

DEEP LEARNING

WITH NEW MATTH

PHILIPP GÄBLER

2022-06-28

DEEP LEARNING

WITH ~~NO~~ MAT^H
AS LITTLE AS POSSIBLE

PHILIPP GÄBLER

2022-06-28

THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG
PILE OF LINEAR ALGEBRA, THEN COLLECT
THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL
THEY START LOOKING RIGHT.



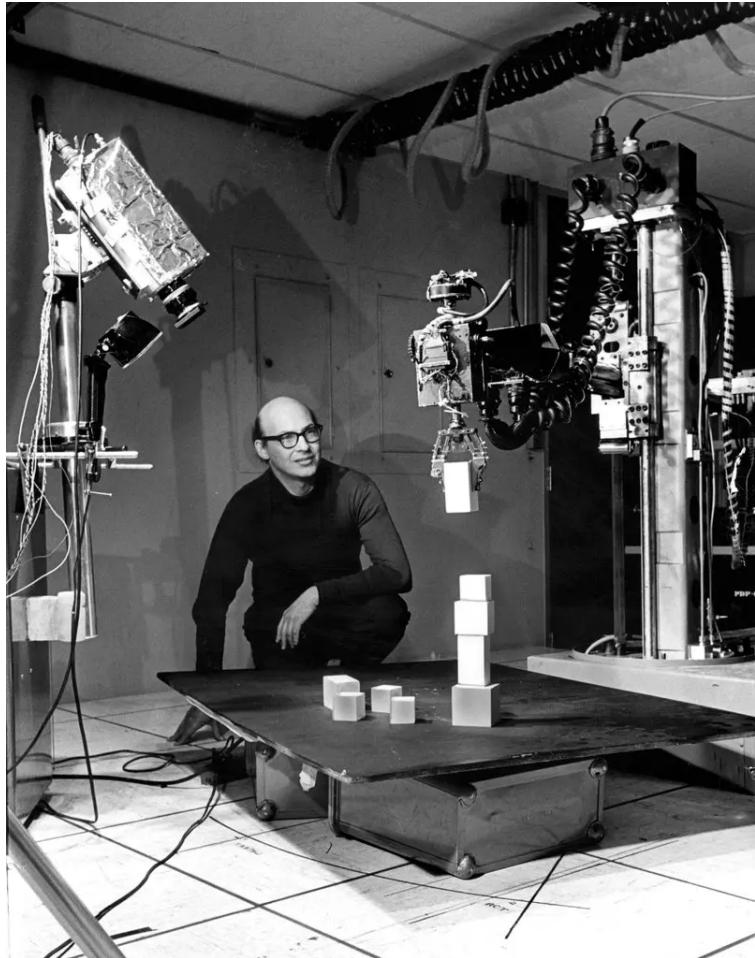
<https://xkcd.com/1838/>

OUTLINE

- ① A BIT OF HISTORY
- ② WHAT PROBLEM(S) DOES DL SOLVE?
 - COMPUTATIONAL
- ③ WHAT DO SOLUTIONS LOOK LIKE?
 - ALGORITHMIC
 - IMPLEMENTATIONAL
- ④ HOW ARE SOLUTIONS IMPLEMENTED?
- ⑤ SOME LANGUAGE-RELATED USE CASES

