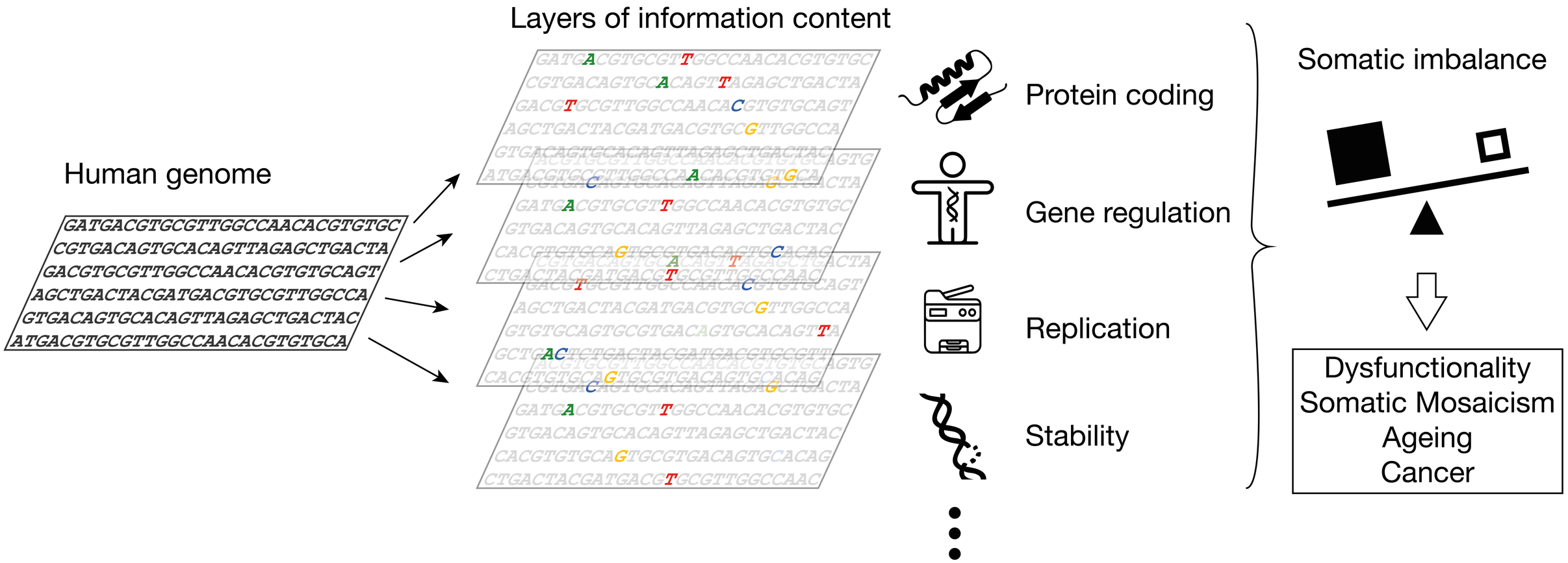
The Human Genome



3.2 Billion letters
119 books
20.000 genes (1-2 %)

Wellcome Trust Collection in London
Image source: Russ London at English Wikipedia, licensed CC-BY-SA 3.0

Information content of DNA



Can we treat DNA as if it were text?

