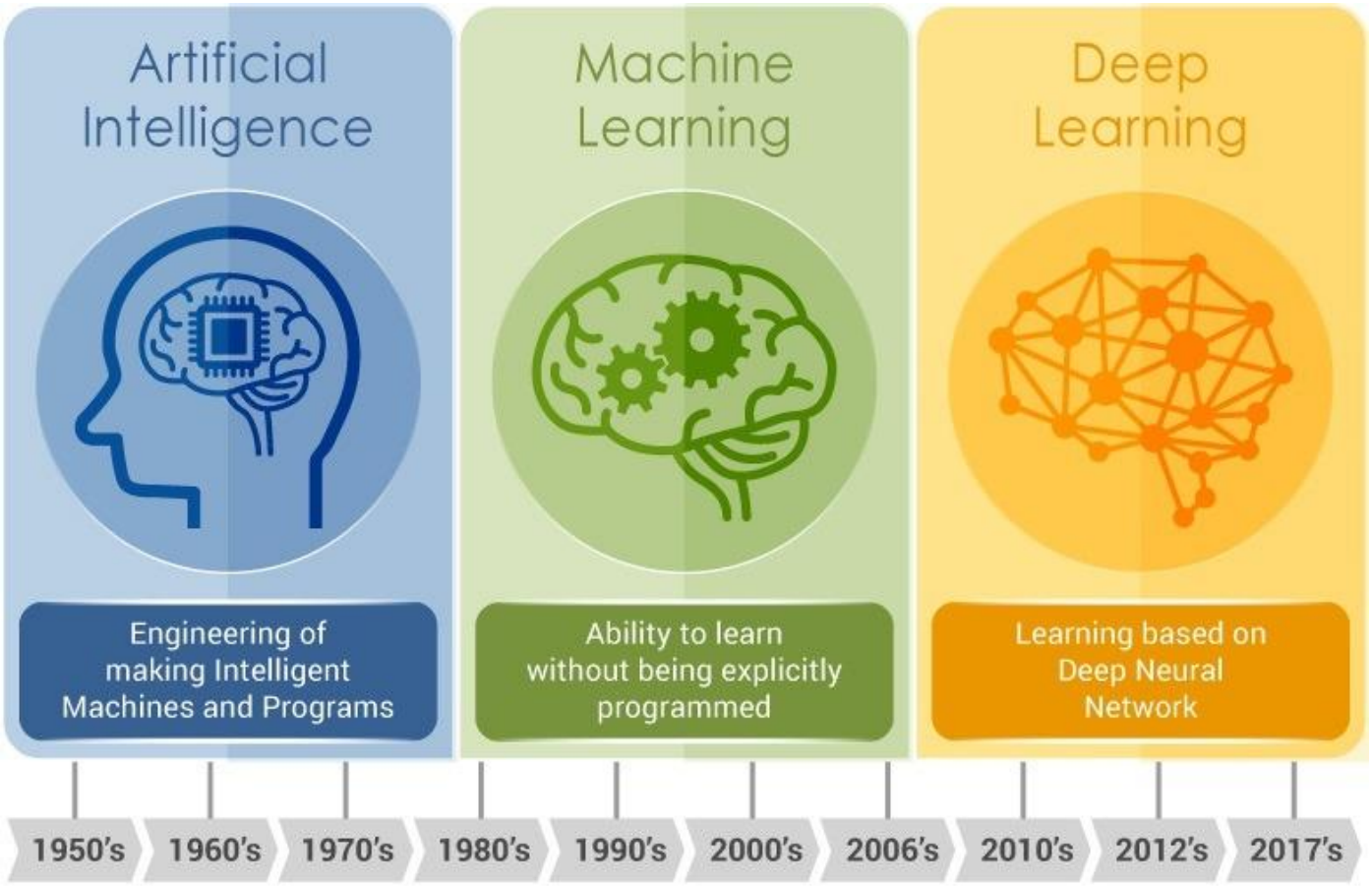
LARGE LANGUAGE MODELS FOR GENOMICS

Raphaël MOURAD, Associate Prof.
MIAT INRAe
Université Paul Sabatier, Toulouse III

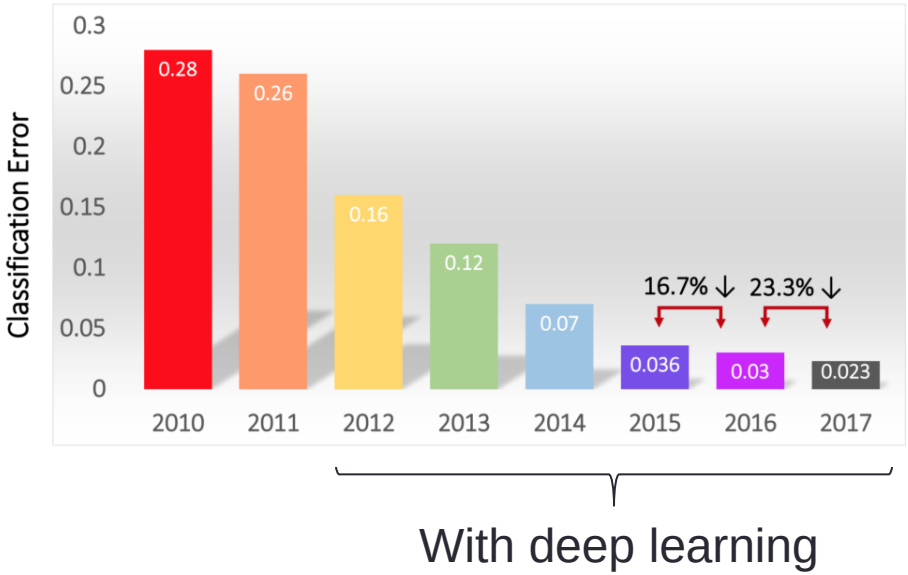
QUICK INTRO TO LARGE LANGUAGE MODELS

Deep learning as a branch of AI



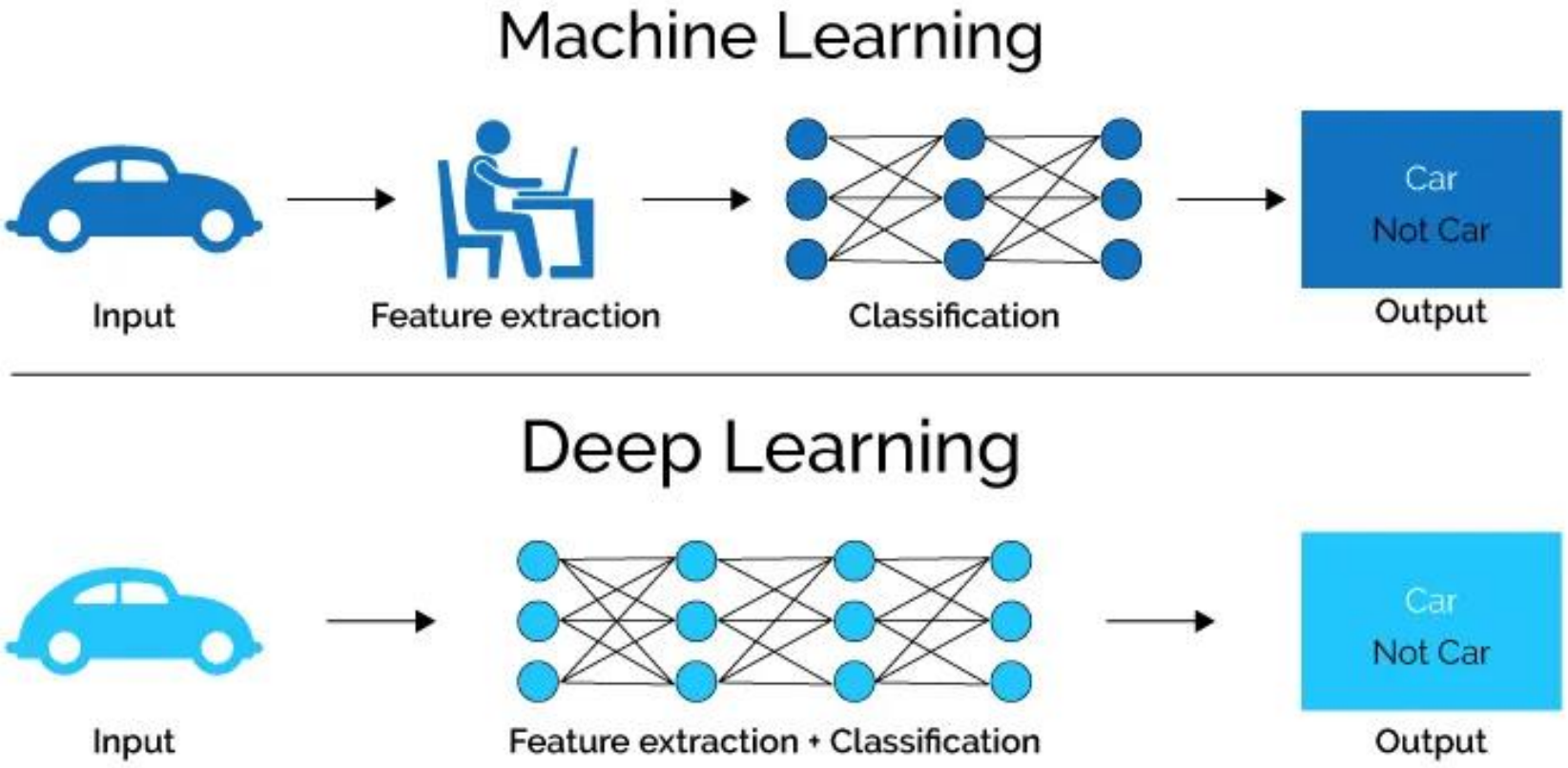
Success of deep learning since 2012: Example of computer vision