1

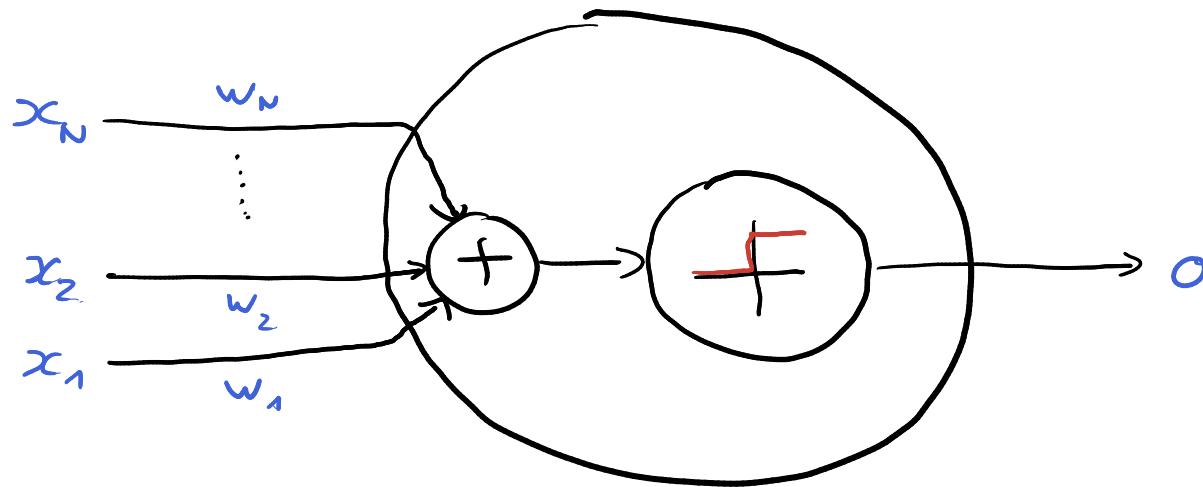
A BIT OF
HISTORY



<https://www.nytimes.com/2016/01/26/business/marvin-minsky-pioneer-in-artificial-intelligence-dies-at-88.html>

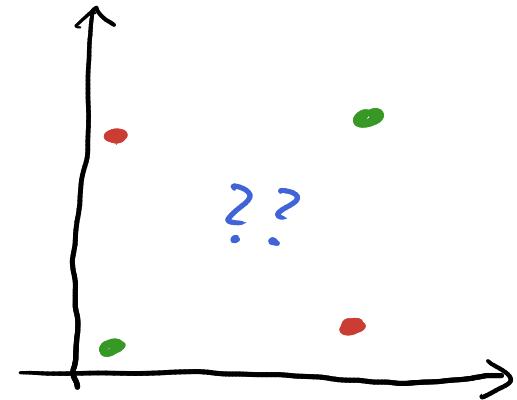
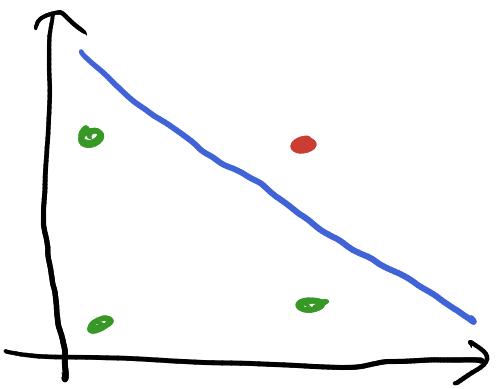
EVĀPĀXĀY ĶĀV TĀ PERCEPTRON

- McCULLOCH & PITTS, 1948 : NEURON MODEL



- ROSENBLATT, 1958 : ENTHUSIASM
- MINSKY & PAPERT, 1969 : SHOCK OF "PERCEPTRONS"

AI WINTER



PROBLEM OF LINEAR SEPARABILITY

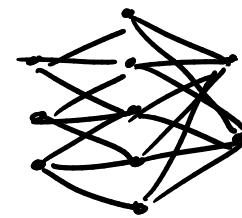
NEW SPRING

- RUMELHART & McCLELLAND (1986): PDP

- ↳ CONNECTIONISM

- ↳ BACKPROPAGATION

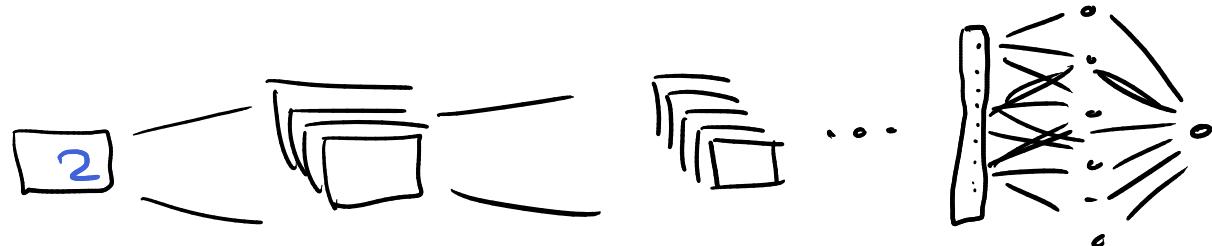
- ↳ SINGLE-LAYER → MULTI-LAYER



- STATISTICAL LEARNING / MACHINE LEARNING
OVER (SYMBOLIC) AI

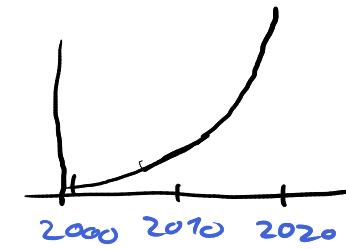
TOWARDS DEEP LEARNING

- MORE LAYERS

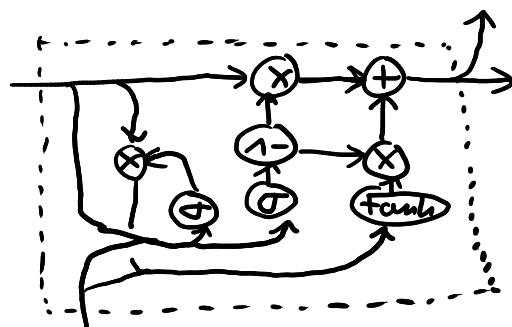


- MORE DATA

- MORE COMPUTE (MOORE'S LAW)



