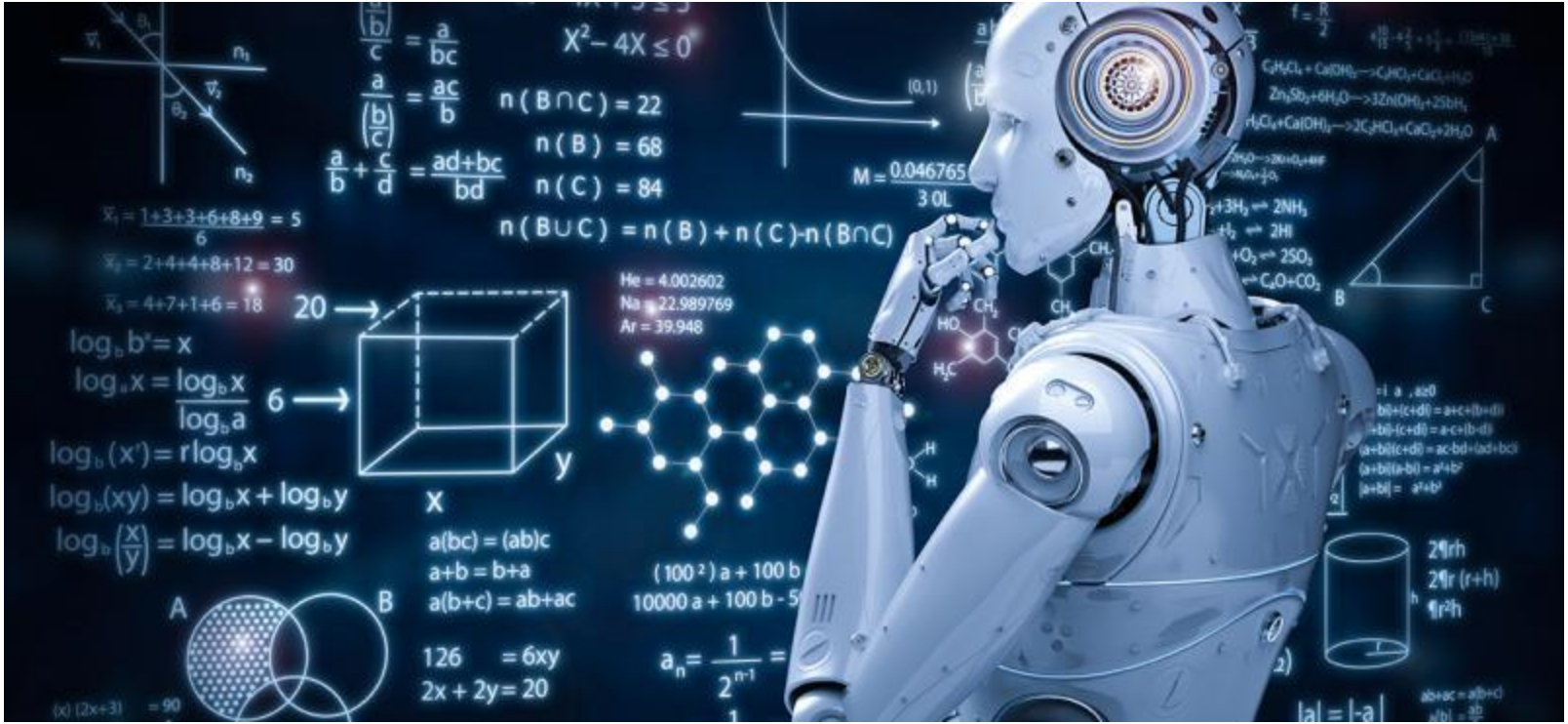
Part III – Machine learning (ML)

1



https://www.informatec.com/en/machine-learning

What is machine learning?

- Machine Learning is the most popular technique of **predicting** the **future** or **classifying information** to help people in making necessary decisions.
- Machine Learning algorithms are trained over instances or examples through which they learn from **past experiences** and also **analyze** the **historical data**.
- Therefore, as it trains over the examples, again and again, it is able to identify patterns in order to make predictions about the future.

https://data-flair.training/blogs/machine-learning-tutorial/

Machine learning, Deep learning, and Artificial Intelligence (AI)

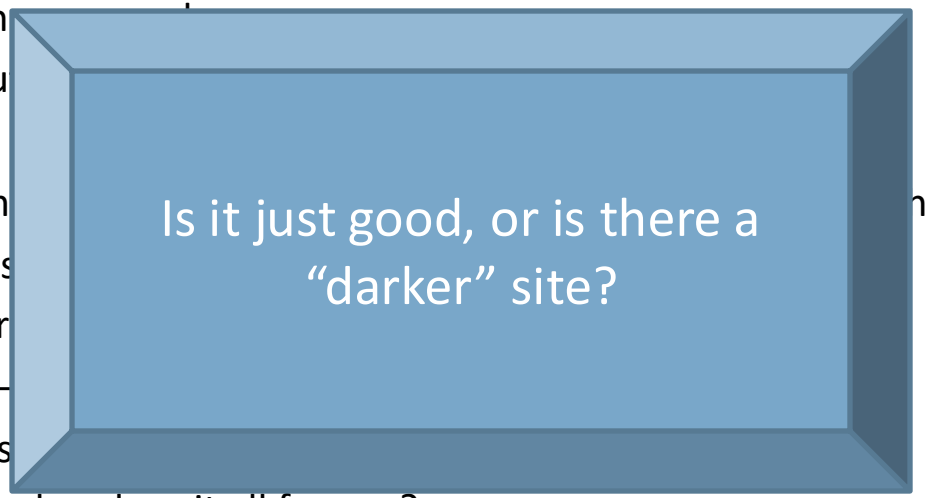
- **Machine Learning:** branch of AI - knowledge analysis that automates analytical model building
- **Deep learning:** broader family of ML methods supported artificial neural networks with representation learning
- **AI** is the greater pool that contains an amalgamation of all the above-discussed technologies.

Where do you encounter ML/AI in your day-2-day live?

ML is everywhere

- Correction when you type on your phone
- Speech recognition
- Which advert to pu
- SPAM filter
- Prediction which m
- Predict your friends
- Autonomous car dr
- single cell analysis -
- predicting protein s
- ChatGPT – and friends – does it all for you?

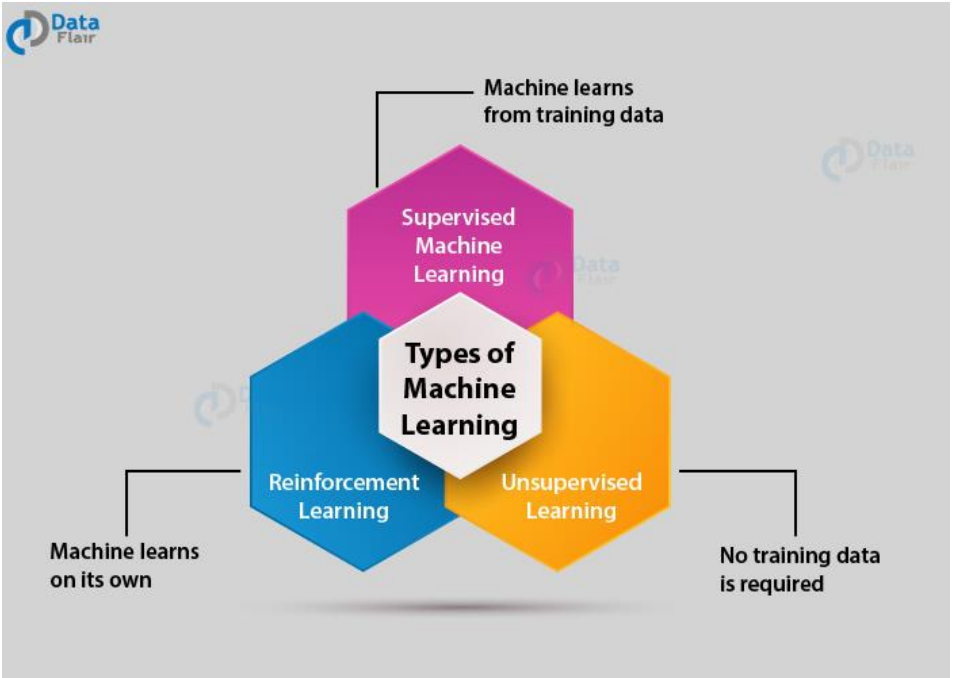
- **ALL BASED on collection of large data amount!**