Image classification
(ImageNet challenge)

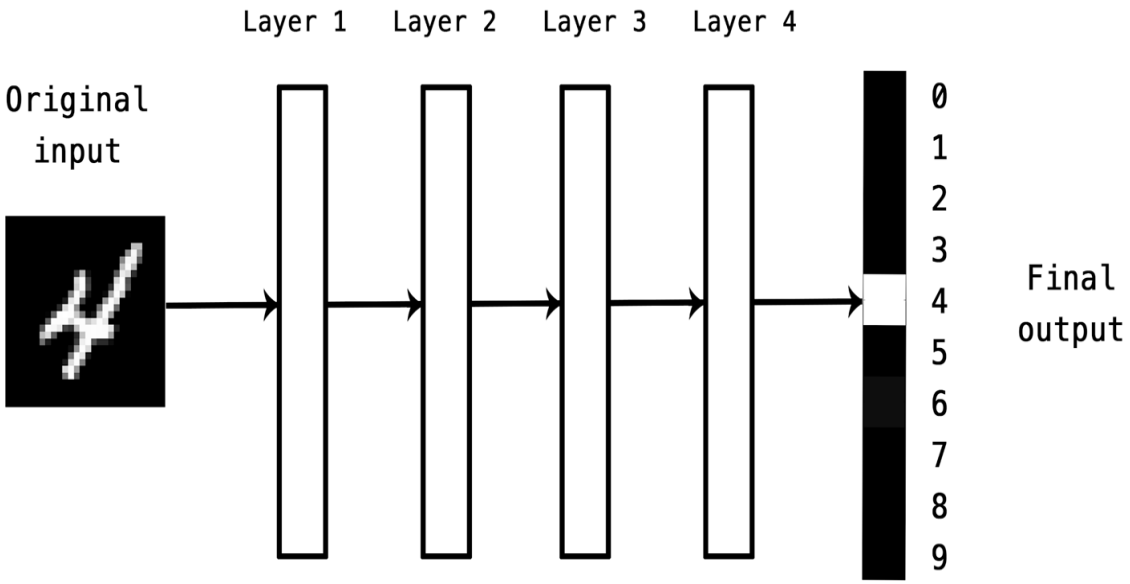


- 2012: AlexNet (convNet)
- 2013: ZFNet
- 2014:
  - VGGNet (deeper, simpler)
  - InceptionNet (faster)
- 2015: ResNet (deeper)
- 2016: Ensemble networks

Difference between machine and deep learning

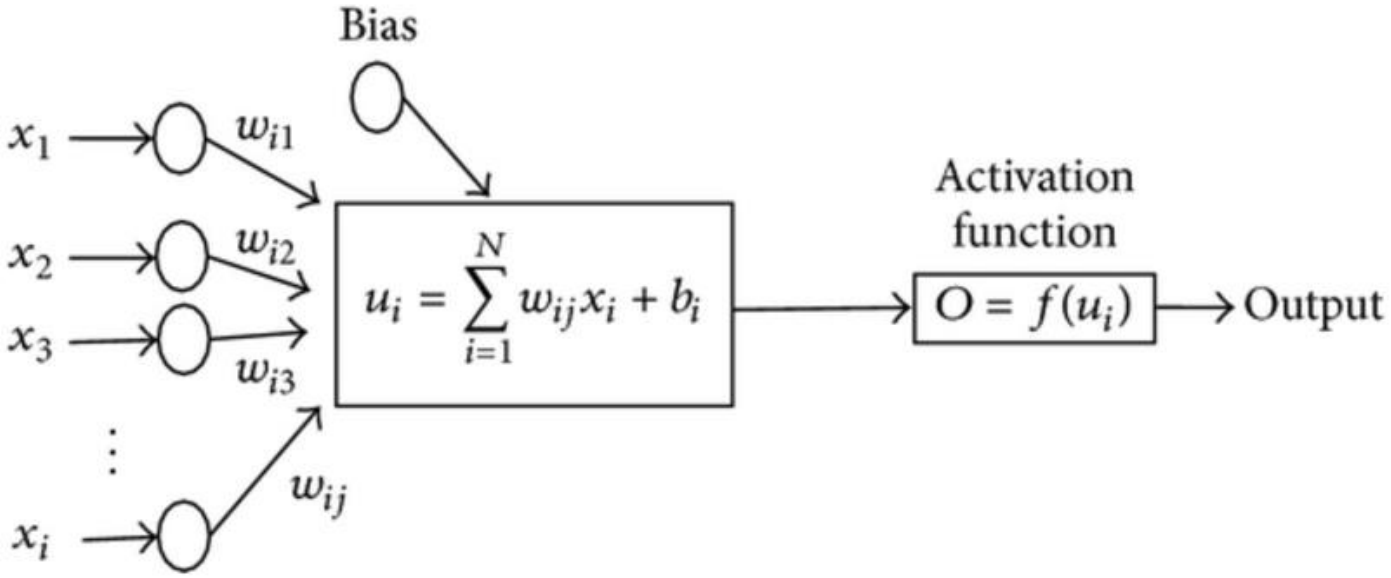


Deep learning as neural networks



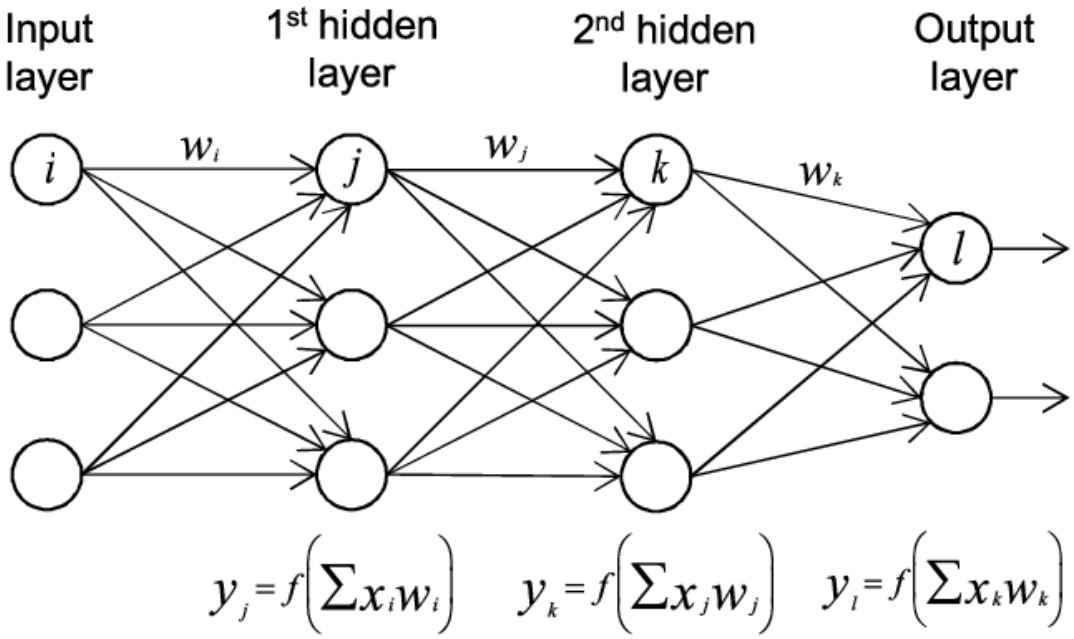
- Deep learning is based on a **deep** neural network which is the stacking of different neuronal layers to predict a final output.