- BETTER TUNING & TRAINING



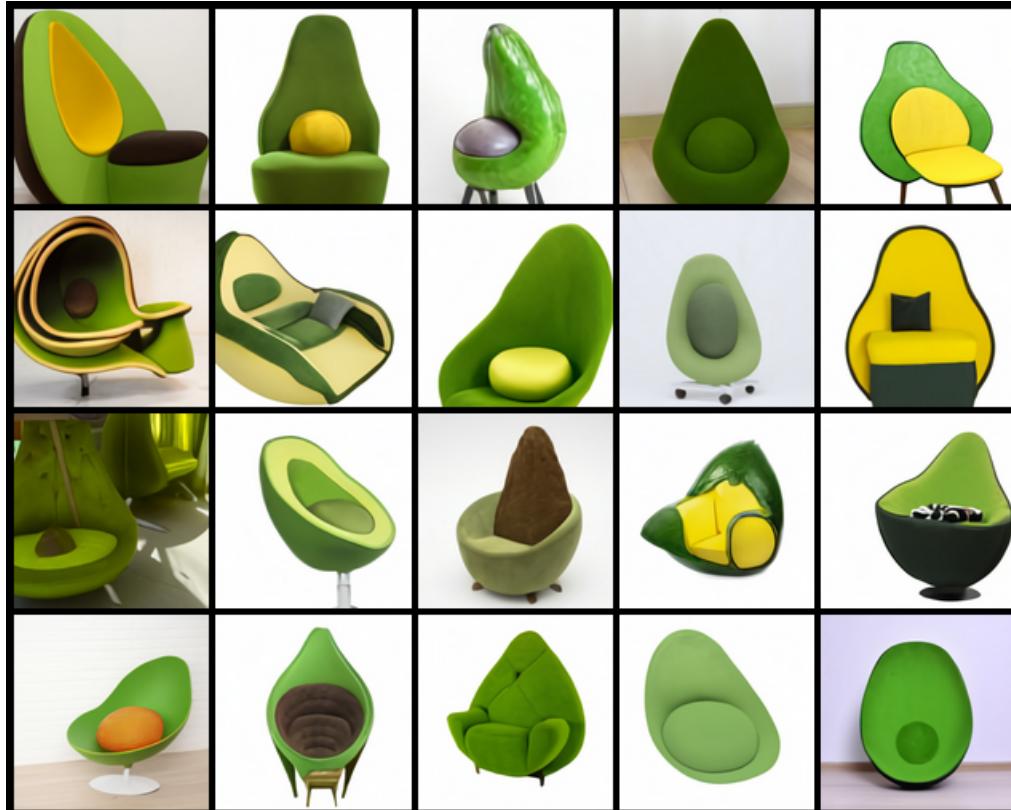
TAKE-AWAYS 1

- ARTIFICIAL - NO NEURONS ANYMORE!
- AI AND ML HAVE SPLIT - DL ≠ INTELLIGENCE
- GROWTH OF TECHNOLOGY IMPORTANT
- DL IS SUB SYMBOLIC (\approx CONNECTIONISM)

②

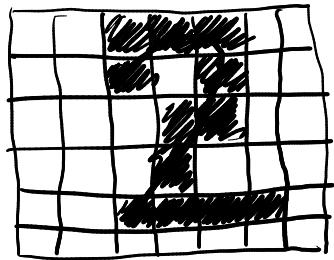
WHAT PROBLEMS DOES
DEEP LEARNING
SOLVE?

"AN ARMCHAIR IN THE SHAPE OF AN AVOCADO"