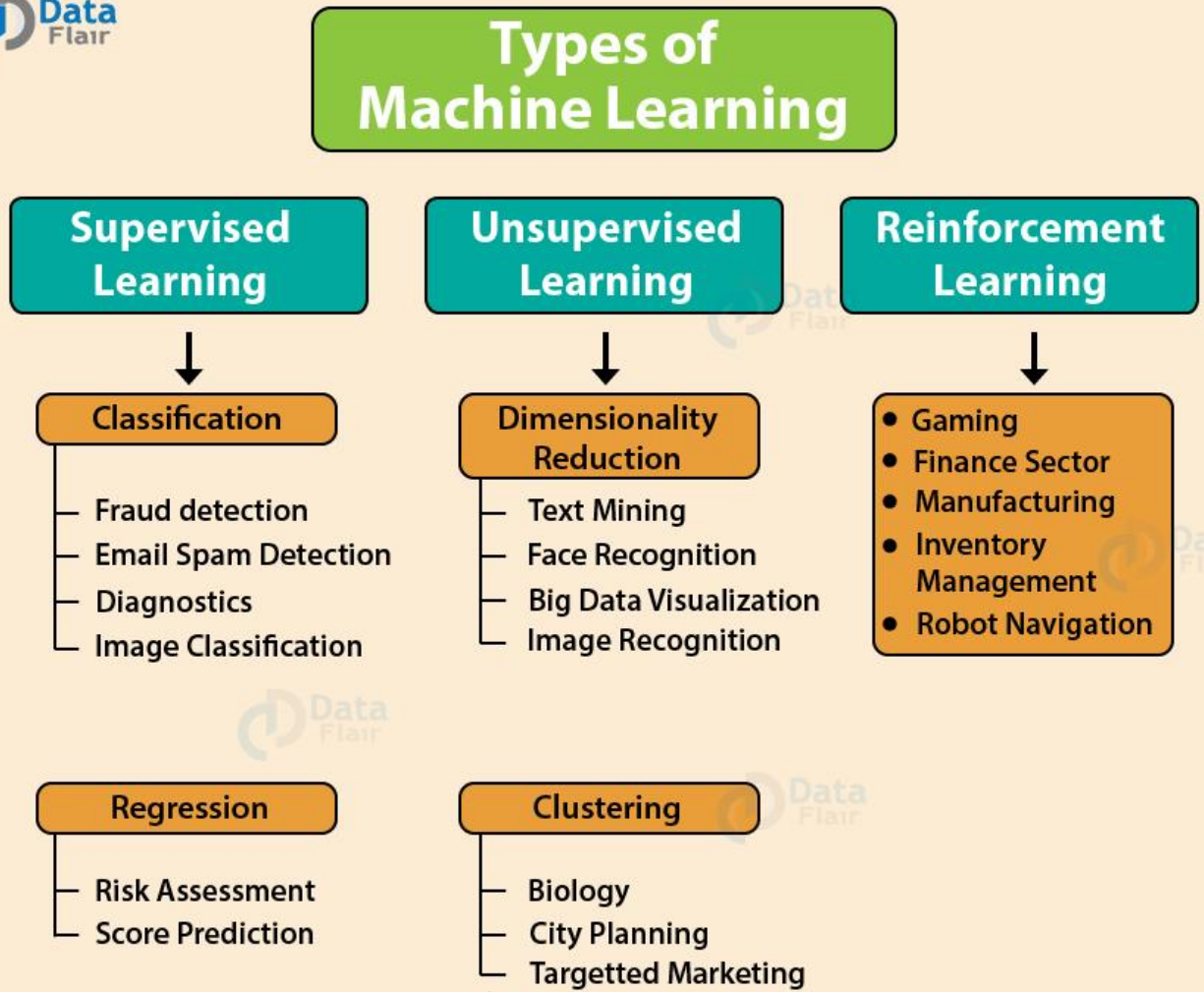
My approach

- Realise, it is everywhere
- What are the dangers?
- General concept of types
- No math ☹
- Some examples

7



https://data-flair.training/blogs/machine-learning-tutorial/



https://data-flair.training/blogs/machine-learning-tutorial/

Machine Learning Algorithms

1. **Regression -** Linear Regression
2. **Decision Tree Learning –** supervised methods to take decisions based on user input.
3. **Support Vector Machines -** to classify data into two categories or classes.
4. **Association Rule Learning -** customer will buy item Y if he has purchased the item X
5. **Artificial Neural Networks (ANN) –** mirror the brain – layers of neuros –
6. **Inductive Logic Programming -** produce a rule-like learning model.
7. **Reinforcement Learning –** self driving car: series of trials and errors to get better
8. **Clustering –** unsupervised methods – we did that!
9. **Similarity and Metric Learning -** learns to map similar objects together
10. **Bayesian Networks -** P (A/B) = P (B/A)*P (A)/P (B)
11. **Representation Learning -** algorithms are able to preserve the input data and essential information
12. **Sparse Dictionary Learning -** extension of representation learning
13. **Transformers** – pay attention
14. **Natural Language Processing (NLP)** – makes chatGPT talk to you

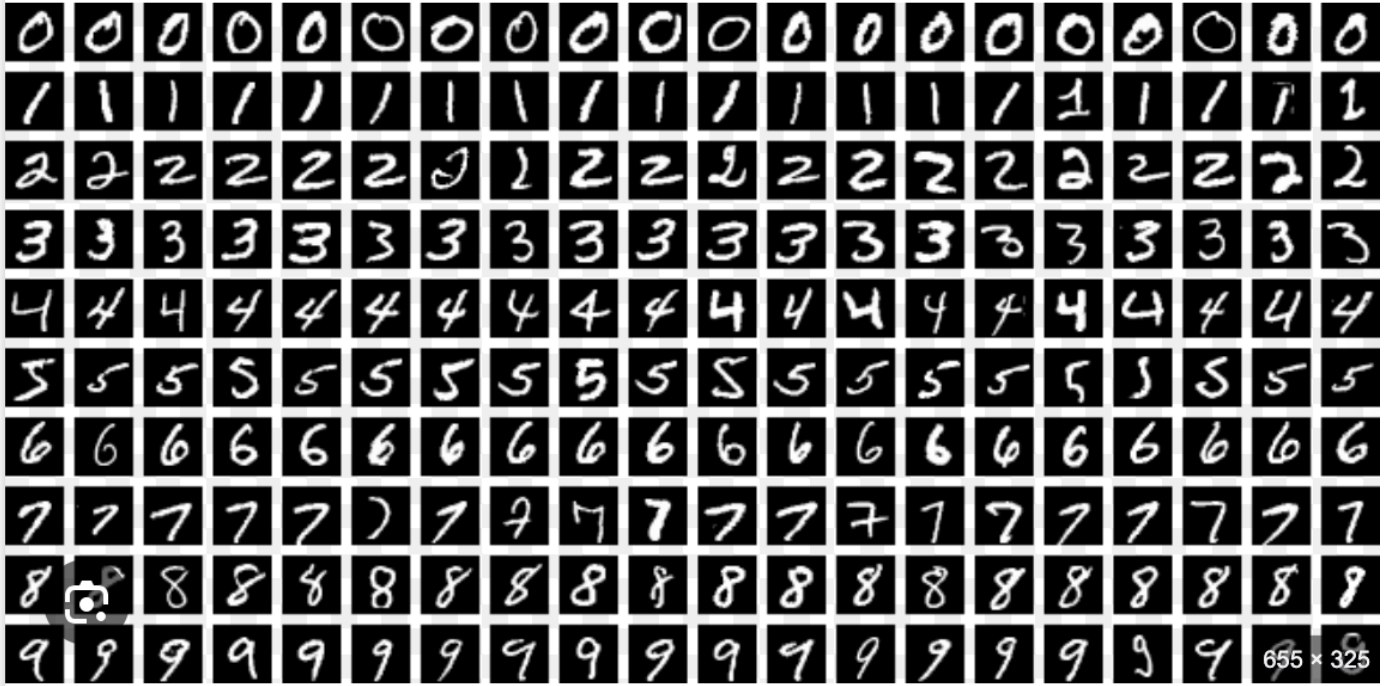
So far

- Clustering
- T-SNE / UMAP

- How can we use AI/ML in der interpretation of our data?

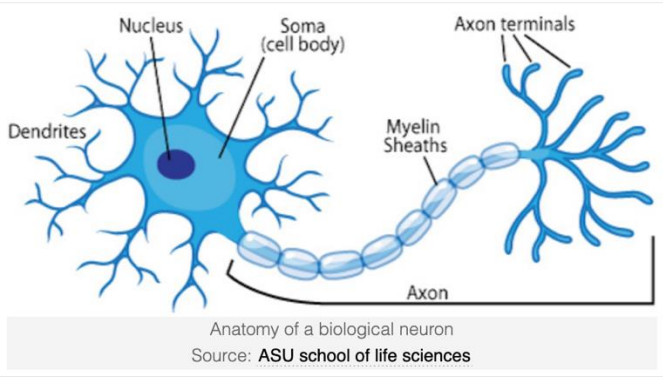
Classifier – perceptron - a simplified approach

- A classifier of numbers from 0-9
- How easy?