Neural networks



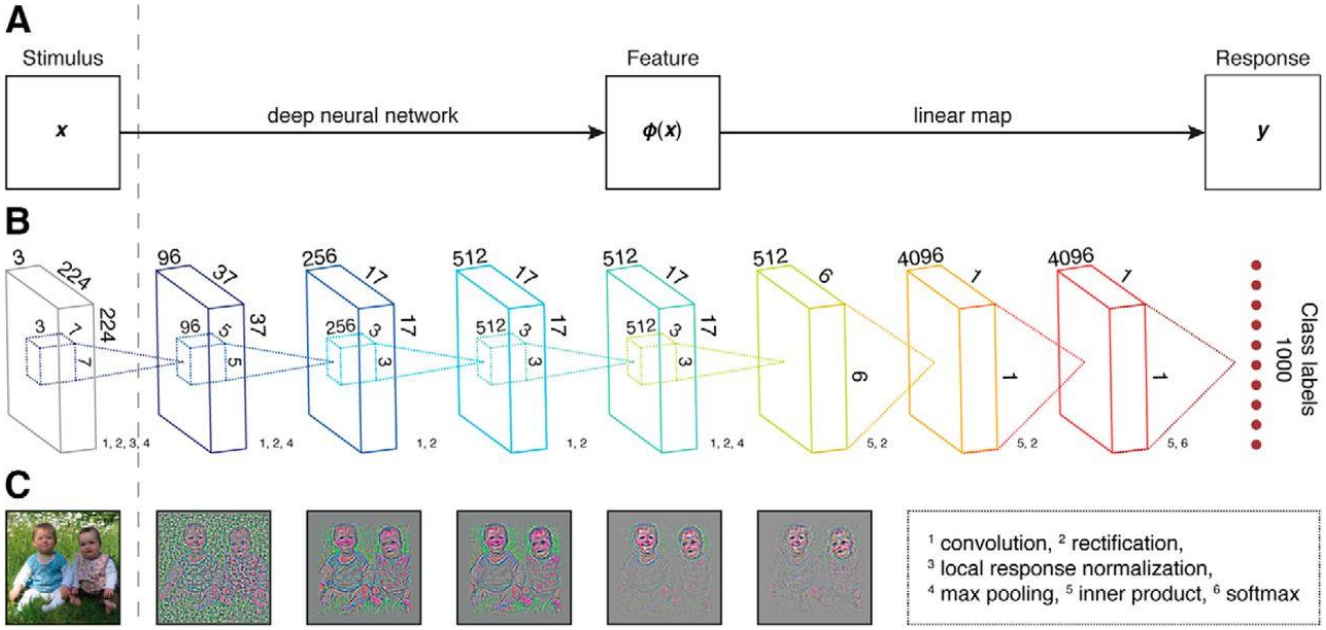
- In a neural network, multiple inputs x_i are combined through a linear combination (with weights w_i), and then an activation function is used for a non-linear transformation to obtain the output.

Deep neural networks



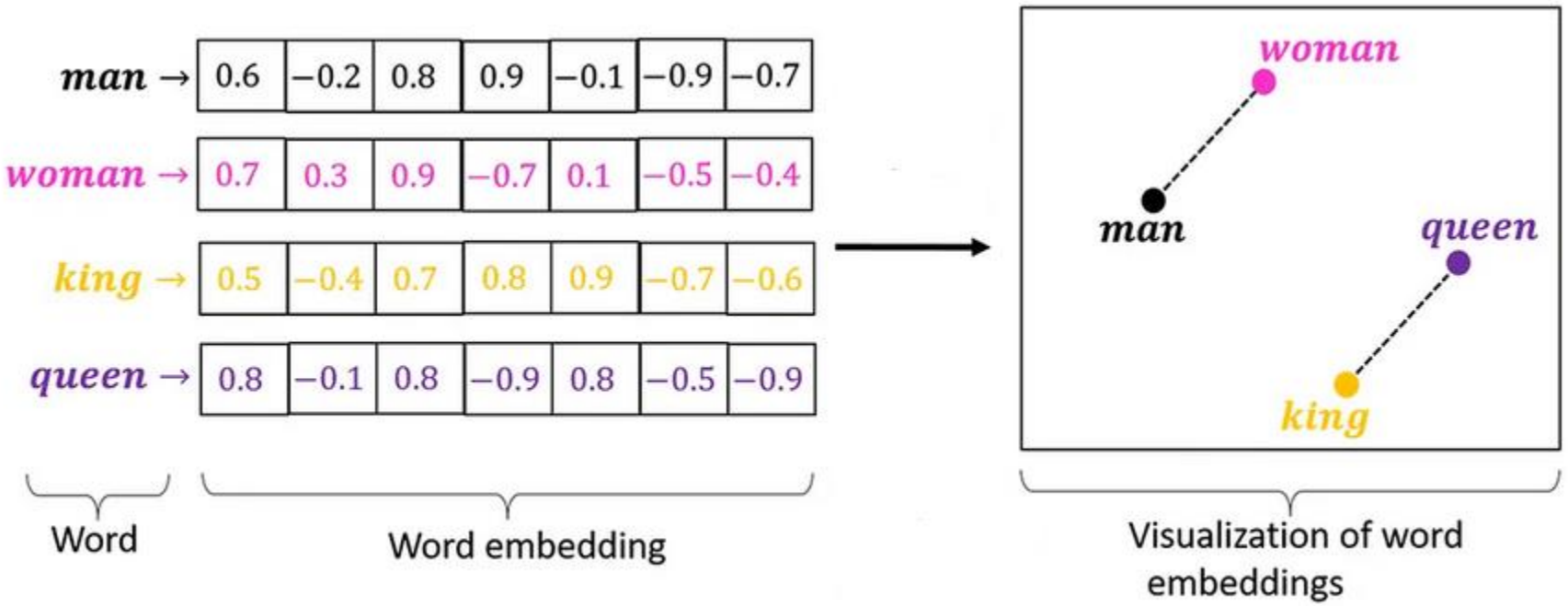
- A deep neural network (DNN) is an neural network (NN) with multiple layers between the input and output layers. Each hidden layer linearly combines the output from the previous layer and then does a non-linear transformation.

Principle of deep learning : stacking many different sorts of layers



- Building a deep neural network is like assembling a lego toy where every lego brick is a layer.

Word embedding



- Embeddings are vectors representing word meaning (word semantic). Embeddings can be produced by an LLM (last layer of the model).

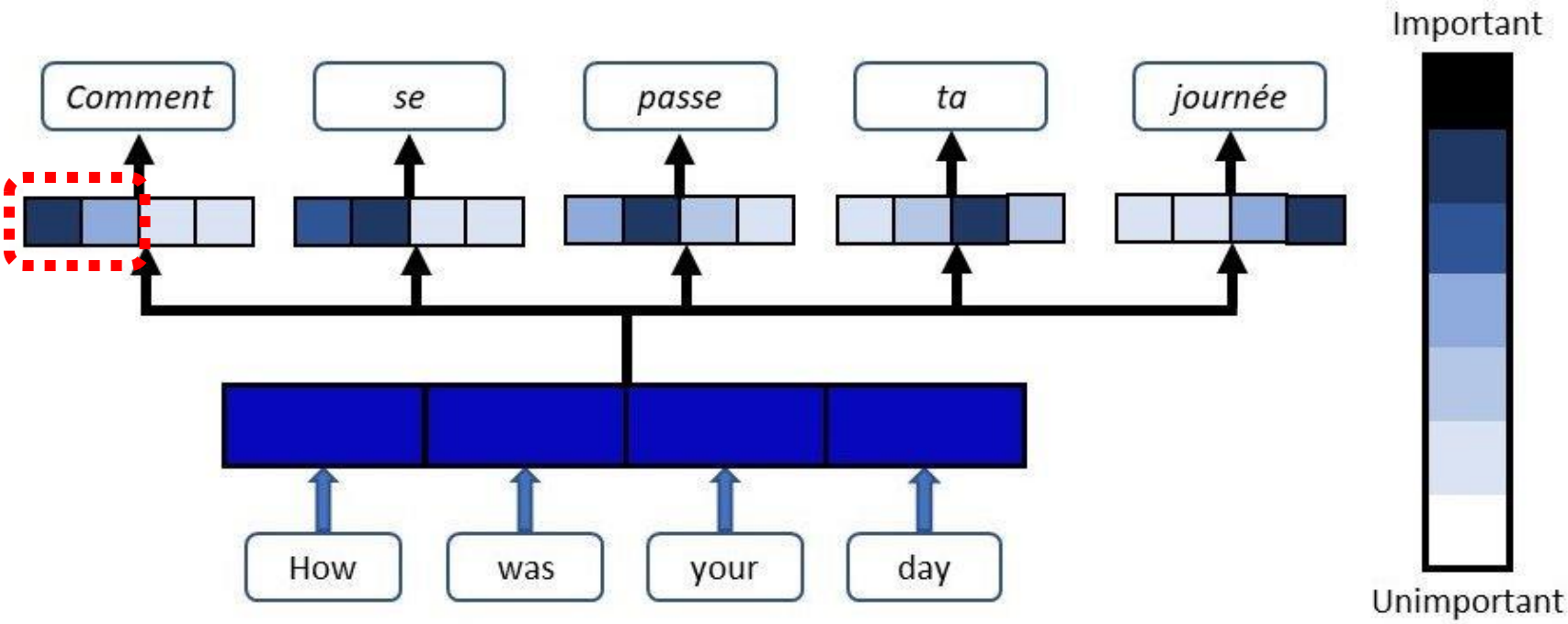
Word embedding

- The embedding of a word is computed given its context (eg the sentence or the paragraph of the word).

- For instance, the word “original” can mean:
  - "authentic, traditional", or
  - "novel, never done before“.

- The embeddings of the same word but with different meanings depending on the context will have different values.