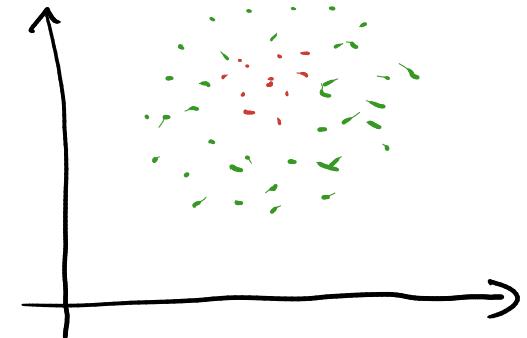


CHARACTERISTICS OF DEEP LEARNING

HIGH-DIMENSIONAL

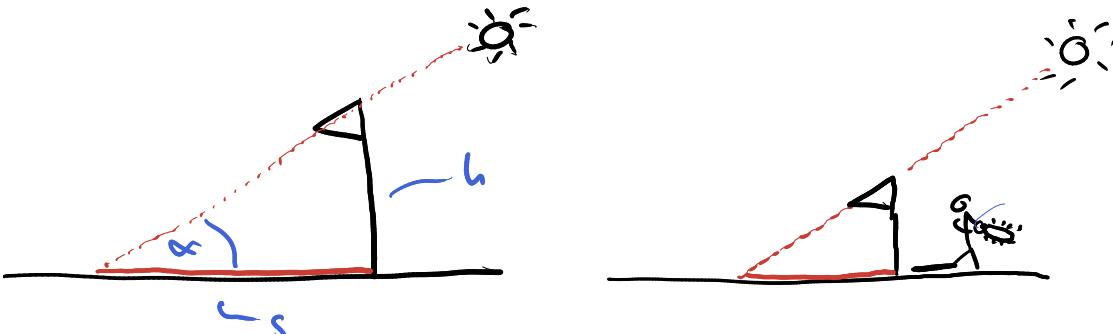


NON-LINEAR

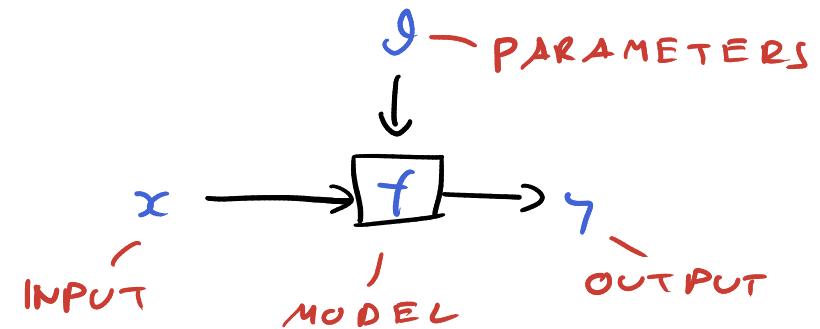


PREDICTION