DNABERT-2

GROVER

HyenaDNA

Nucleotide Transformer

Proformer

Enformer

⋮

DNA-BERT

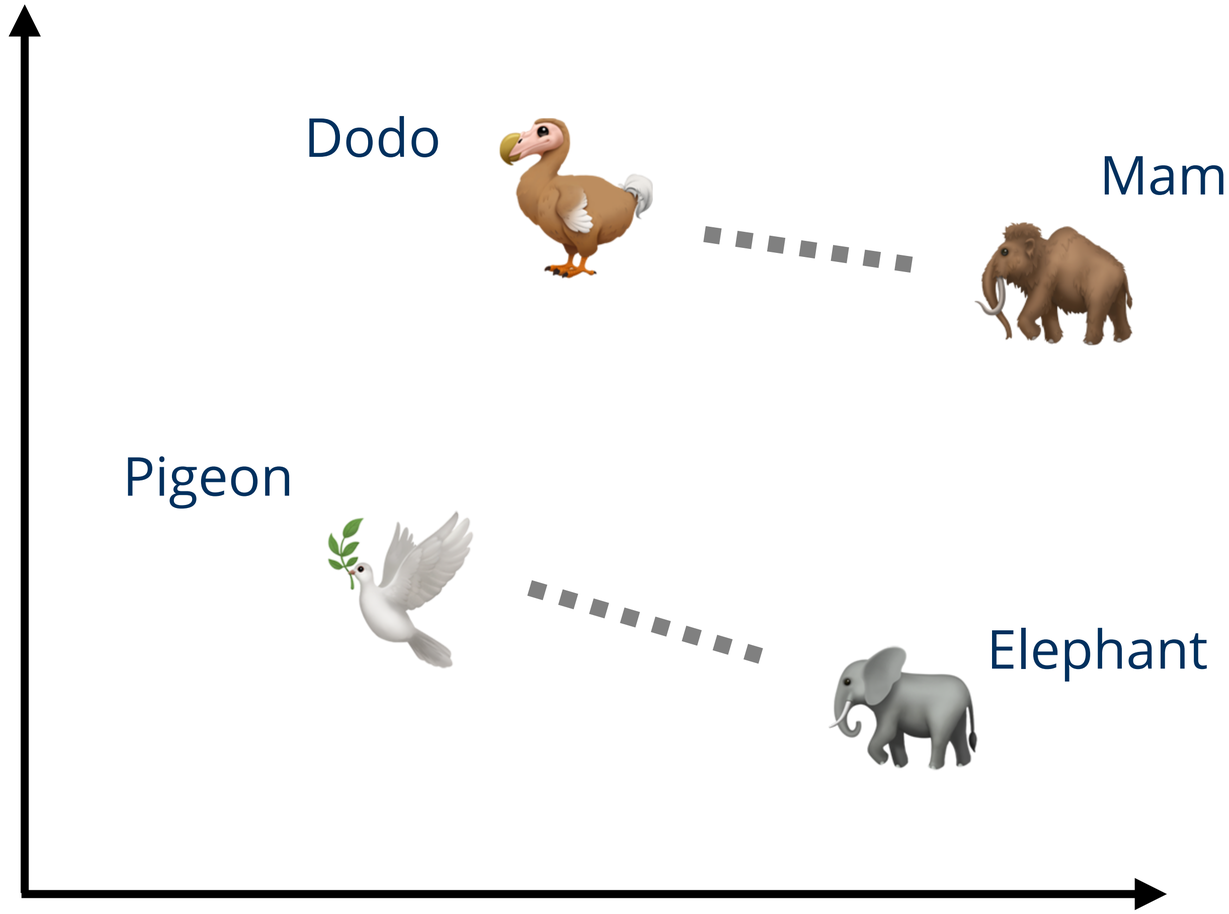
Yi *et al*: https://academic.oup.com/bioinformatics/article/37/15/2112/6128680
Sanabria *et al*: https://www.biorxiv.org/content/10.1101/2023.07.11.548593v1

DNA-BERT

Sanabria *et al*: https://www.biorxiv.org/content/10.1101/2023.07.11.548593v1

Word2Vec embedding

Algorithm for semantic word associations through static embeddings



- shallow neural networks
- input is large corpus of text
- Each unique word is assigned a position in the vector space
- Different methods can be used, e.g. continuous bag-of-words (CBOW) “fill in the blank” task.