Attention



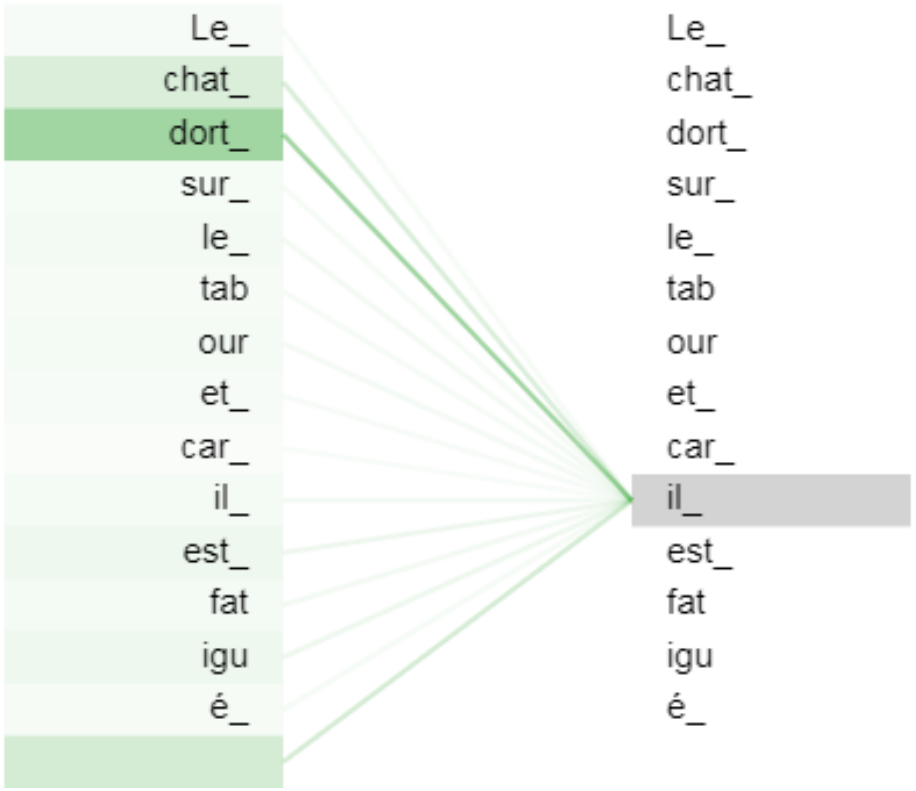
- Here, attention is used to weight different words from English to translate into French.
- For instance, to translate “how” to “comment”, you don’t only need the word “comment” (high weight) but you need other words such as the word “was” (moderate weight).

Self-attention in transformer model

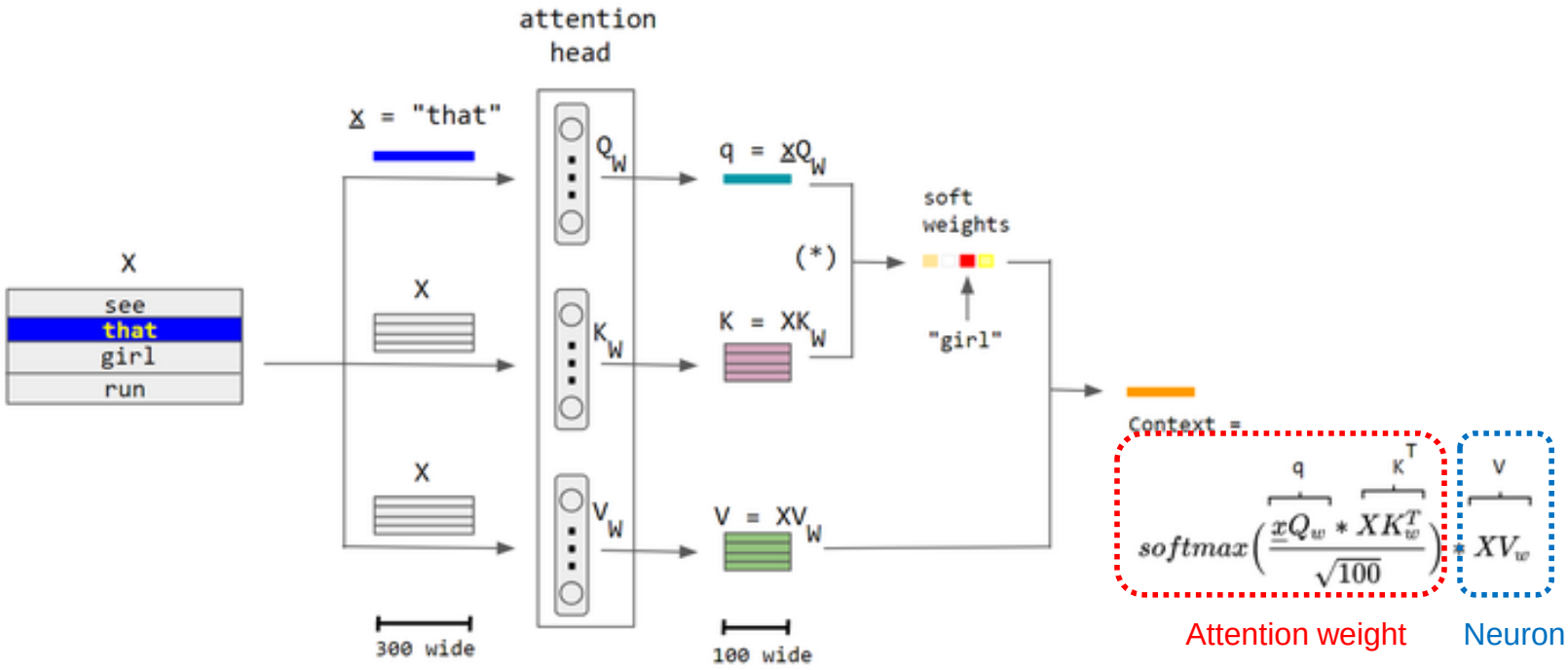
Self-attention is similar to attention, except that it is applied to the input sequence itself.

Self-attention allows to model long-range dependencies between any word in a sequence.

Self-attention is not directional as compared to RNN (LSTM or GRU), allowing parallel computing.

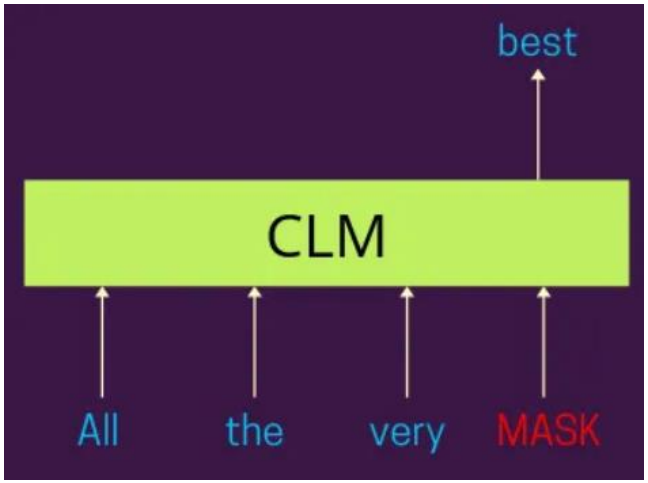


How to compute self-attention weights



Causal language modeling (CLM) vs masked language modeling (MLM)

Output

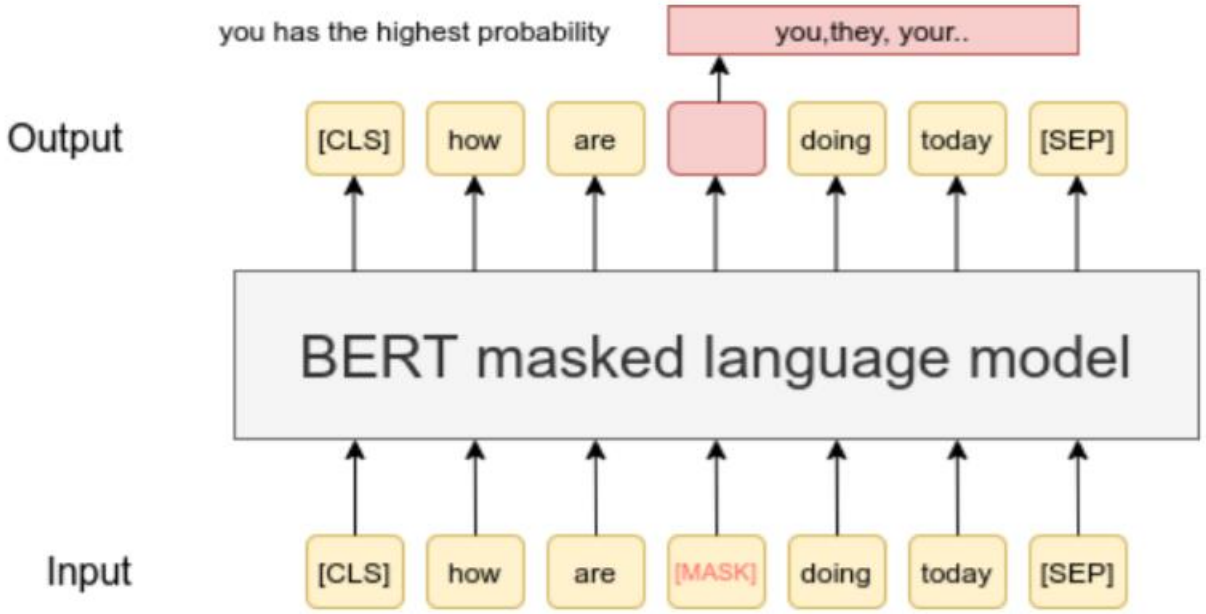


Input:
Unidirectional (left -> right)

Output

Input
Bidirectional

BERT model (masked language model)



- In a masked language model, the task is to predict some words that were masked using the context of the words.
- Together with self-attention, it was used for the BERT model (Bidirectional Encoder Representations from Transformers).