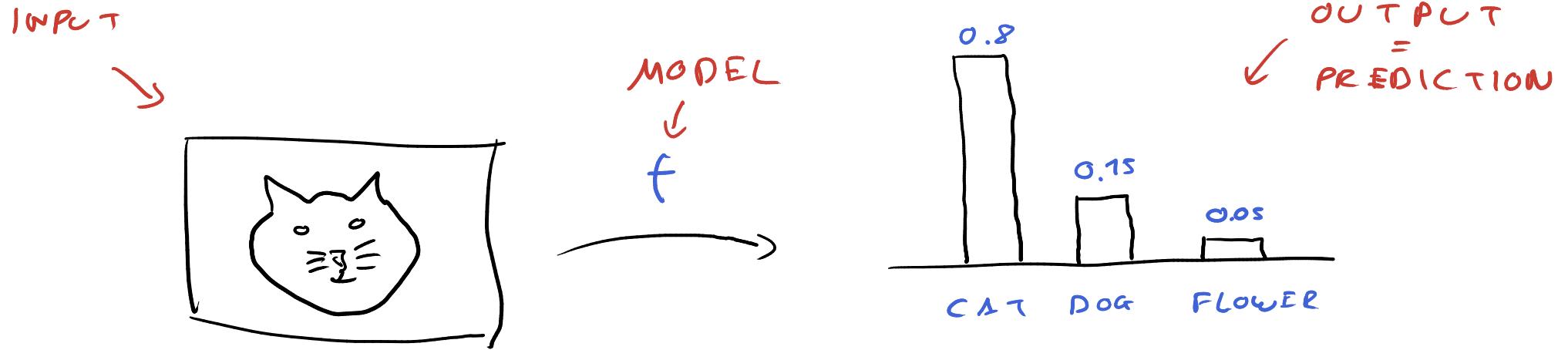
ASSOCIATIVE



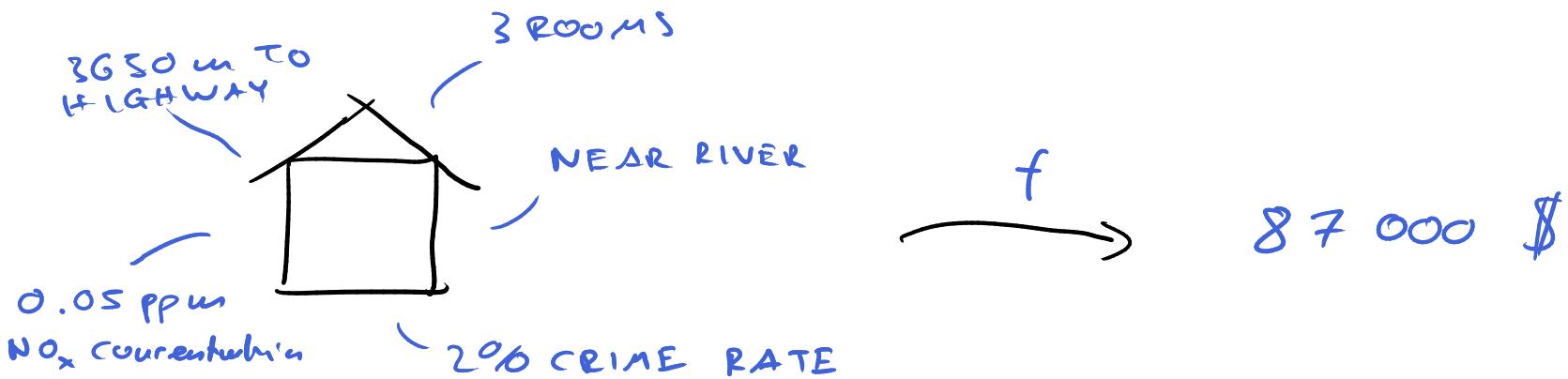
END-TO-END TRAINING,
PARAMETRIC



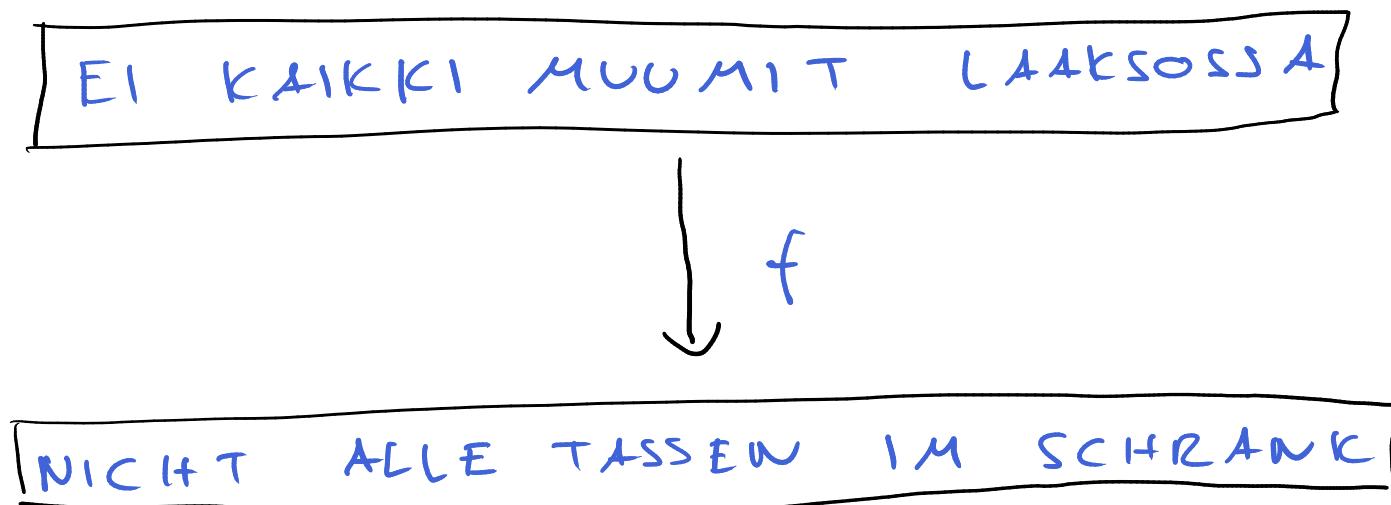
CLASSIFICATION