Using next-k-mer prediction to compare models

Sanabria *et al*: https://www.biorxiv.org/content/10.1101/2023.07.11.548593v1

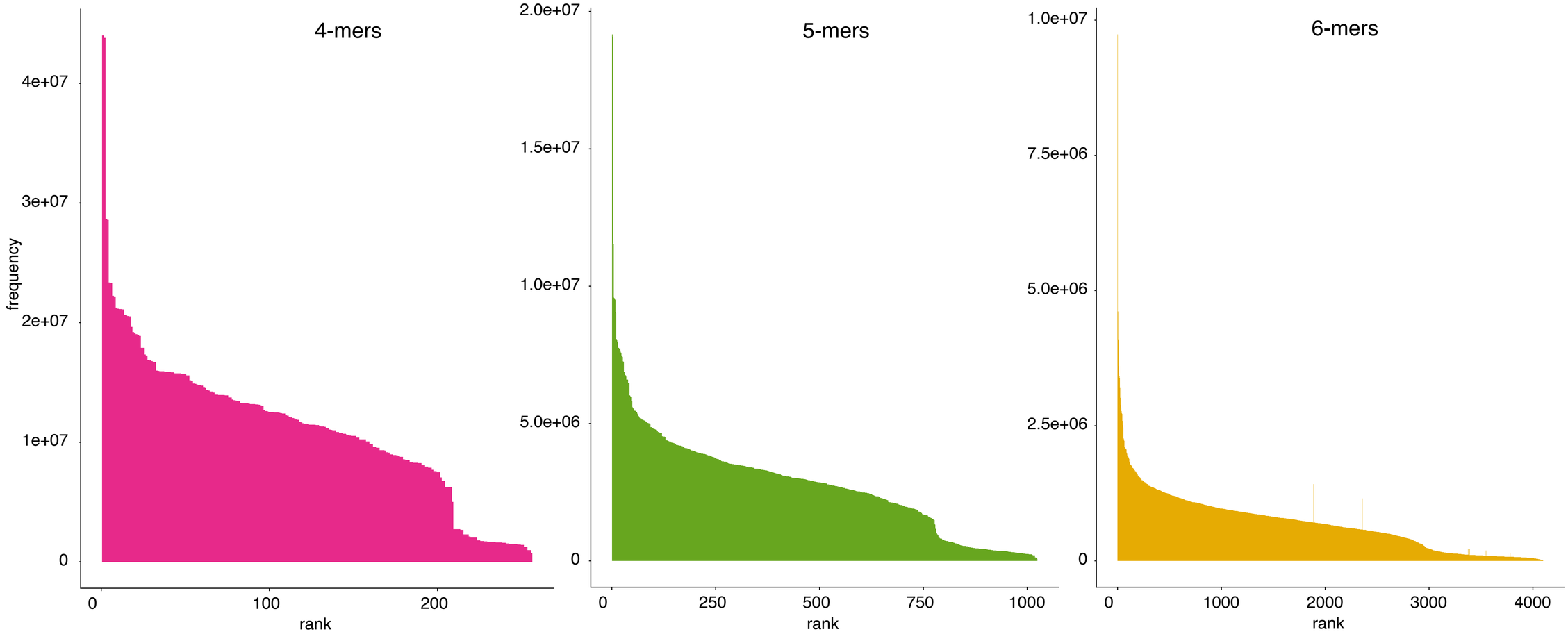