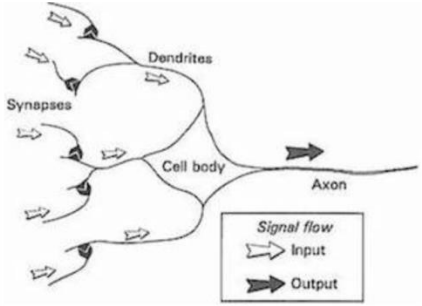


What we need…

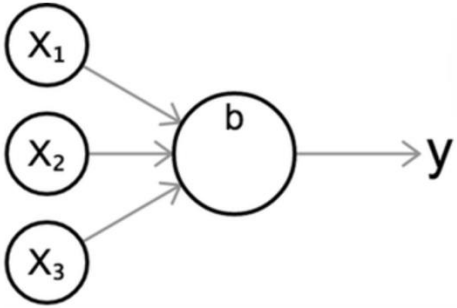
- We give it a number and want out, what number it is.
- Difficult as not clear
- Let's use a neural network



Anatomy of a biological neuron
Source: ASU school of life sciences



Simplified neuron body within a network
Source: Gurney, 1997. An Introduction to Neural Networks



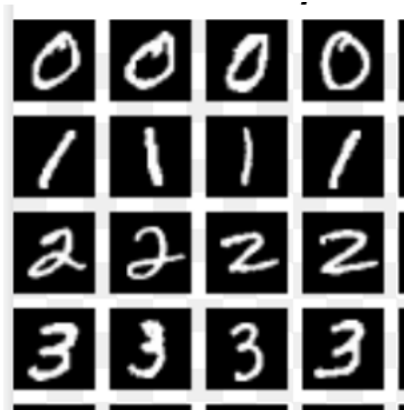
Artificial neuron

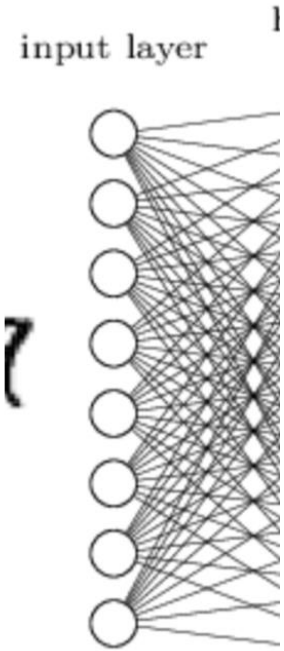
https://ml4a.github.io/ml4a/neural_networks/

https://www.kaggle.com/code/heeraldedhia/mnist-classifier-first-deep-learning-project

What is the input?

- Yes the number?
- What how are we going to give it to the network?

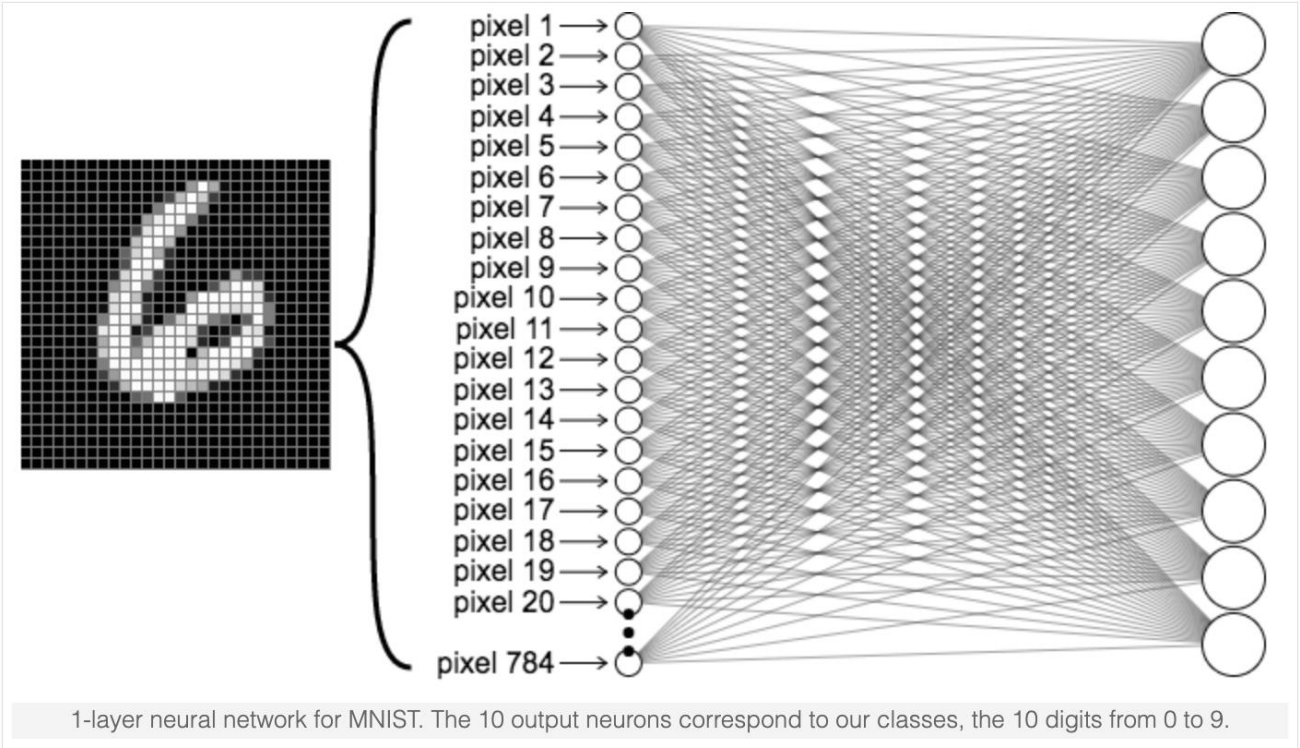




We need an ”embedding”

Any ideas?

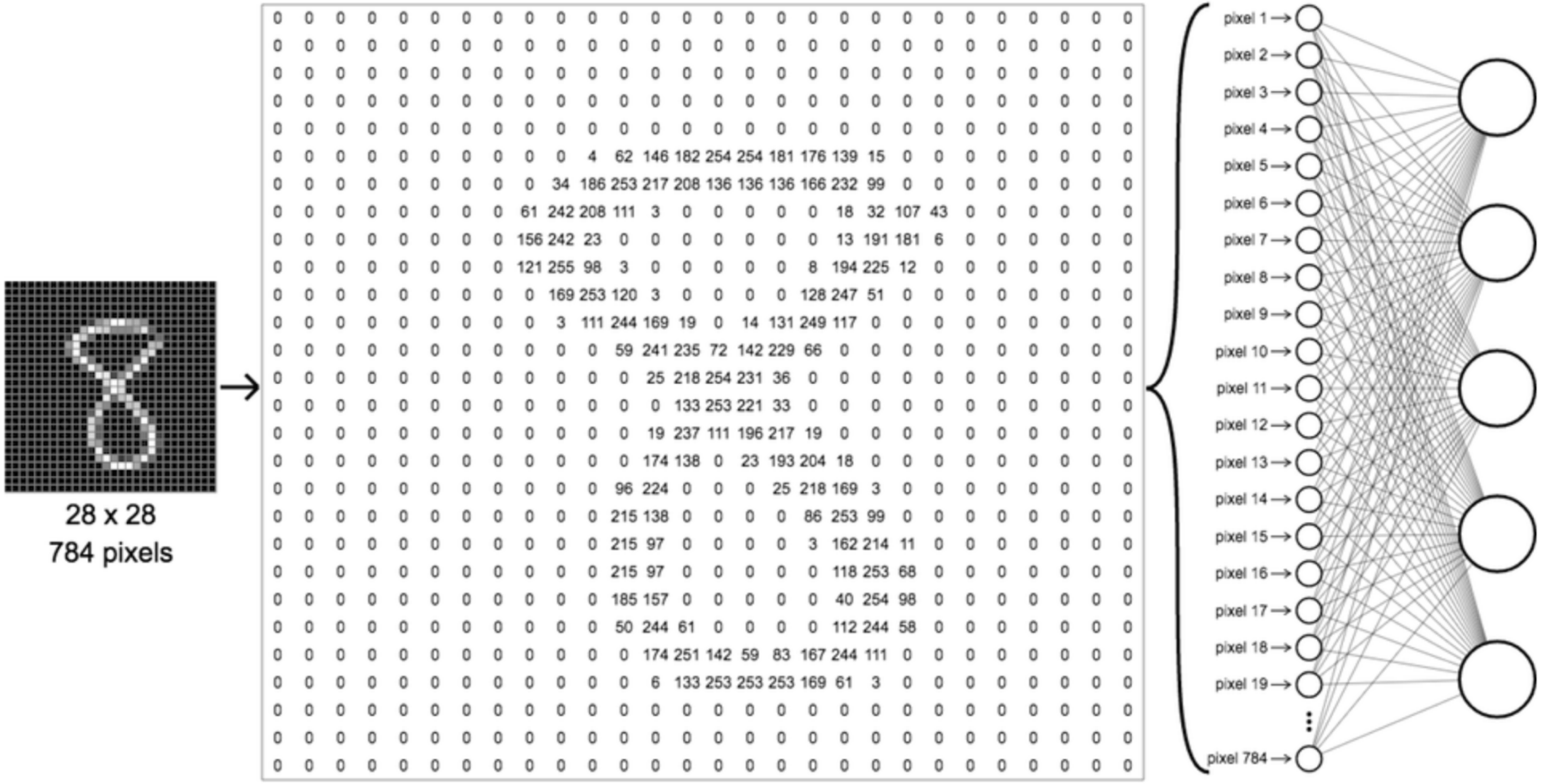
Dimension of image + gray values



1-layer neural network for MNIST. The 10 output neurons correspond to our classes, the 10 digits from 0 to 9.