GPT-1 model (causal language model)

- GPT-1 implements the transformer architecture (self-attention).

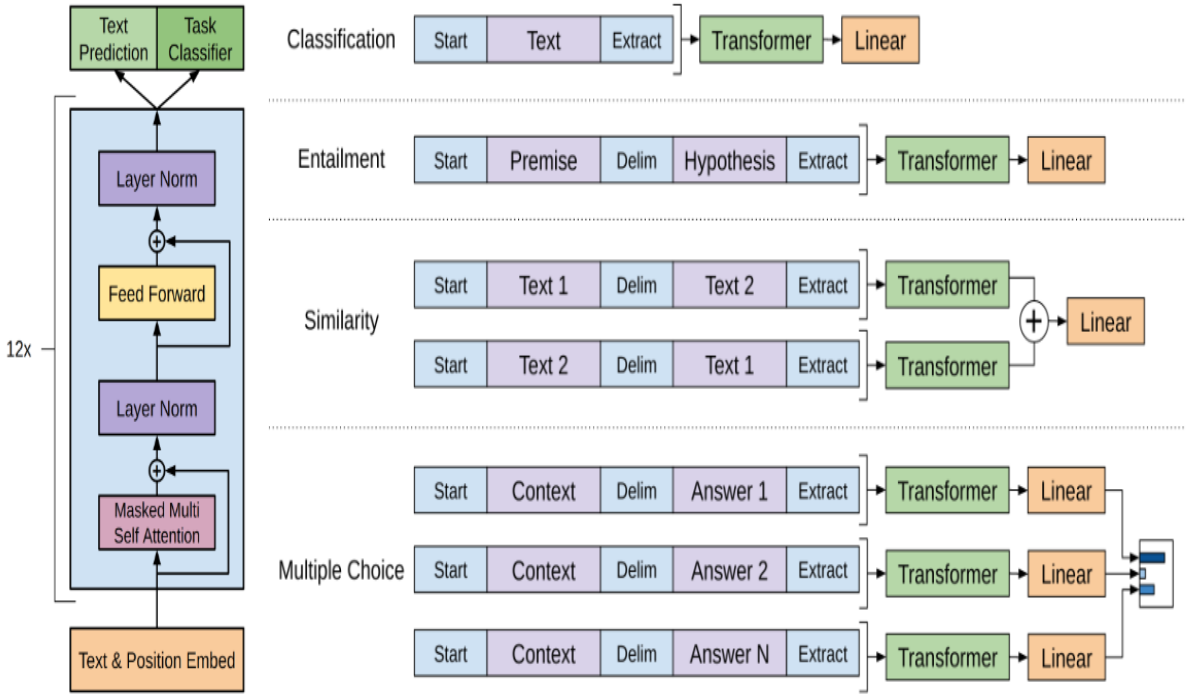


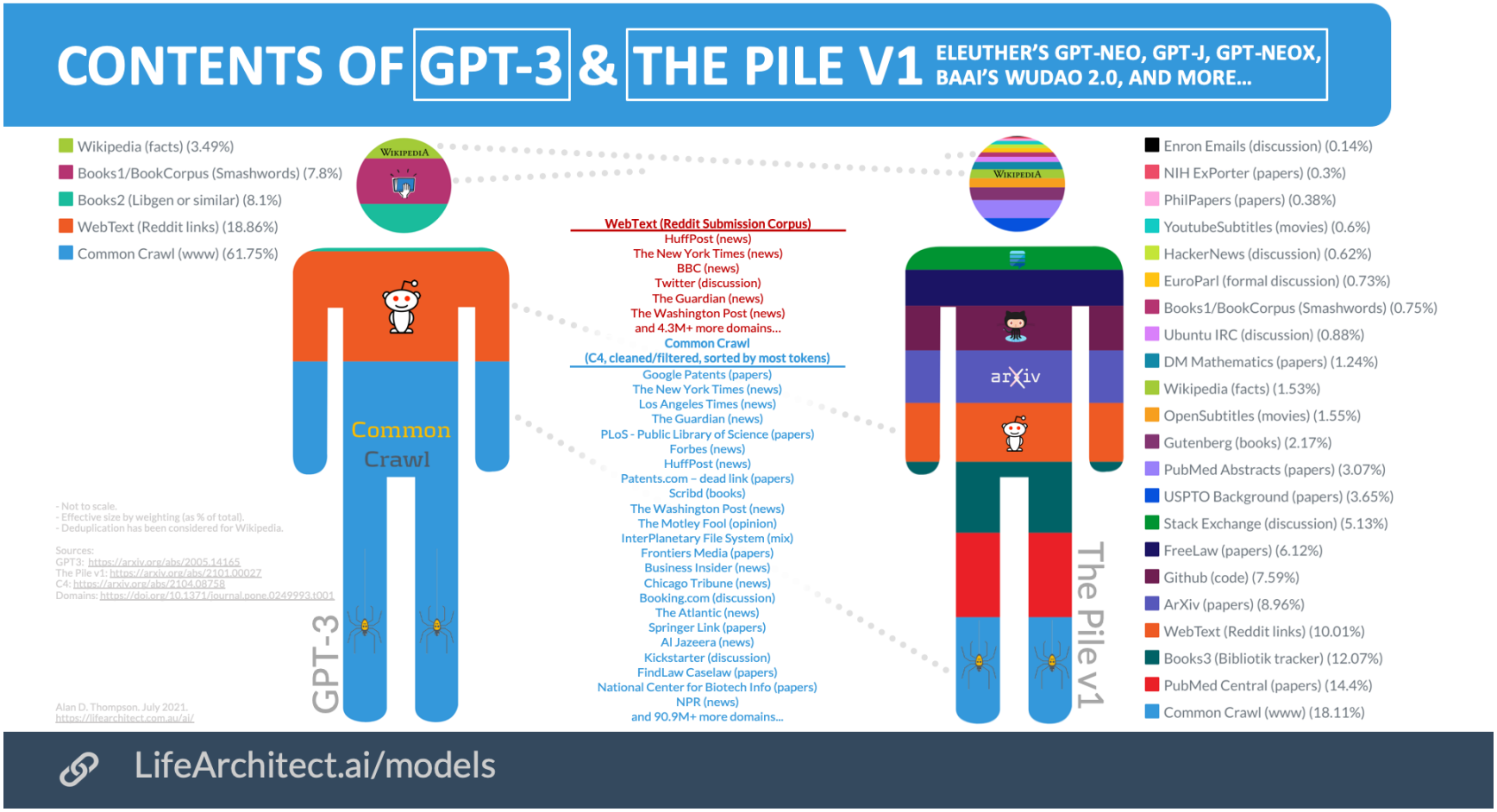
Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

GPT-1, GPT-2 and GPT-3

	GPT-1	GPT-2	GPT-3
Parameters	117 million	1.5 billion	175 billion
Decoded Layers	12	48	96
Hidden Layer	768	1600	12288
Context Token size	512	1024	2048