Overlaps lead to learning how to spell
How can we avoid this?

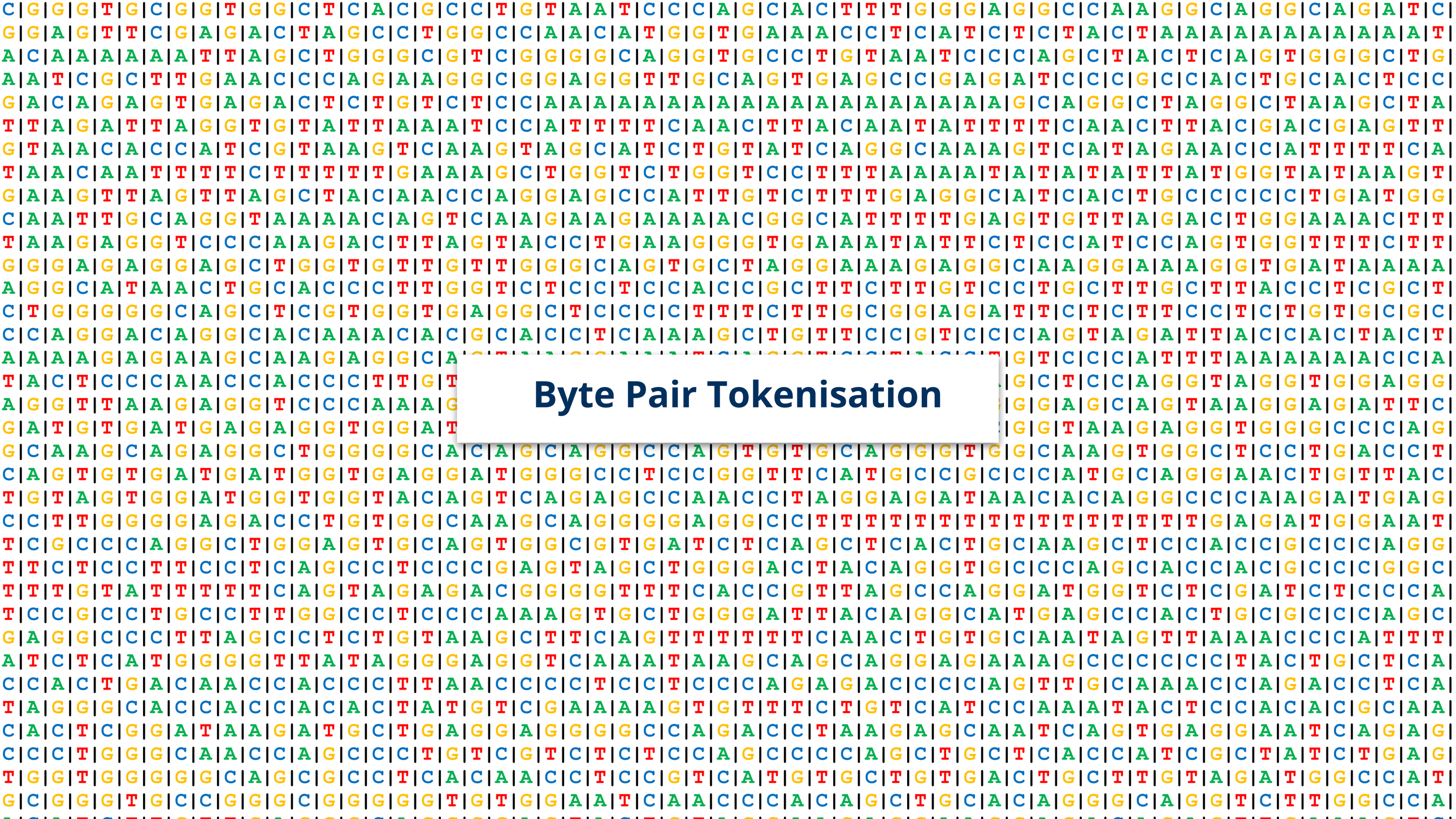
GROVER: Genome Rules Obtained Via Extracted Representations