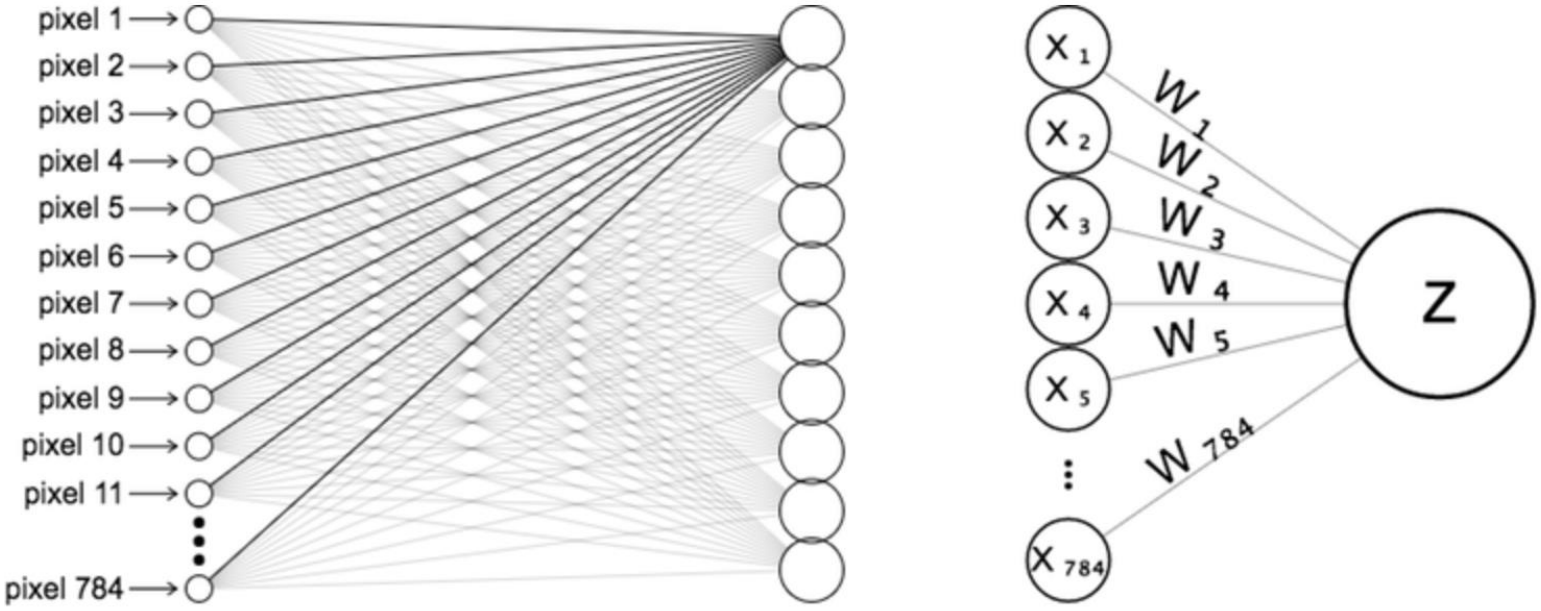
https://ml4a.github.io/ml4a/looking_inside_neural_nets/

Or back to 2 D



How to input an image into a neural network

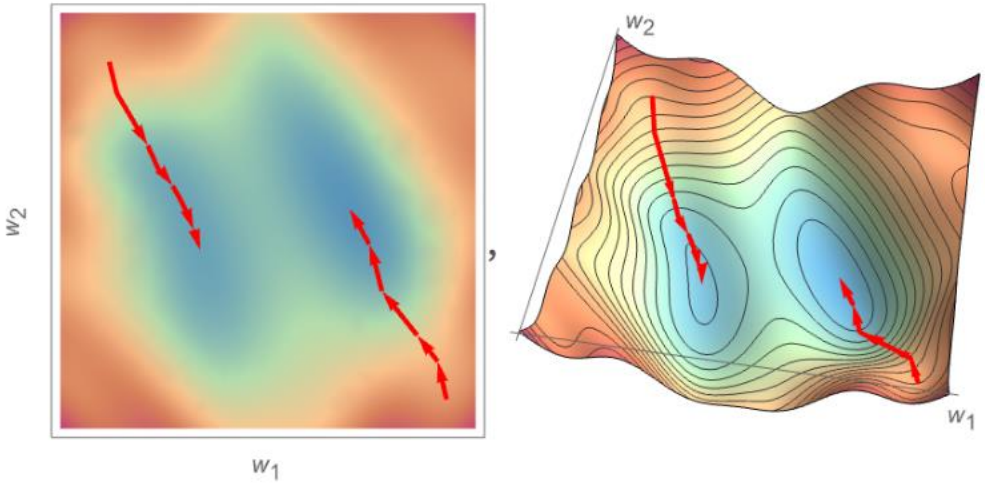
Which output fits to the correct numbers?

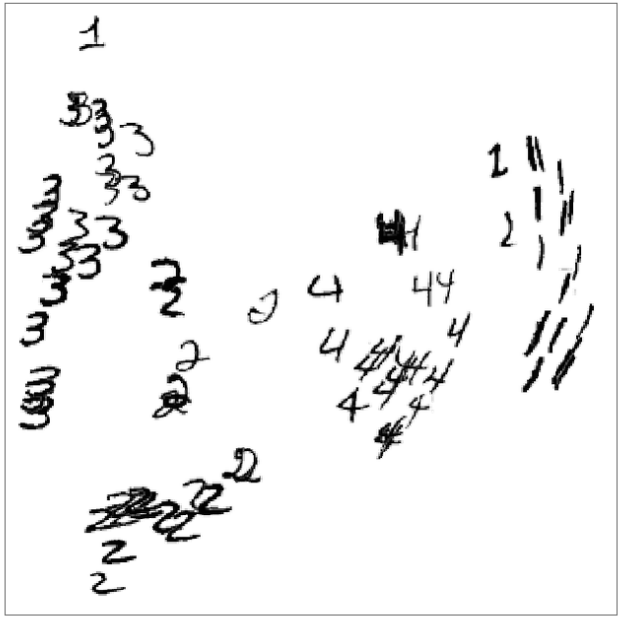


Highlighting the weights connections to a single output neuron

Learning

- Split the data into training and testing
- Train, iteratively, optimise a cost function
- Try to find the global minimum of the cost function
- Hope that the system converges





Demo: Looking inside neural nets [demo page] [view source]

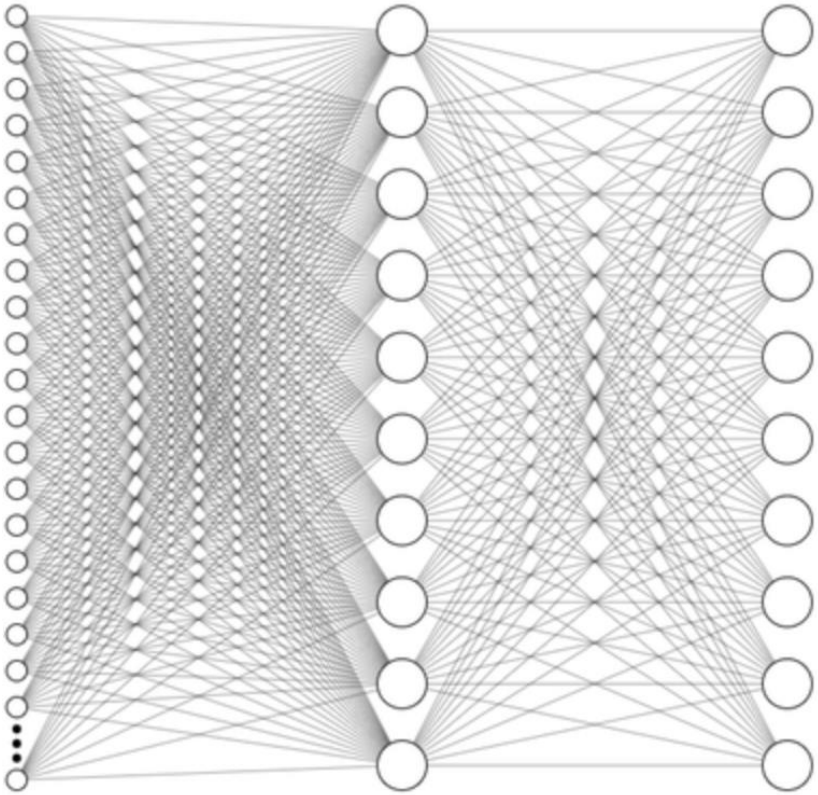
MNIST ordinary | View numbers | save

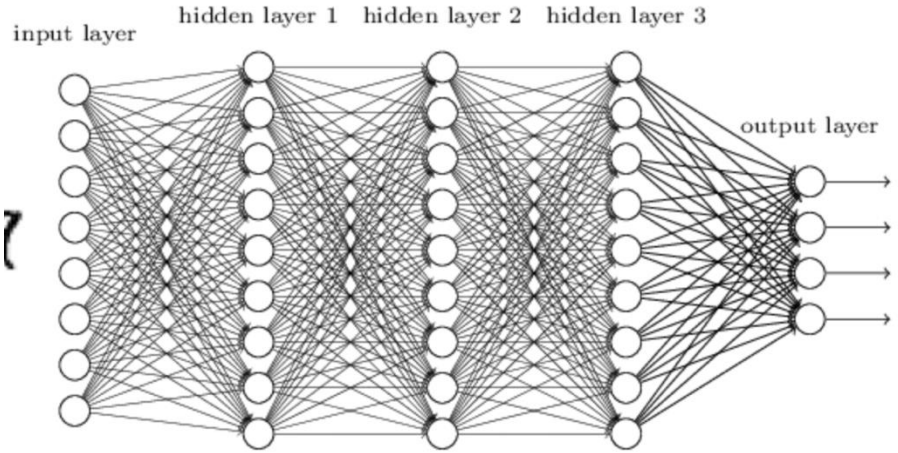
	predicted 0	predicted 1	predicted 2	predicted 3	predicted 4	predicted 5	predicted 6	predicted 7	predicted 8	predicted 9	recall
actual 0	954	0	0	7	1	10	6	3	7	3	96%
actual 1	0	1031	4	3	1	4	1	2	16	2	97%
actual 2	12	21	852	18	11	8	14	20	29	5	86%
actual 3	2	5	9	899	1	71	0	12	23	7	87%
actual 4	2	8	2	2	861	7	7	1	4	89	88%
actual 5	7	5	9	24	3	833	12	8	12	2	91%
actual 6	11	6	2	0	6	31	902	0	8	1	93%
actual 7	3	10	5	3	7	7	1	1041	0	14	95%
actual 8	6	4	3	29	2	31	1	9	882	21	87%
actual 9	7	3	1	7	10	11	1	44	4	873	91%
precision	95%	92%	96%	91%	95%	82%	95%	91%	90%	86%	accuracy 91%

3 misclassified as 9

How to improve???