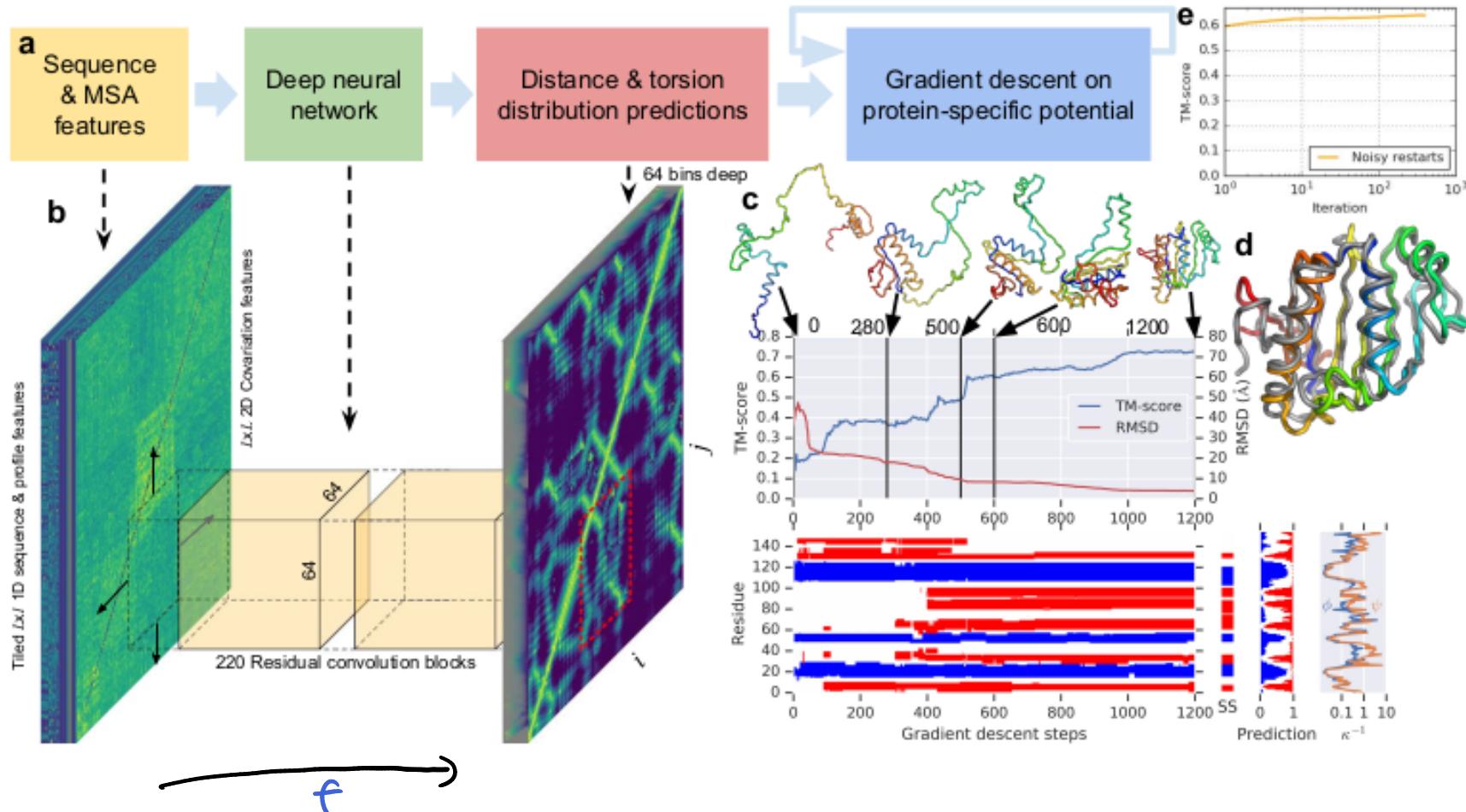
REGRESSION



SEQUENCE PREDICTION



HIGH-DIMENSIONAL MODELLING

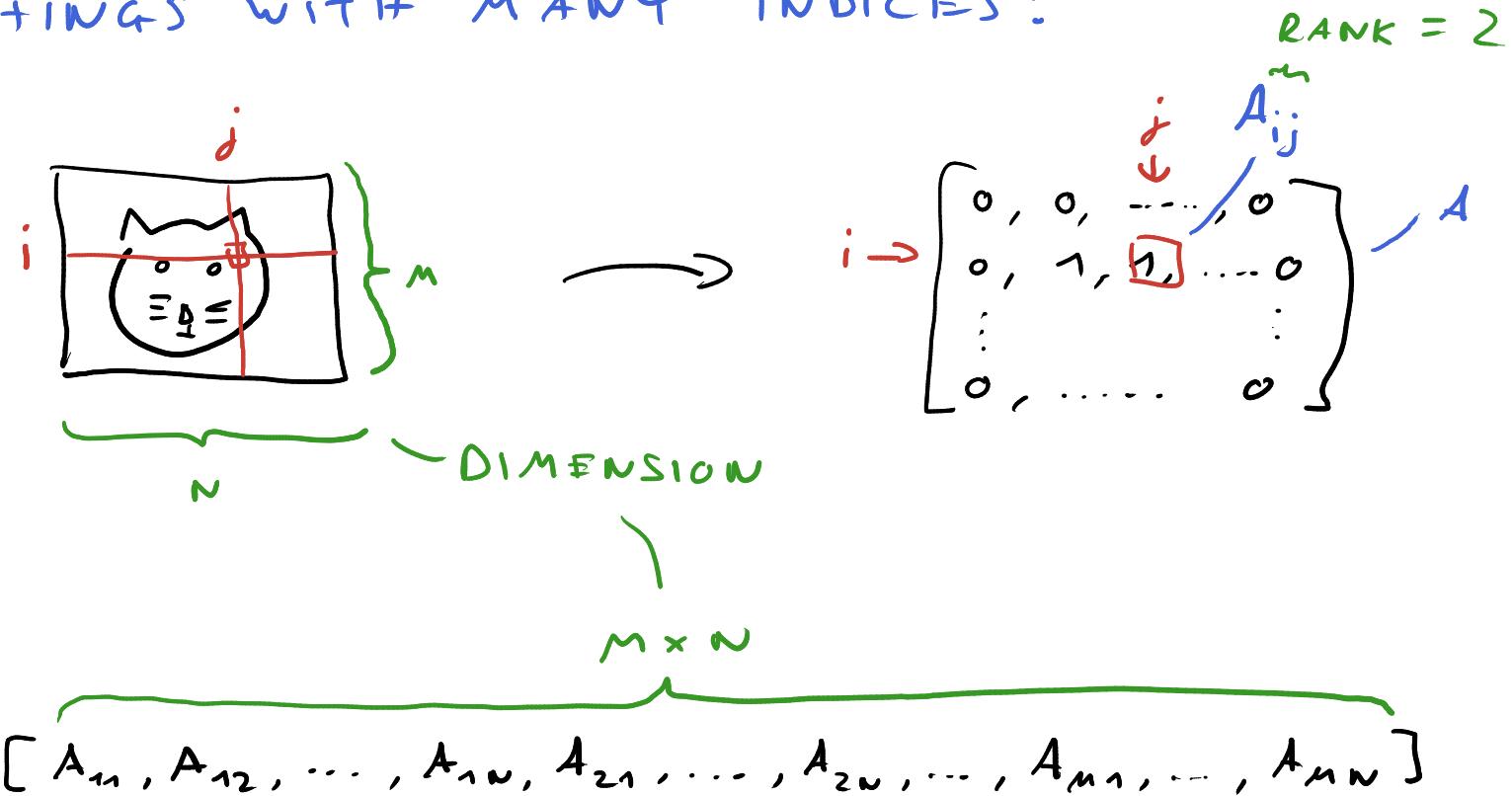


<https://doi.org/10.1038/s41586-019-1923-7>



TensorFlow

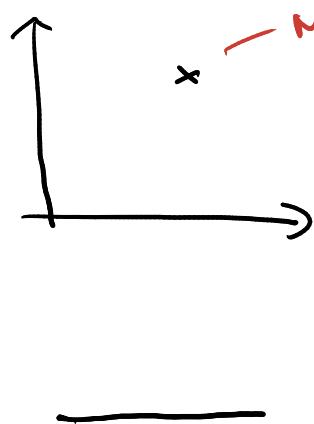
JUST THINGS WITH MANY INDICES:



EVERYTHING IS LINEAR ALGEBRA!

TENSOR SHAPES

RANK 1



N-DIMENSIONAL POINT

RANK 2



B/W

RANK 3



RGB

RANK 4



[2.0, -0.3]