Sanabria *et al* BioRxiv

What is a word?

TACCATGTACCTGGAGCCATGTTGCTCTCCTCTAGCAAAGATGCAGCAGATGCAGTTGGGGCTGCCACT




Byte Pair Tokenisation


T|T-> TT


A|A-> AA


T|G-> TG


A|G-> AG


C|C-> CC

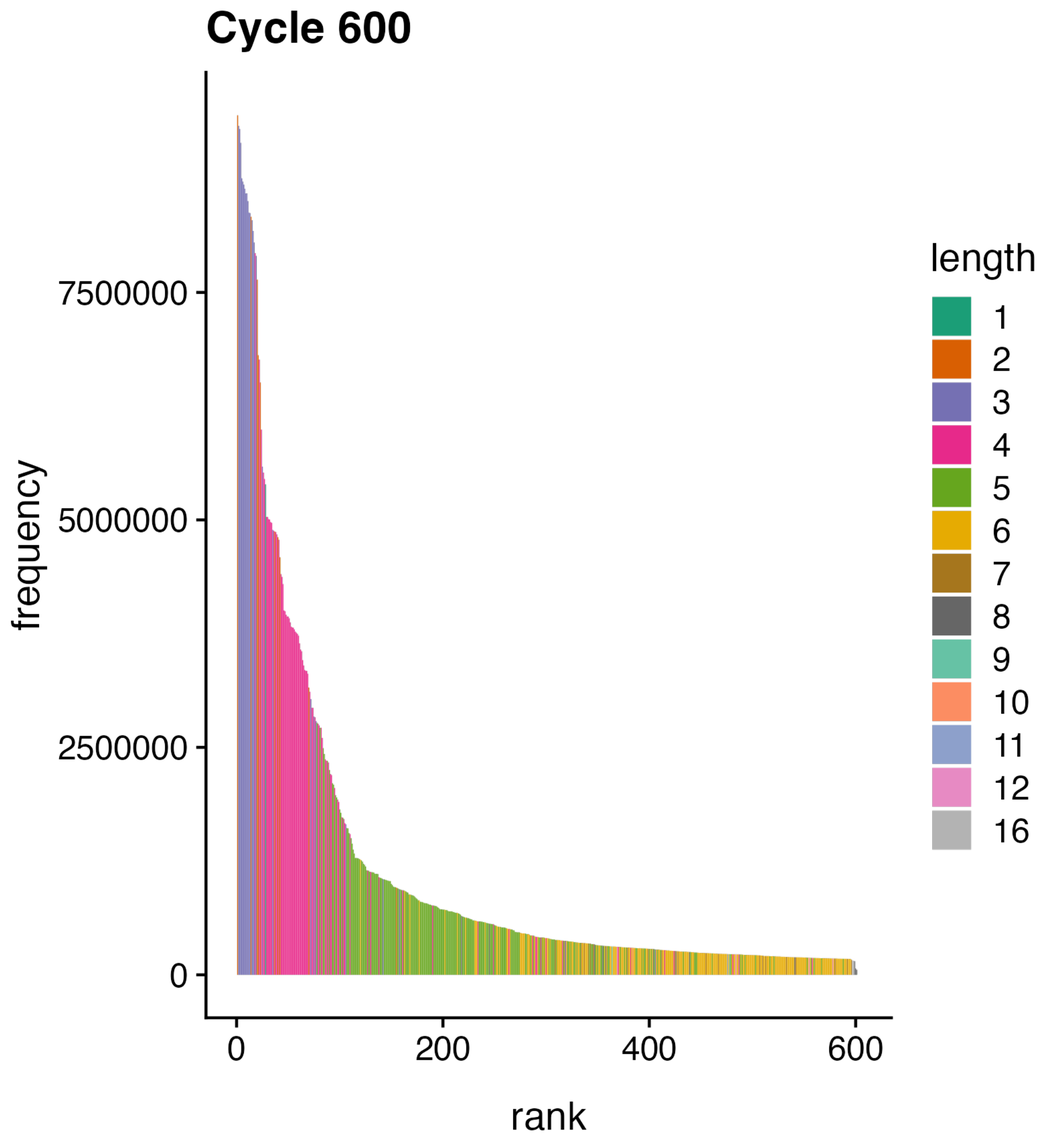

T|C-> TC


A|C-> AC


G|G-> GG


A|TT-> ATT

600 cycles builds the “best” vocabulary



Sanabria *et al*: https://www.biorxiv.org/content/10.1101/2023.07.19.549677v1

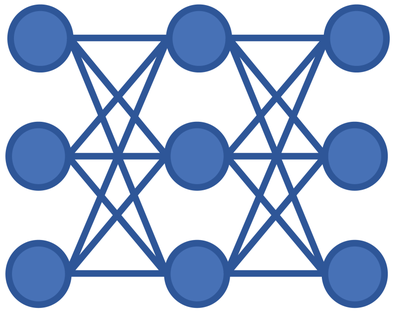
What does the vocabulary look like?

Sanabria *et al*: https://www.biorxiv.org/content/10.1101/2023.07.19.549677v1

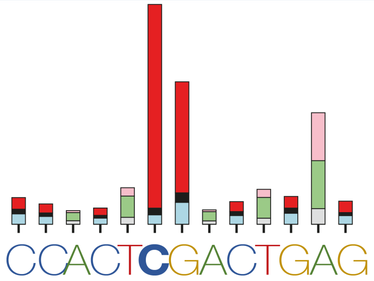
What can we use GROVER for?

Vocabulary of the human genome

Language model



Language rules/ biological meaning



Elements . . . Mutagenesis . . . DNA breakage

https://huggingface.co/PoetschLab/GROVER

What is GROVER learning?

It learns to differentiate token characteristics

Sanabria *et al* BioRxiv

What is GROVER learning?

It learns to see tokens in context

Sanabria *et al* BioRxiv

What is GROVER learning?

It learns some context

Sanabria *et al* BioRxiv

Fine-tuning tasks to understand genome biology

Sanabria *et al* BioRxiv

What does GROVER learn?

- Self-supervised context for text-completion
- token sequence
- token contexts
- biological functionality of sequence

How can we use it to discover biology?

Sanabria *et al* BioRxiv

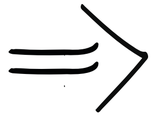
Is Genome Instability is encoded in the DNA sequence?