Adding an internal layer!





99.4 % correctness

Why not do more internal layers?

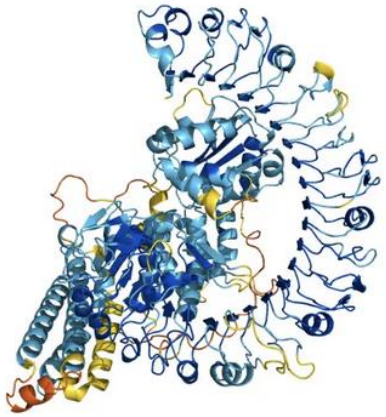
If doing this, you need to read about

- Sample size
- Overfitting: a model **learns patterns that are specific to the training data but don’t generalize**

Example of AI: Alpha fold v1

AlphaFold v1 (2018)

- Used **convolutional** and **recurrent (LSTM)** networks.
- Architecture was relatively shallow by today's standards: around **8–10 layers** in its main modules.
- Achieved impressive accuracy for its time but had limits on long-range residue interactions.



CASP (Critical Assessment of Structure Prediction)

Accuracy: It predicted **~25 out of 43 protein structures** *more accurately than any other method*.

Average GDT_TS: ≈ 58.9

Second-best method

- The **second-ranked group** at CASP13 was typically listed as **Zhang Lab** (University of Michigan, team "Zhang-Server").
- Their models were the best for about **3 to 5** targets (depending on subcategory).
- **Average GDT_TS: ≈ 49.6** (FM category)

AF version 2

- **AlphaFold v2 at CASP14 (2020)**
- **Competition:** CASP14 (2020)
- **Result:** Transformed the field entirely.
- **Average GDT_TS: 92.4** (out of 100) across all categories — *essentially experimental accuracy*.

So here we can see that AI – deep neural networks, do ”something correct”

AlphaFold v2 (2021)

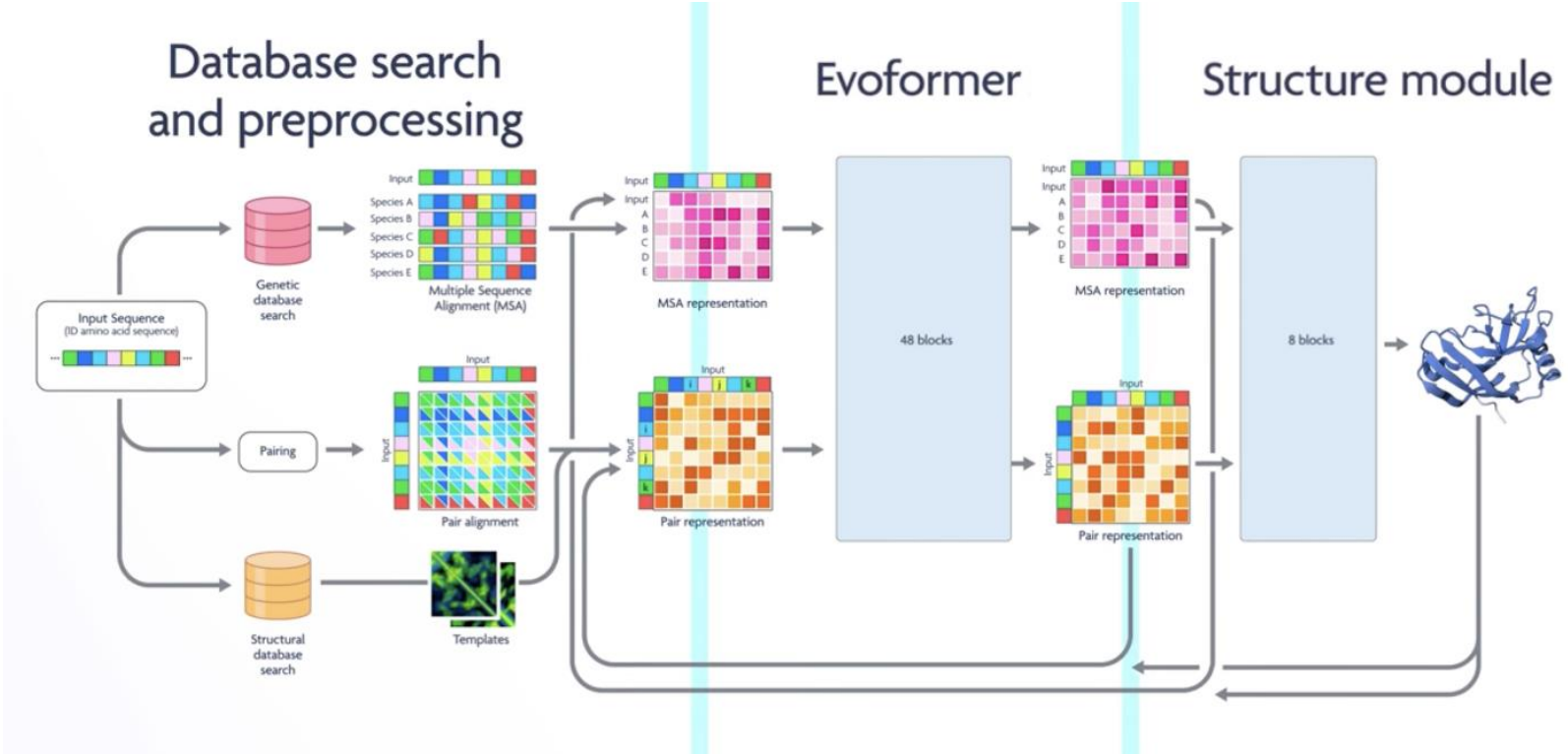
- **Completely redesigned** with transformer-style modules.
- Uses two key components:

  Evoformer – a deep stack of *transformer blocks* processing sequence and pairwise residue information.

  Structure Module – a geometric deep-learning component that predicts 3D atomic coordinates.

- Each Evoformer contains about **48 transformer blocks**, and each block includes **multiple attention layers** (so, effectively hundreds of sublayers when expanded).
- So depending on how you count, AlphaFold2 has on the order of **hundreds of neural layers**, far deeper than a standard MLP.

The trick is the transformer and putting things one after the other



- Transformers let the model *learn what to pay attention to*.

Attention Is All You Need

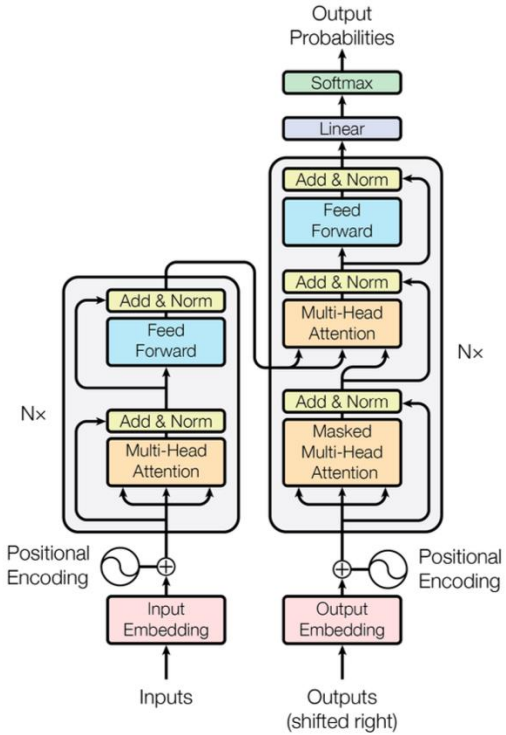


Figure 1: The Transformer - model architecture.

I still need to read this paper!

What is the attention part exactly?

- “The cat sat on the mat because it was soft.”
- The It refers to the mat – that is not easy to get the of neural network
- For each token (word) it computes three vectors: **Query (Q)**, **Key (K)**, and **Value (V)**.

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

How does ChatGPT works?