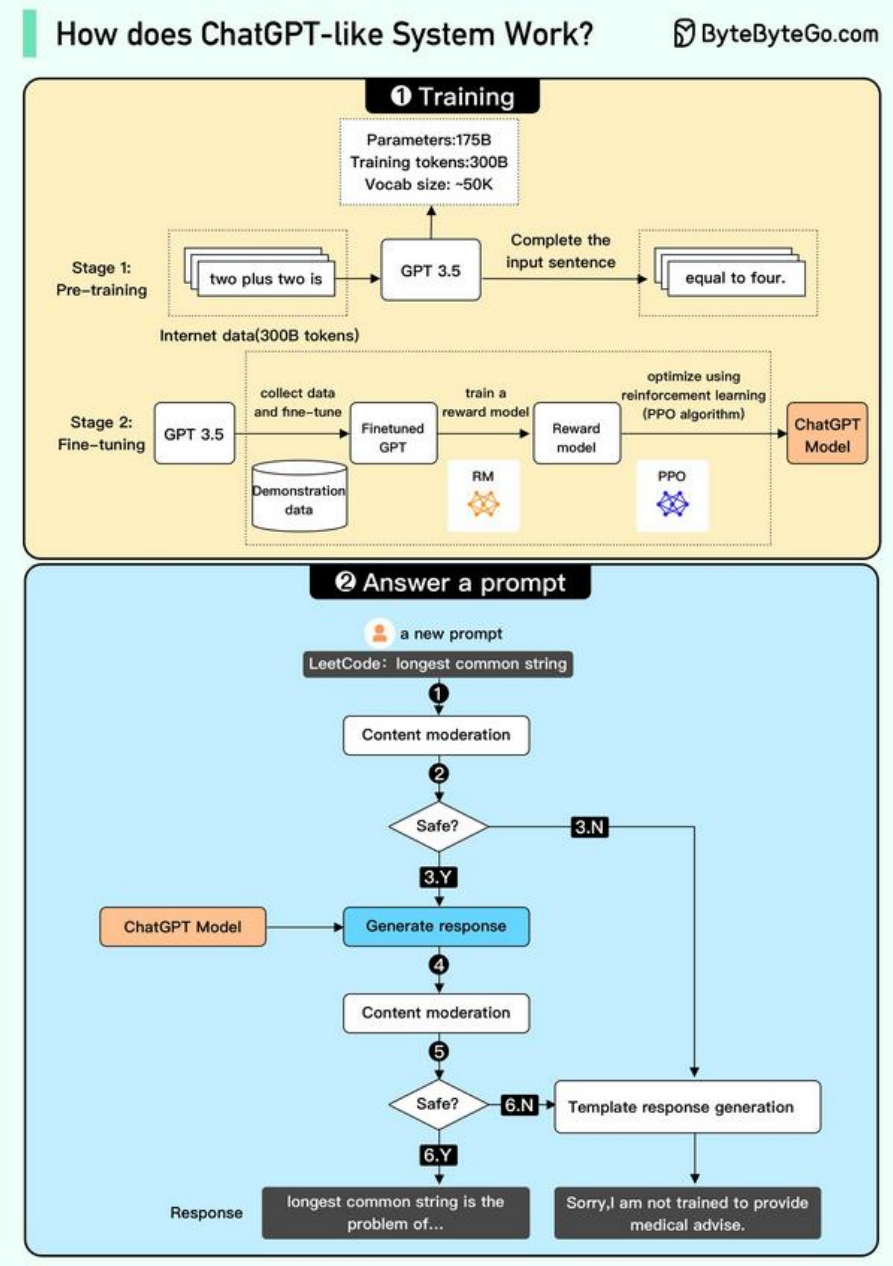
- Number of parameters increasing over time.

GPT-3 training data



chatGPT

- Based on gpt3.5 (pretraining), with fine-tuning (stage 2) and reinforcement learning with human feedback (in blue).



DEEP LEARNING FOR GENOMICS

Human genome

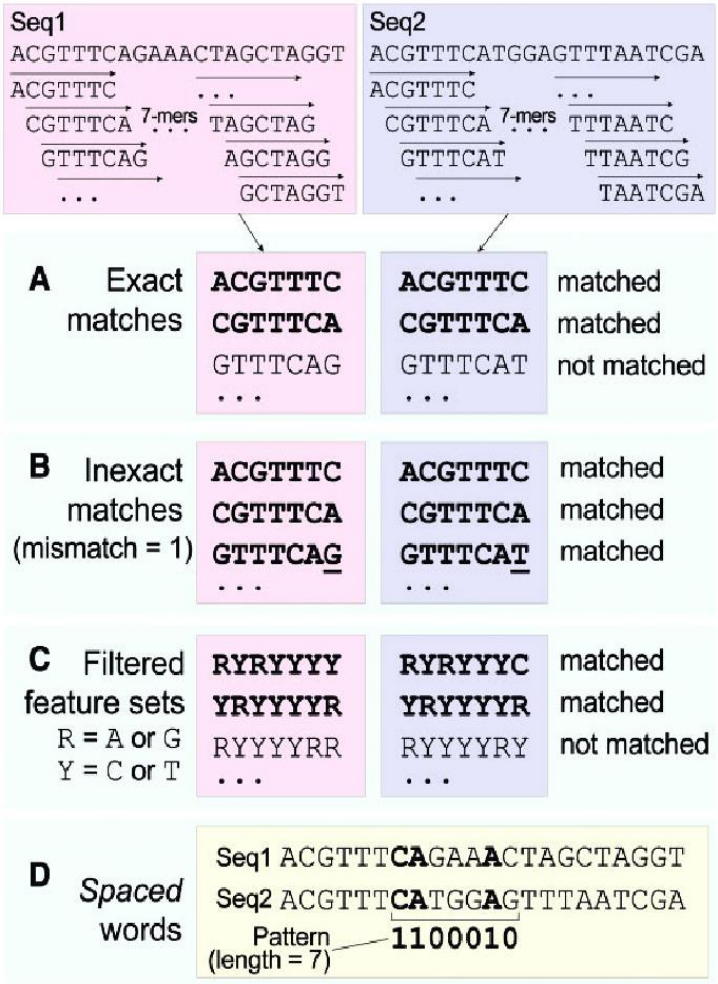
- The human genome is composed on 22 autosomal chromosomes, two sex chromosomes and a mitochondria. It is composed of 3,3 billion DNA letters (A, T, G, C).

- Bellow is an example of a DNA sequence:

TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCTAACCCTAACCCTAACCCTAACCCCTAACCCTAACCCTAAACCCTAAACCCTAACCC
TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCCTAACCCTAACCCTAACCCTAACC
GAGGAGAACGCAACTCCGCCGTTGCAAAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCC
CACATGCTAGCGCGTCGGGGTGGAGGCGTGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCG
AAGCCTACGGGCGGGGGTTGGGGGGGCGTGTGTTGCAGGAGCAAAGTCGCACGGCGCCGGGC
GCTTGCTCACGGTGCTGTGCCAGGGCGCCCCCTGCTGGCGACTAGGGCAACTGCAGGGCTCTCT
GCACGCCCACCTGCTGGCAGCTGGGGACACTGCCGGGCCCTCTTGCTCCAACAGTACTGGCGGA

How to represent a DNA sequence for a large language model?

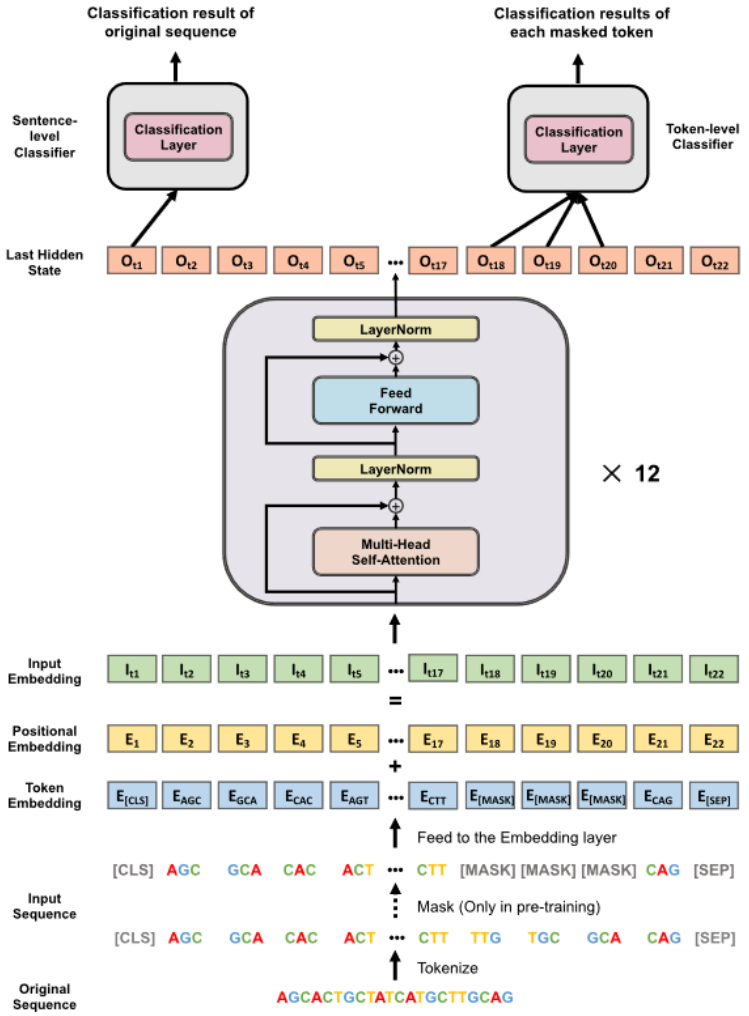
- K-mers (for instance: ATCTC or ATTTC, …): very powerfull approach since almost no prior information (except k) is needed to build the features.