How genomes age

damage over time

mutation over time

Epigenetic changes



Loss of functionality

+

Survival of the fittest

Mario Aguilar

Fine-tuning tasks to predict the frequency of DNA double strand breaks

Fine-tuning tasks to predict the frequency of DNA double strand breaks

Fine-tuning tasks to predict the frequency of DNA double strand breaks

Summary

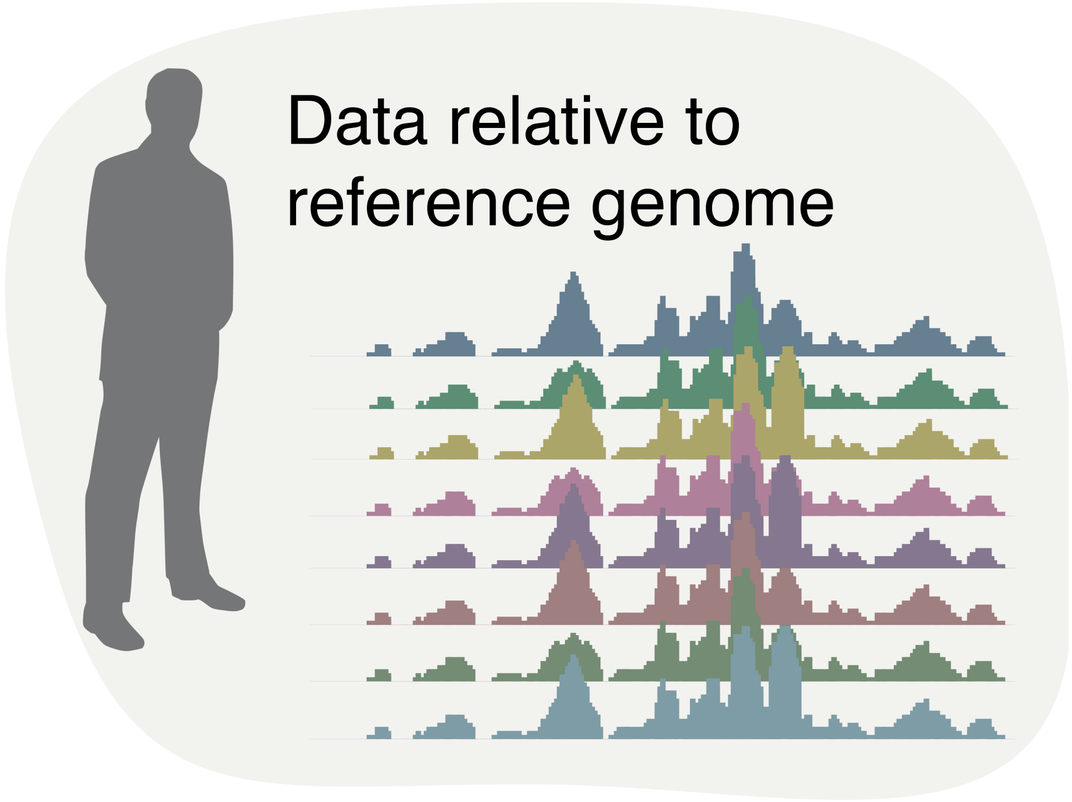
- DNA language models learn sequence content and context

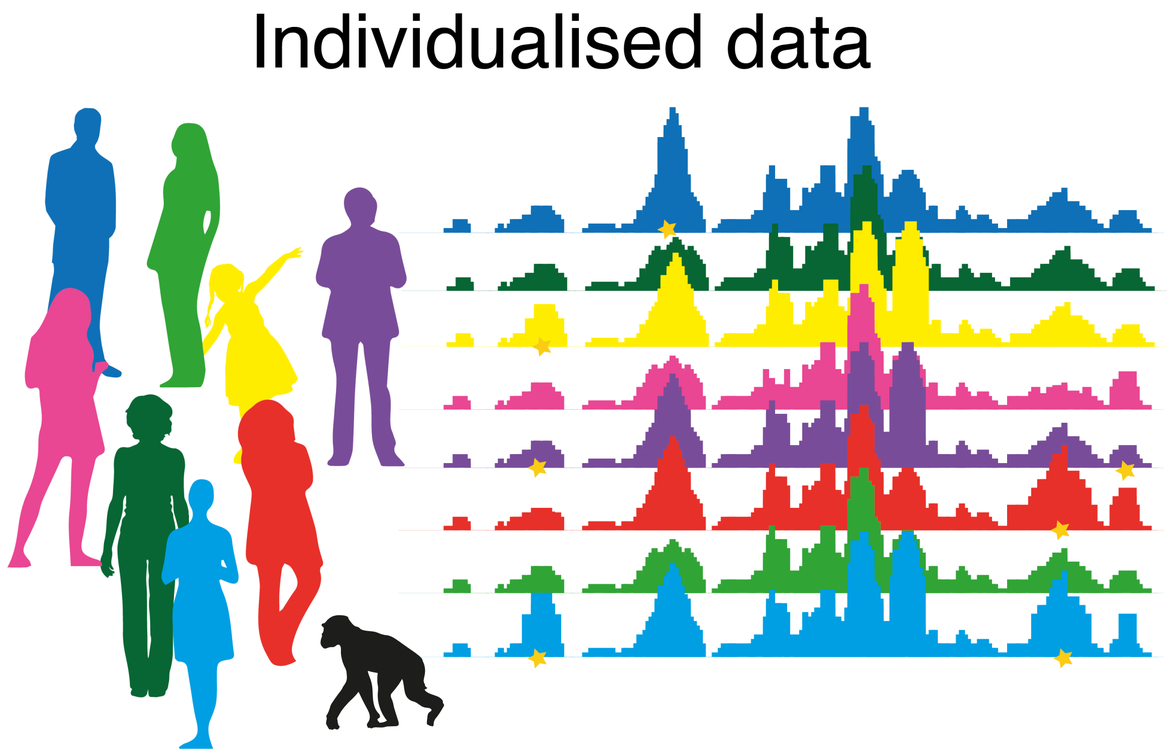
- They can be implemented with two strategies:
  - self-supervised pretraining and fine tuning
  - direct learning of the question at hand