- https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

- Just decided “randomly” what the next word will be

a ranked list of words that might follow, together with “probabilities”:

The best thing about AI is its ability to

learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%

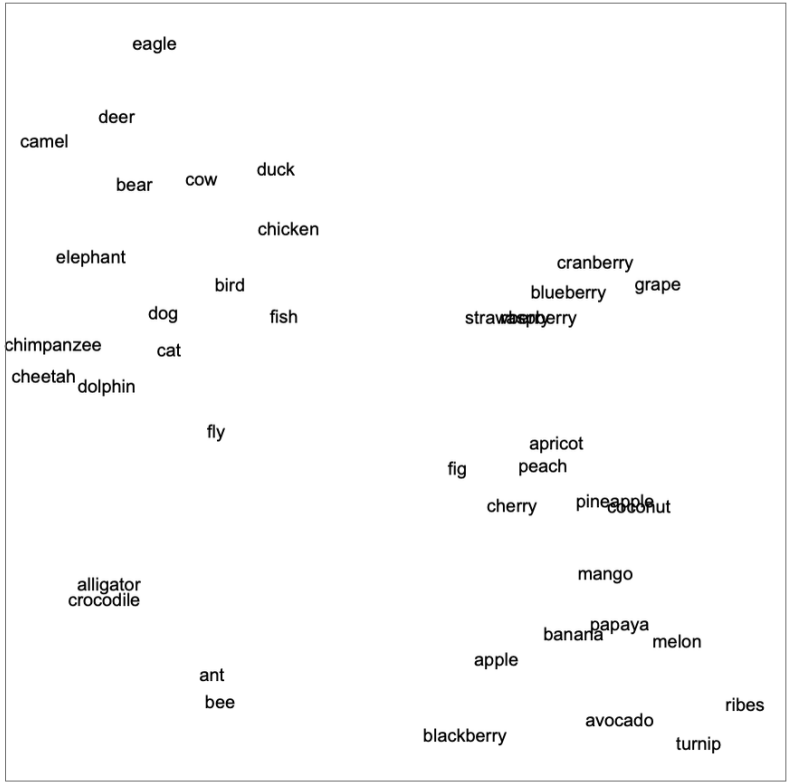
- Probability is learn from billions of texts of the internet
- There is a random value, which word to take in the list… so every answer is the same

The Concept of Embeddings

- Words get transformed into numbers
- if we project down to 2D

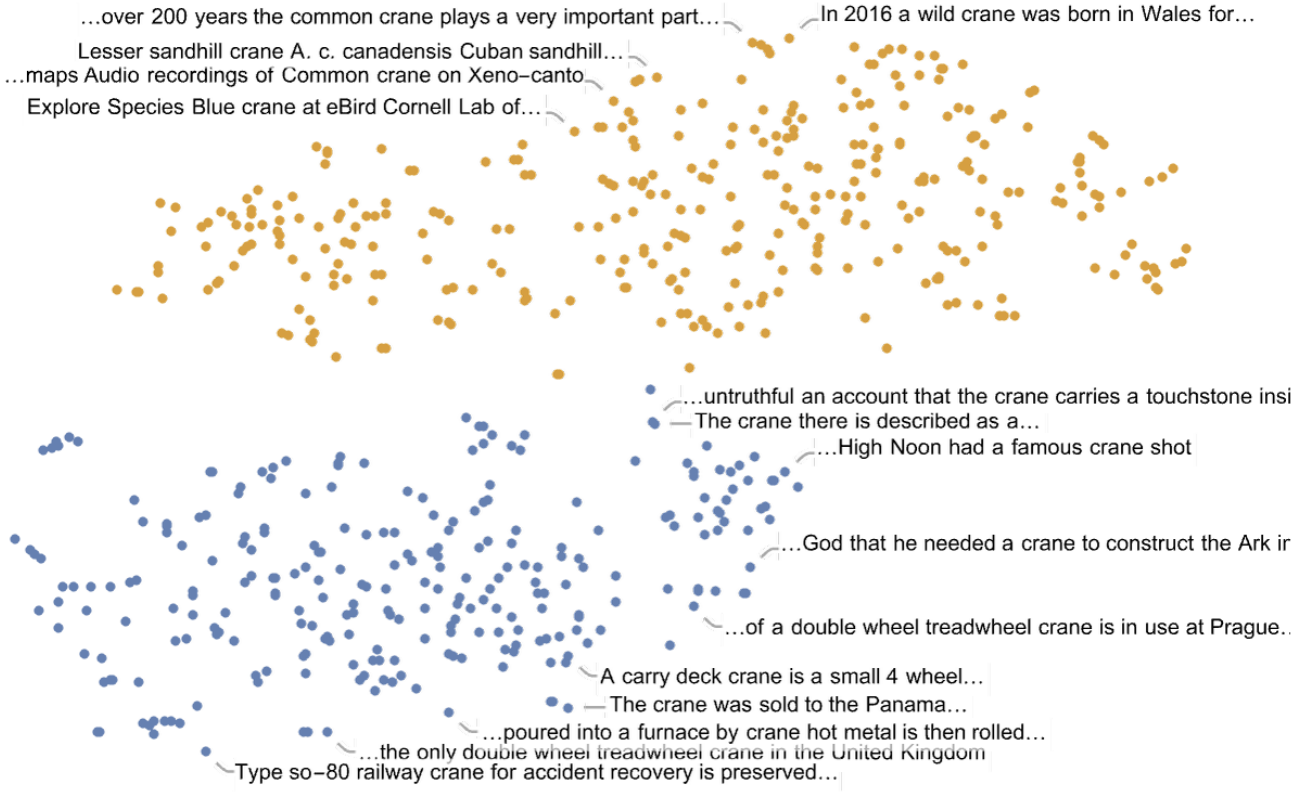
- Train on 5 billion words from the web and some words like crocodile and alligator will often appear almost interchangeably in otherwise similar sentences

- 175 billion neural net weights

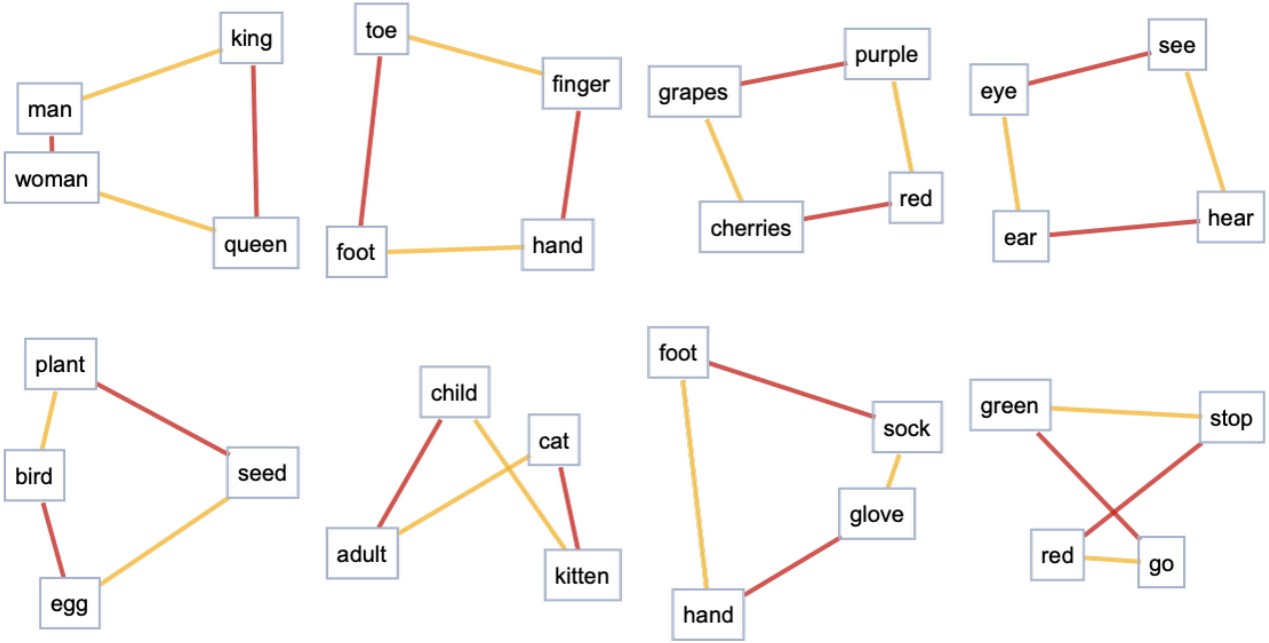


https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

Store more concepts



Is there for example some kind of notion of “parallel transport” that would reflect “flatness” in the space? One way to get a handle on that is to look at analogies:



And, yes, even when we project down to 2D, there’s often at least a “hint of flatness”, though it’s certainly not universally seen.

But why am I telling you that?

- Could we train a LLM based on scRNA-Seq data, that ”knows” which gene to express next?