DNABERT

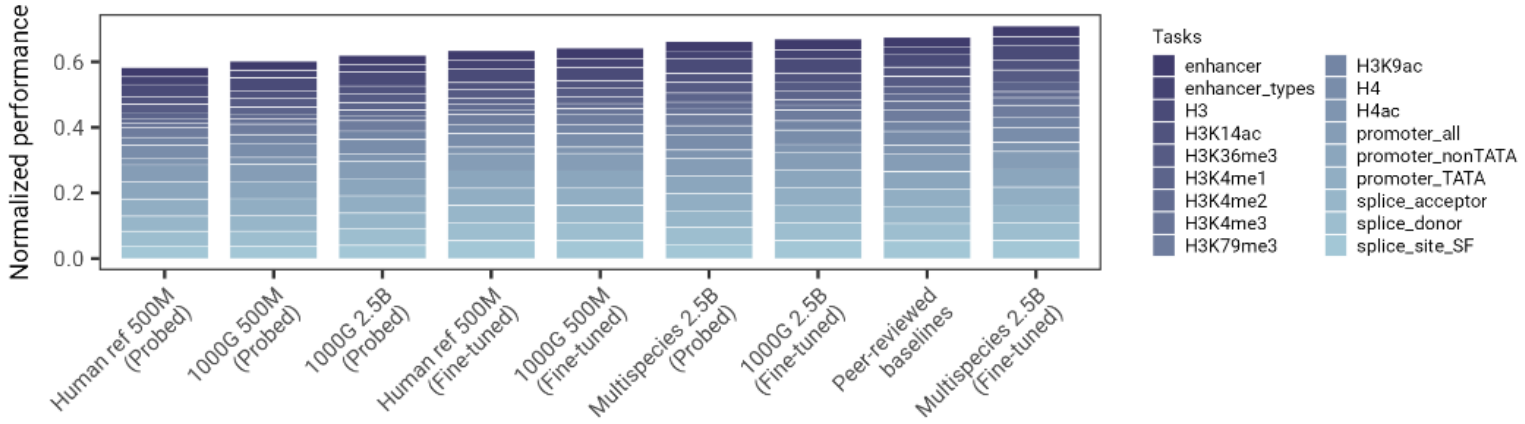
- The self-attention model DNABERT is trained by masking some kmers in the DNA sequence and then by trying to predict them using the other k-mers in the DNA sequence (context).

- At the end, the model provides features that encode DNA sequences in a very efficient way for any predictive task.



Nucleotide Transformer

- BERT model trained not only from the human genome (3.3 Gb) like DNABERT, but trained from thousands of human and animal genomes!



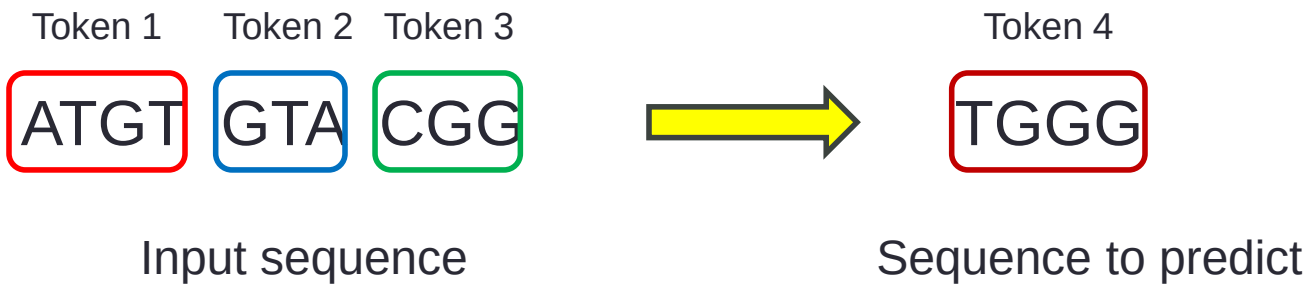
- The prediction performance increases with the number of genomes used during training.

Mistral-DNA

- Mistral-DNA is an LLM based on Mixtral-8x7B-v0.1 model.
- Causal language model.
- Based on Byte pair encoding (BPE) tokenizer:
  - AGCCTTTCTCT -> AGC***Z***TT***ZZ***, where ***Z***=CT
  - Allows to identify most frequent k-mers
- Sparse mixture of Experts: reduces number of parameters used during inference (prediction)

Mistral-DNA

- Prediction:

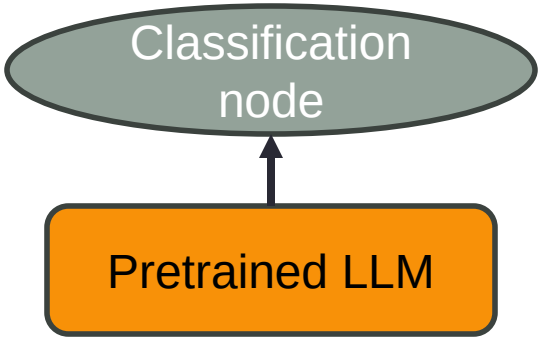


- The model is « pretrained » using this strategy. Given a DNA sequence, it tries to predict the next k-mer (=next word).

APPLICATIONS

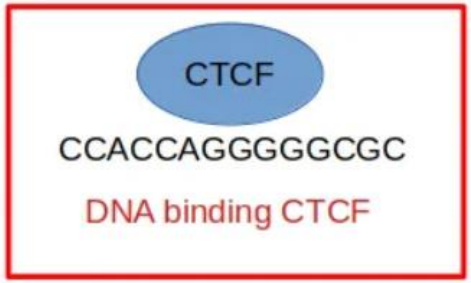
1) Finetuning for classification

- 1st step: modify the model

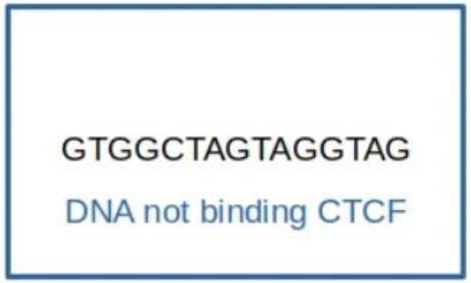


Add classification node at the top of the LLM.

- 2nd step: train the model on labeled sequences (=finetuning).

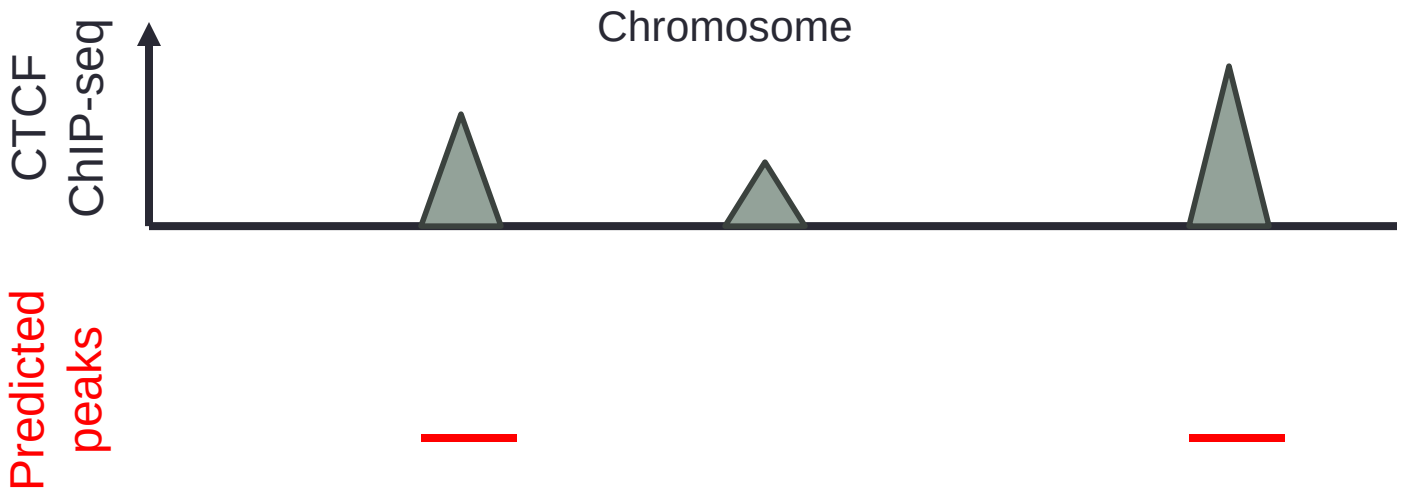


Label 1



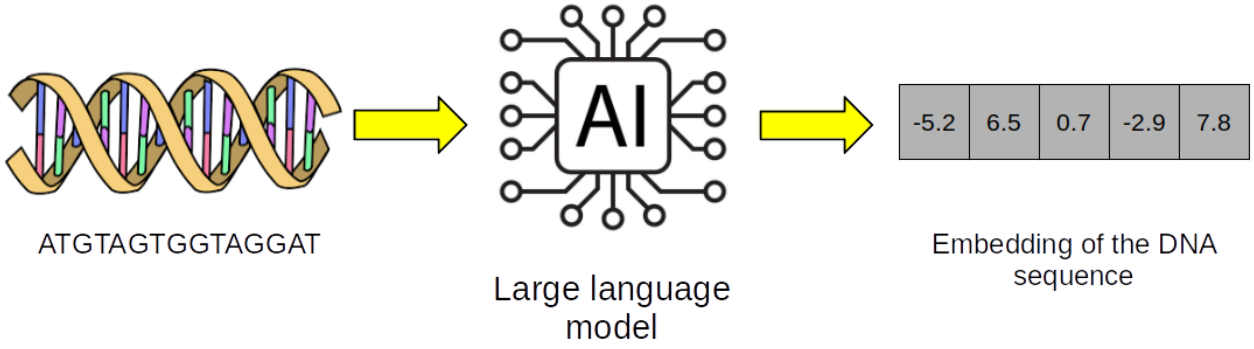
Label 0

1) Finetuning for classification



- The finetuned model can be used to predict CTCF binding peaks along the genome.