- Combining models can help to differentiate information content between sequence only and epigenetic information

- GROVER can help with a variety of questions in genomics

Translation - individualisation of omics data

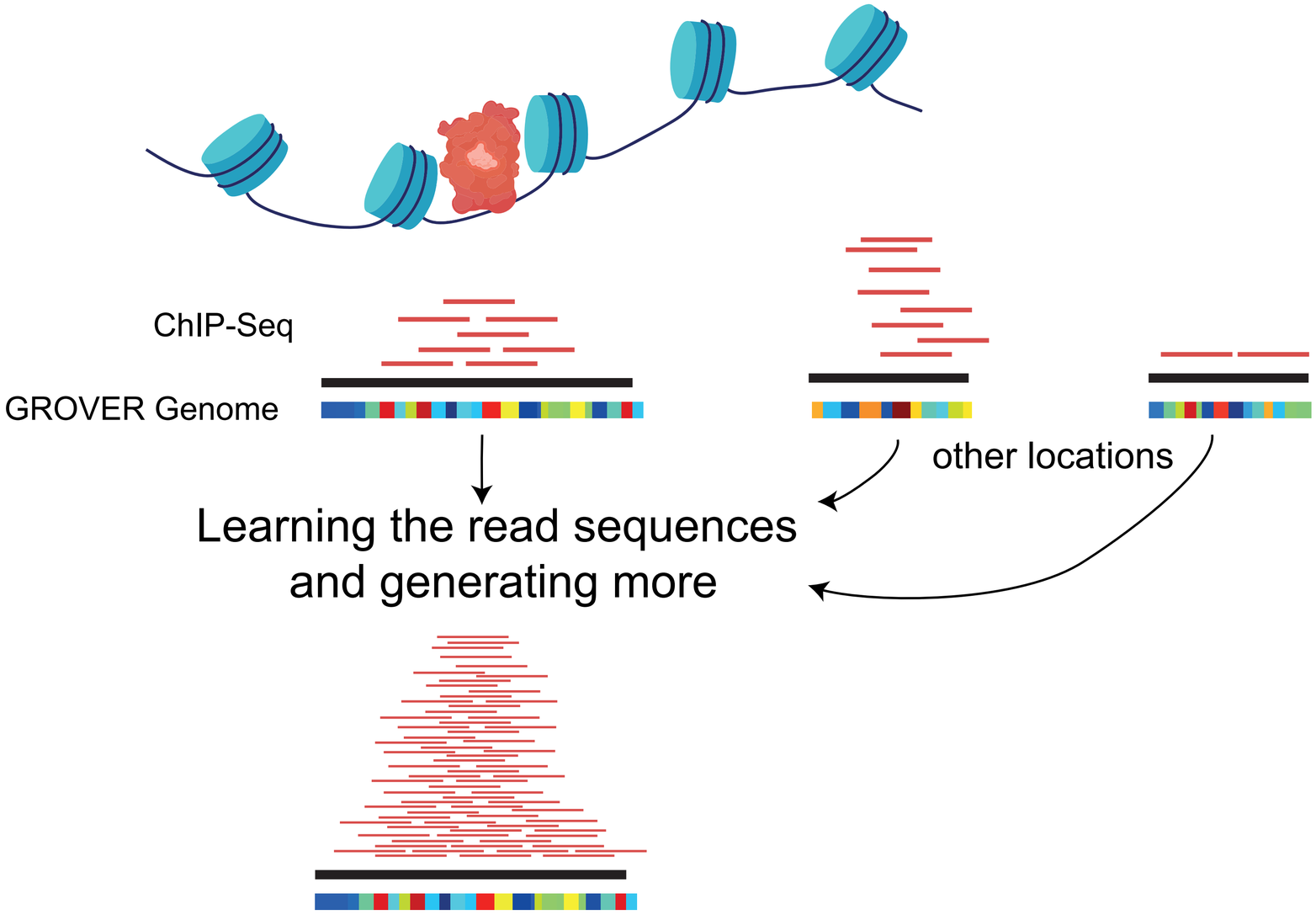




Translation AI - translation of genomics data across species

annotation
regulatory genomics data
⋮

Generative AI - Boosting of genomics data



TextToImage - GenotypeToPhenotype

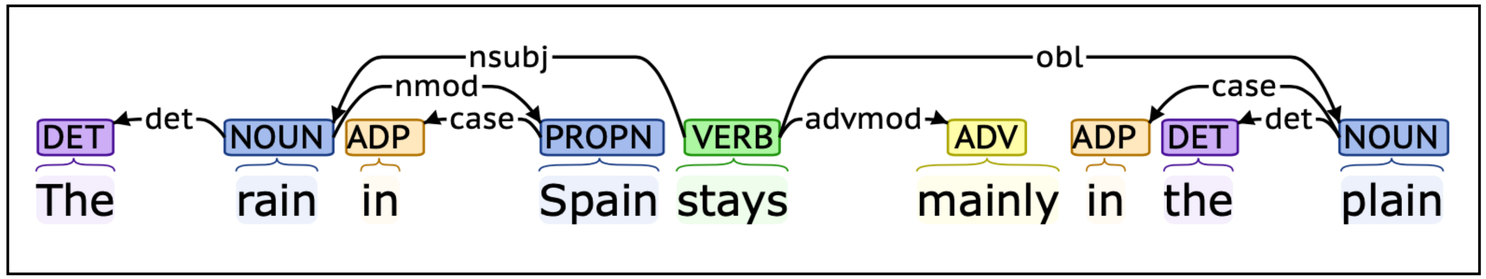