Article

Transfer learning enables predictions in network biology

https://doi.org/10.1038/s41586-023-06139-9

Received: 29 March 2022

Accepted: 27 April 2023

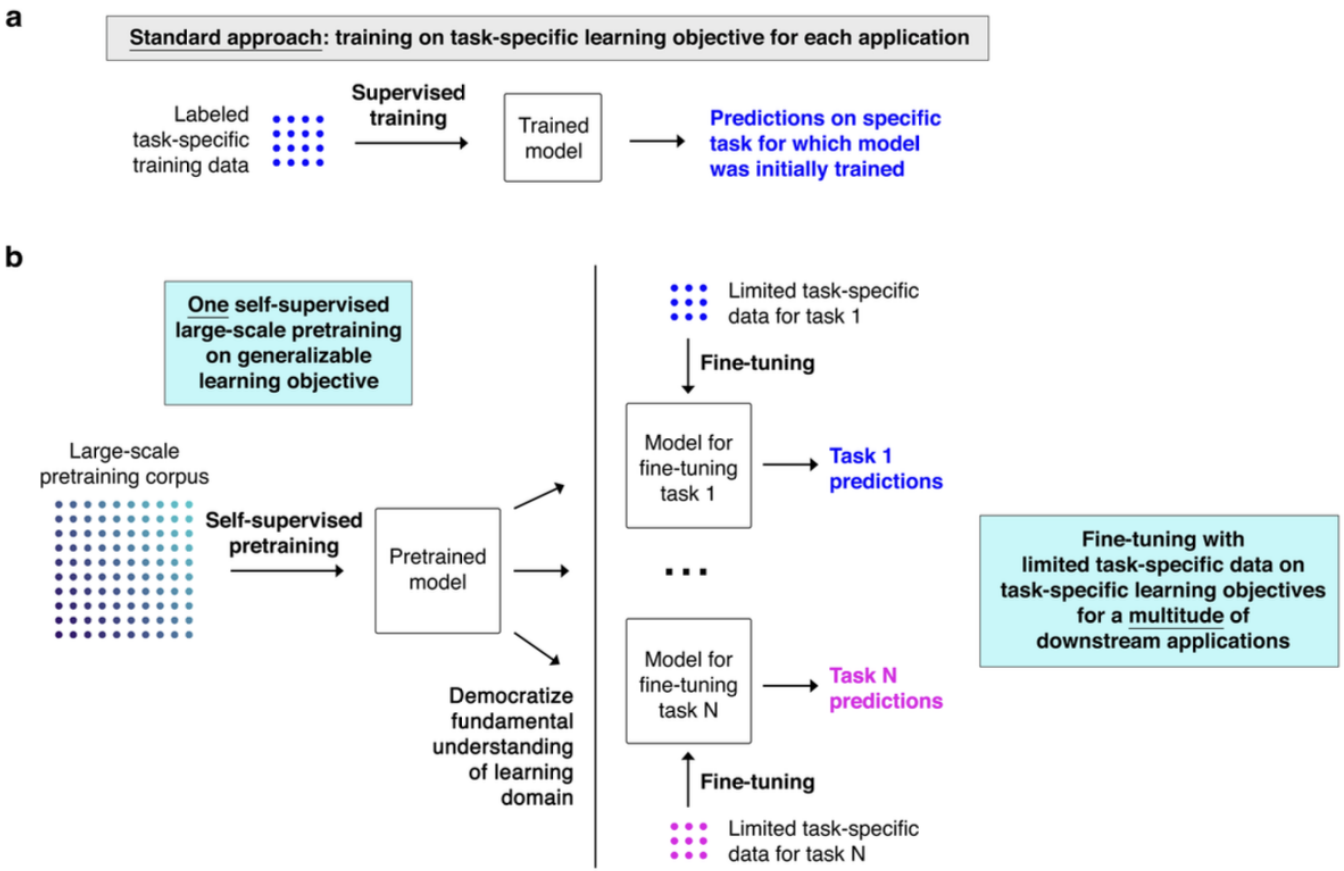
Published online: 31 May 2023

Check for updates

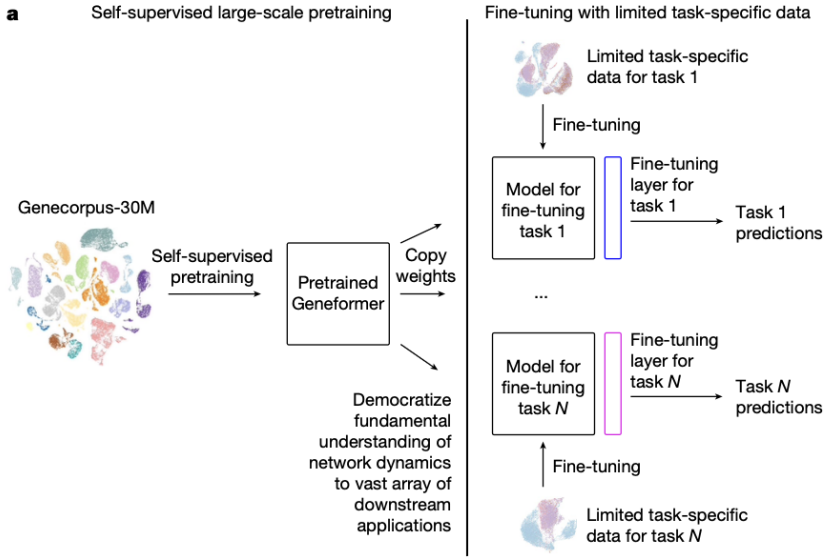
Christina V. Theodoris[1,2,3,4]✉, Ling Xiao[2,5], Anant Chopra[6], Mark D. Chaffin[2], Zeina R. Al Sayed[2], Matthew C. Hill[2,5], Helene Mantineo[2,5], Elizabeth M. Brydon[6], Zexian Zeng[1,7], X. Shirley Liu[1,7,8] & Patrick T. Ellinor[2,5]✉

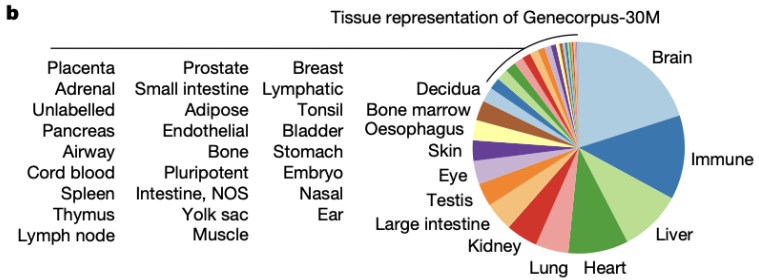
Mapping gene networks requires large amounts of transcriptomic data to learn the connections between genes, which impedes discoveries in settings with limited data,