2) Assessing the impact of a SNP



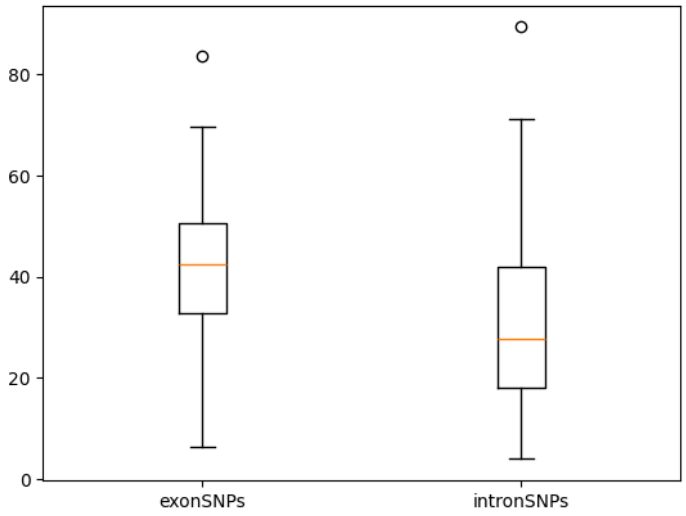
- The LLM is a way to convert a DNA sequence into a vector of numbers (an embedding).
- We can assume that if two sequences are functionally different, they should have different embeddings.

2) Assessing the impact of a SNP

- Easy way to assess the impact of a SNP:
  - Predict the embedding for a sequence with the reference allele **C**:
    - ATGTAGTGGGTACC**C**TGTGTAGAAGCCA
  - Predict the embedding for a sequence with the reference allele **T**:
    - ATGTAGTGGGTACC**T**TGTGTAGAAGCCA

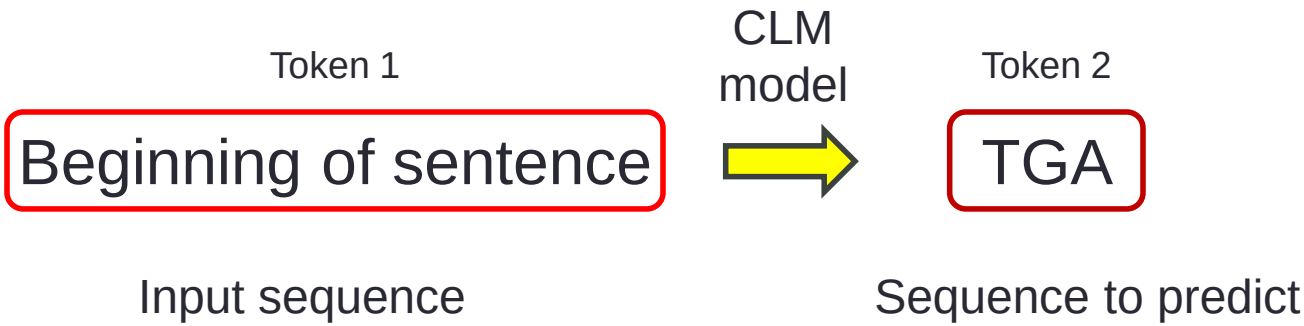
Compute the L2 distance (for instance) between the two embeddings.

On the right, L2 distances computed for a set of SNPs inside exons and a set of SNPs inside introns.

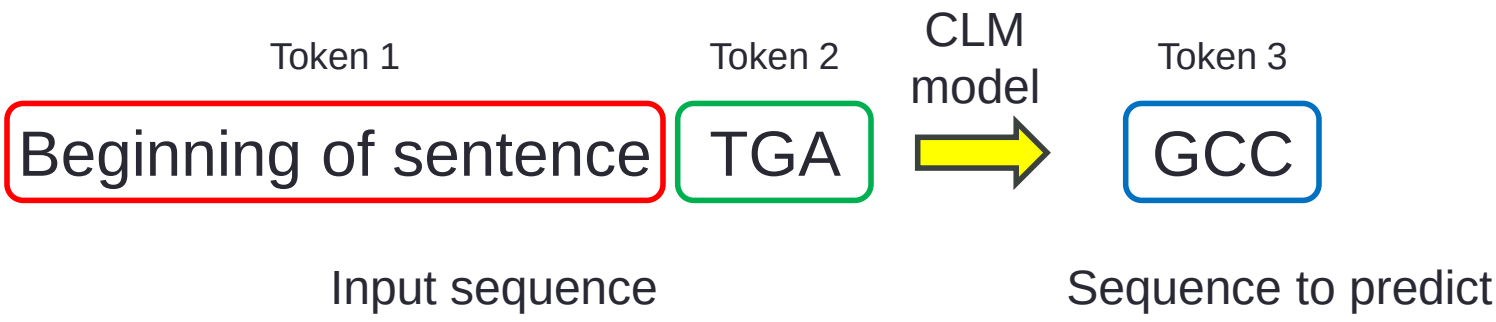


3) Synthetic DNA sequence generation (only works with CLM models)

- Predict a synthetic DNA sequence:



- Repeat same task iteratively:



3) Synthetic DNA sequence generation

```
# THESE ARE 5 ARTIFICIAL YEAST DNA SEQUENCES
artificial_sequences[0:5]
```

```
['CCTTGAATACCCGATGACGAAAAAAGTTCGTTCAACAGAAGAAAGATTTCAAGGAATGAAGACGAAACTATCGAGAAGAGGCAAAGAAAACTGGCGAATT',
 'TCTTCGACTATCGATGATTAGCAAGTAGCGTAAACGGGTTCAAAACTTCGTTCAATATCATTTTTCAATTTAAAAATCACGCTAACGAAGAATAATTTGA',
 'GTACGATAGAACGTTGTTGCTACAAGAATTGGCAATTTCAATTTGCTTTCGTCAATAACGAAAAAATTAATGAGCGTAAATTACCAATCTTCGGTATT',
 'CTTTCGTTGTACGATCTTCCAACGGTTTTTGAAGATTCCGATACTAACTTACAAAGAACGTTTCATCATTAAAATATTGGTTTGACGAATTATCTGTAAA',
 'TCAAGAATCTATACGATACGCATTTATCATATAATTTTTGATTATTGTATATTATTAATAACTTATGCAATAATAATAATATATTTATTAATTATAATTTATATTATAAATATAAA']
```

- For instance, for the 1st sequence, BLAST shows no matching to any known DNA sequence:

Job Title	**Nucleotide Sequence**
RID	WE1JYSHG016 *Search expires on 03-05 20:55 pm* Download All
Program	Citation
Database	core_nt See details
Query ID	lcl|Query_379957
Description	None
Molecule type	dna
Query Length	100
Other reports	

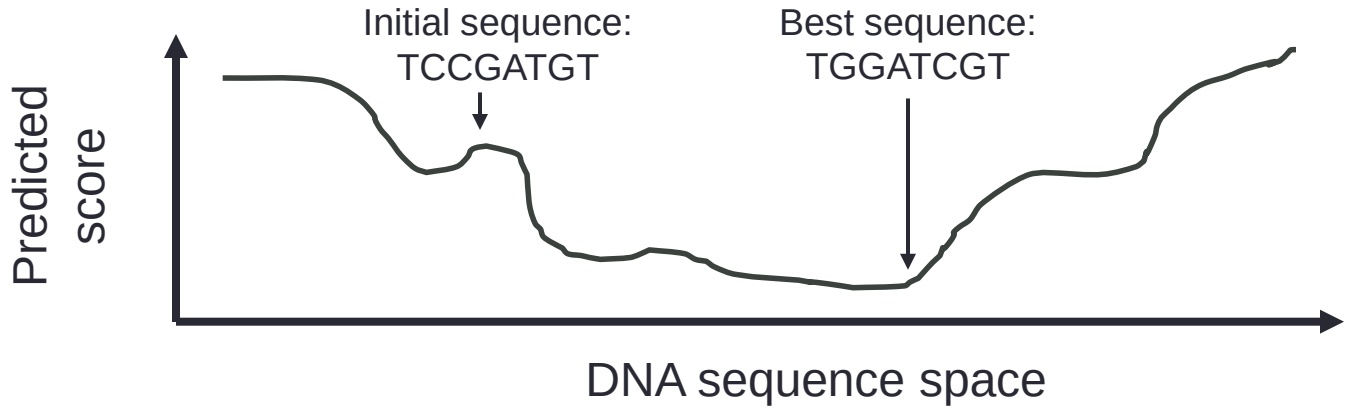
Filter Results

Percent Identity	E value	Query Coverage
to	to	to

Filter Reset

4) DNA sequence optimization

- DNA sequence optimization consists in finding the sequence with a highest (or lowest) score predicted by a previously training model.



- For instance, we have finetuned an LLM for promoter prediction. We can take a given DNA sequence and optimize it to be a strong promoter (high score).

4) DNA sequence optimization

- We can use black box optimization algos:
  - For instance, evolutionary algorithms;
  - Discrete optimization: A T G or C are not quantitative variables;