“DallE2, please give me an illustration of damaged DNA in comic style”:

“...ACGCGTAAAATCGATTTAGCGATTTGCAA...”:

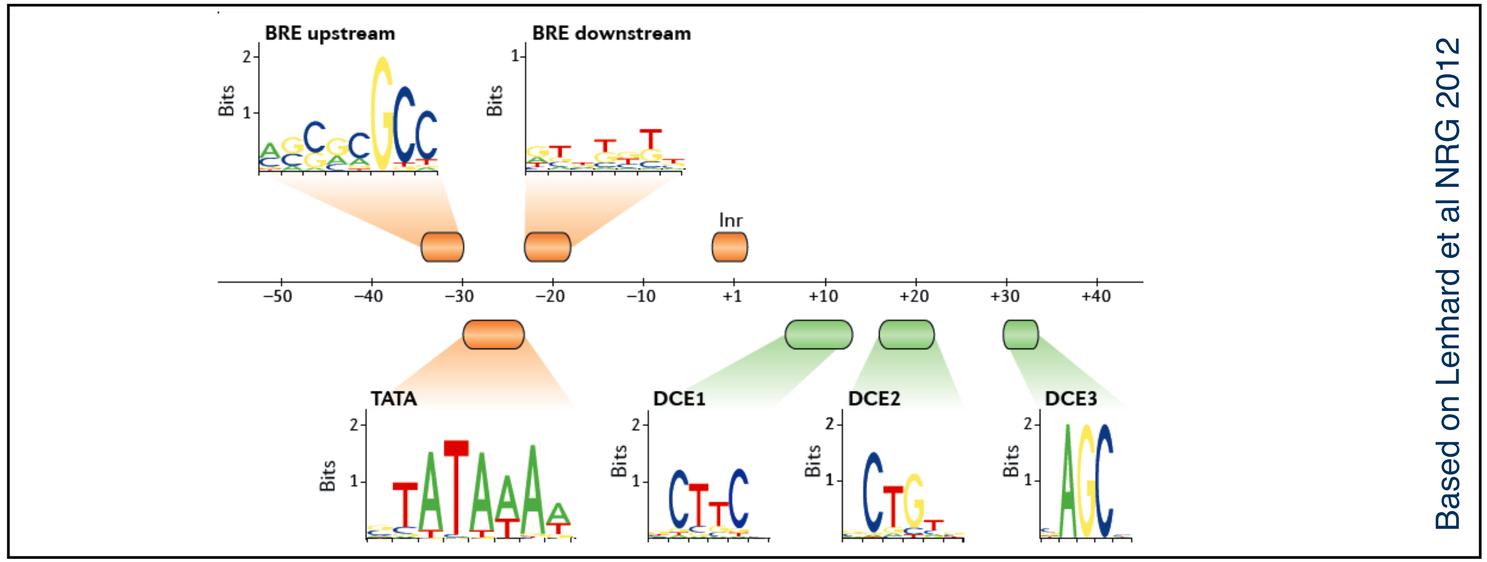
https://www.ebi.ac.uk/training/online/courses/human-genetic-variation-introduction/what-is-genetic-variation/genotype-or-phenotype/

Computational linguistics - Genome linguistics

Stanza v1.3.0



Promoter Grammar



Based on Lenhard et al NRG 2012

GROVER tutorial

- predicting the use of CTCF binding sites