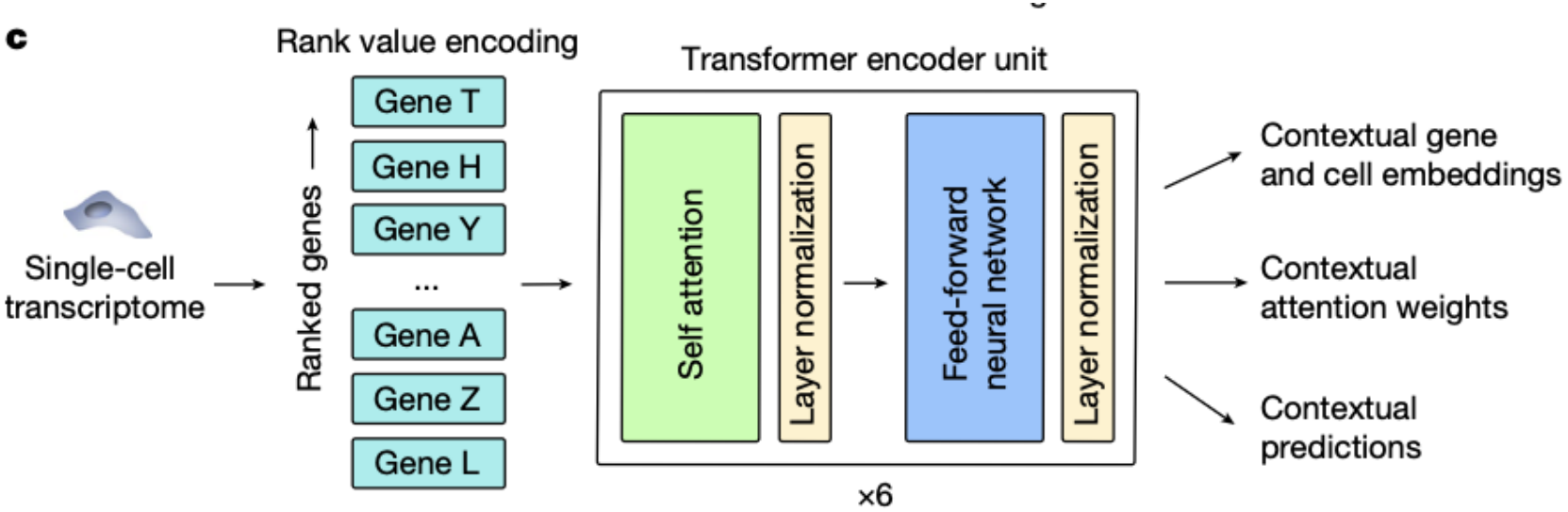
GeneFormer - the model



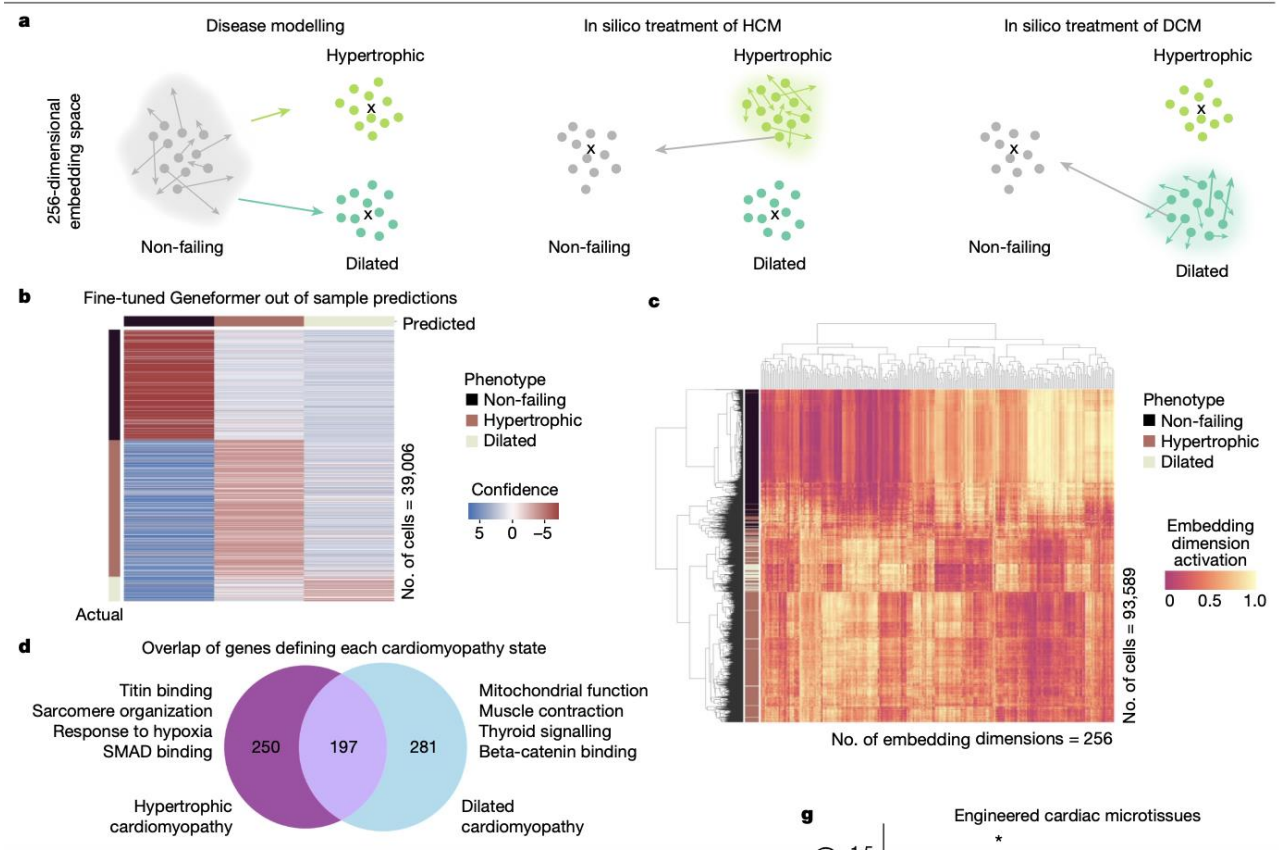
The training data






c
Rank value encoding
Transformer encoder unit
Single-cell transcriptome
Ranked genes
Gene T
Gene H
Gene Y
...
Gene A
Gene Z
Gene L
Self attention
Layer normalization
Feed-forward neural network
Layer normalization
×6
Contextual gene and cell embeddings
Contextual attention weights
Contextual predictions

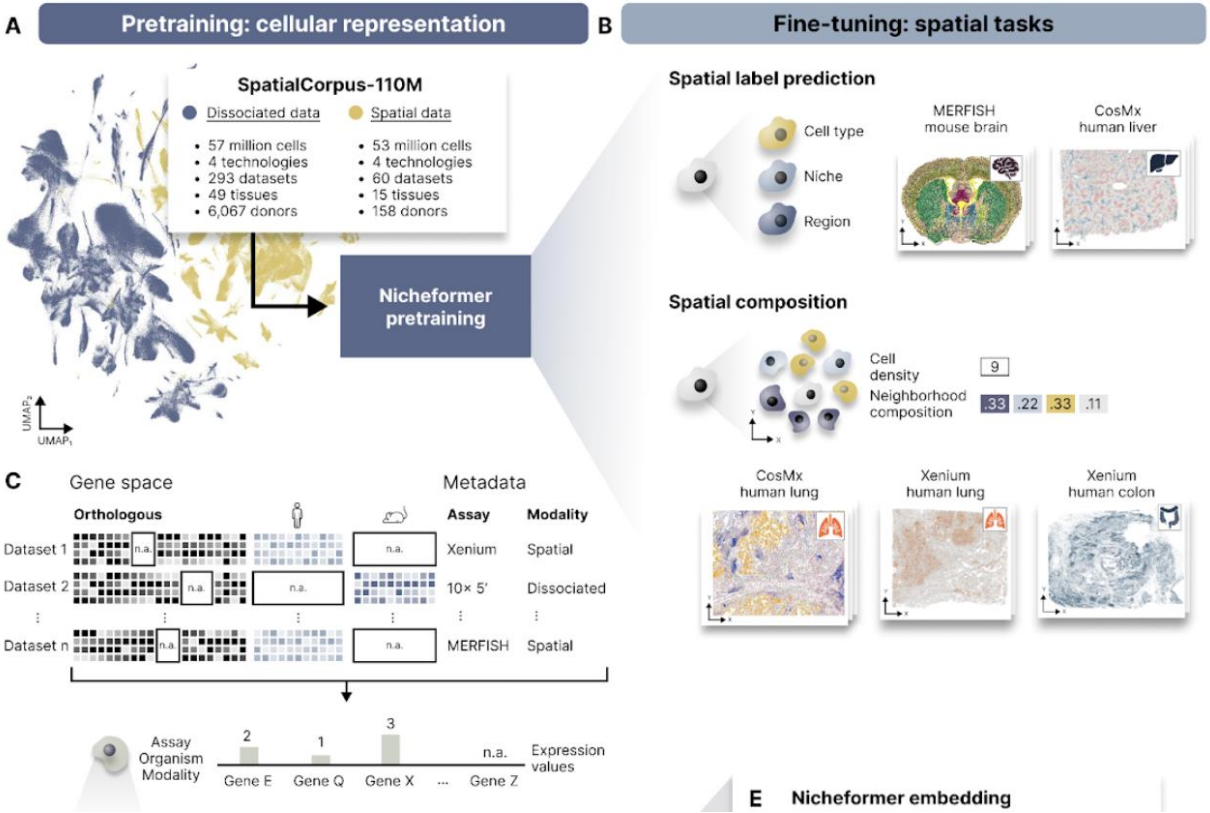
In silico perturbation

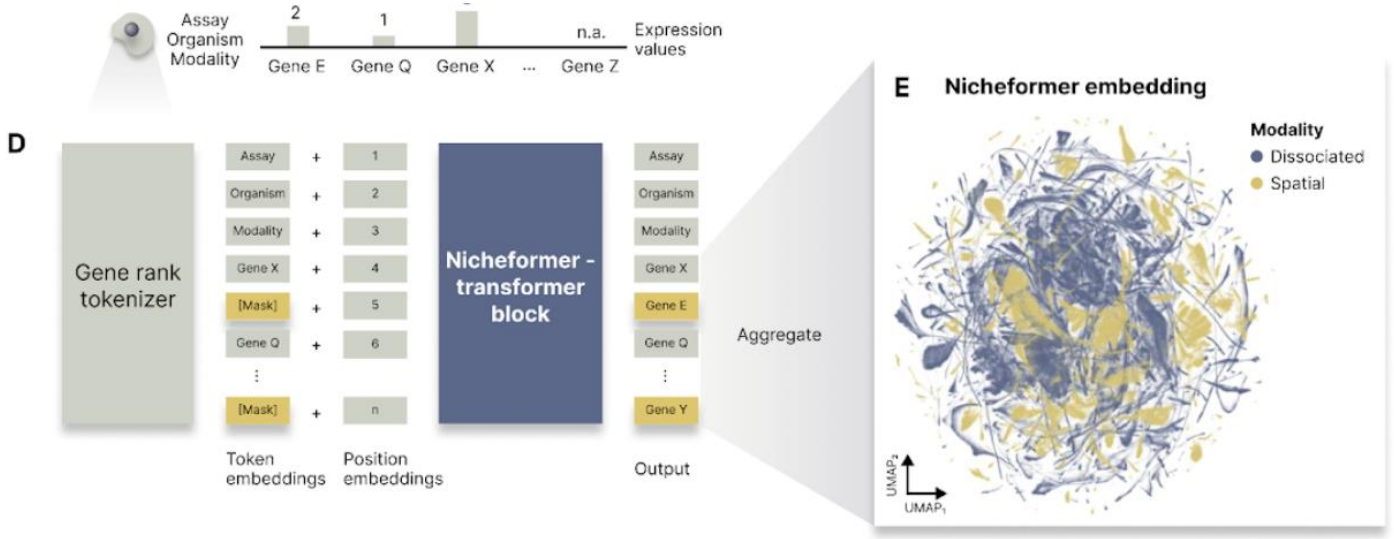


Now, think you have your own data, what could you do?

More datasets to come

- Niechformer – Schaar et al, Theis lab, BioRxiv 20205 - From 90 million to 300 cells?





129

130 **Figure 1 | Nicheformer, a foundation model for spatial transcriptomics.** A) Nicheformer is pretrained on the

Again, how could you used that?

Thank you for inviting me to Montevideo

125 Years BNITM 1900-2025
BERNHARD NOCHT INSTITUTE FOR TROPICAL MEDICINE