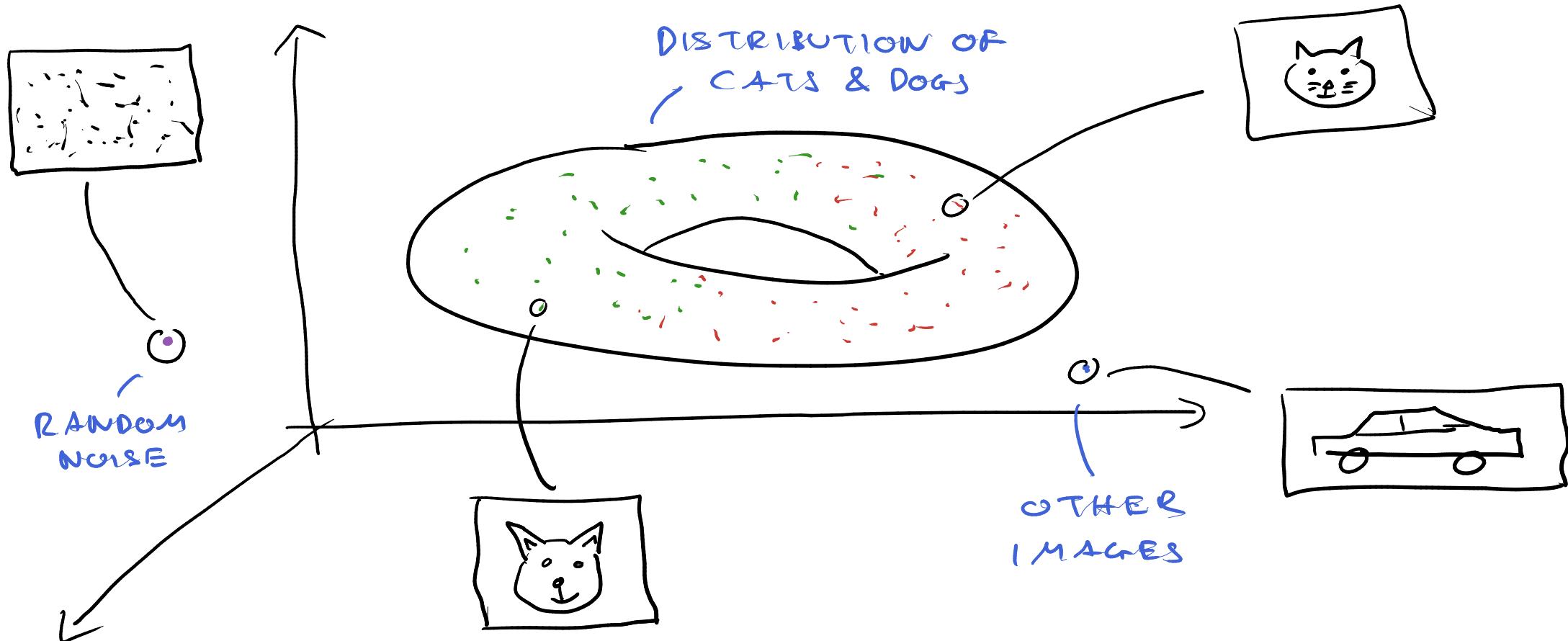
$$\begin{bmatrix} 0, 0, \dots, 0 \\ 0, 1, 1, \dots, 0 \\ \vdots \\ 0, \dots, 0 \end{bmatrix}_x$$

$$\begin{bmatrix} 0, 1, \dots, 0 \\ 0, 2, \dots, 0 \\ 0, 0, \dots, 0 \\ 0, 1, 1, \dots, 0 \\ \vdots \\ 0, \dots, 0 \end{bmatrix}_r^B$$

$$\begin{bmatrix} \dots \\ \dots \\ \dots \end{bmatrix}_c^B$$

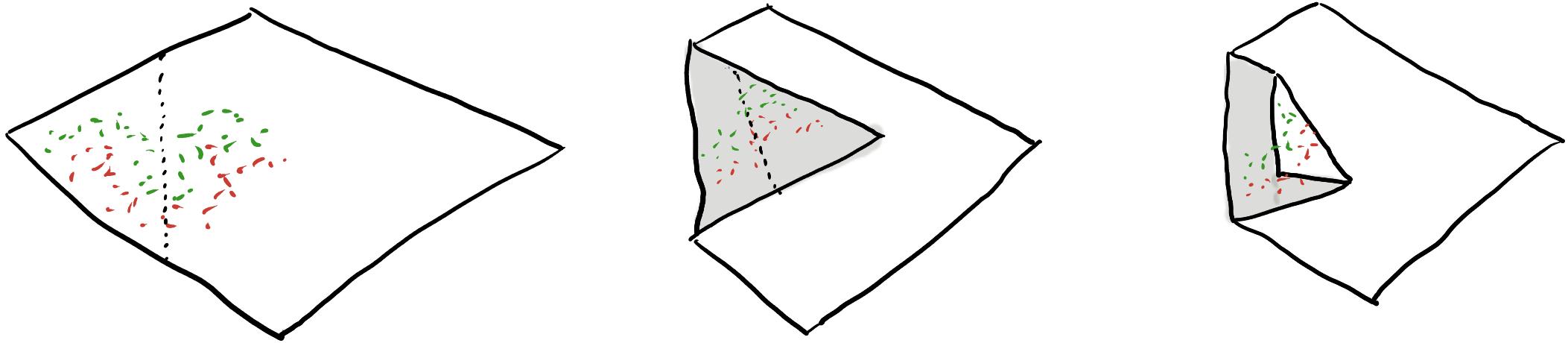
1
2
3
 $c \in B$

MANIFOLD HYPOTHESIS

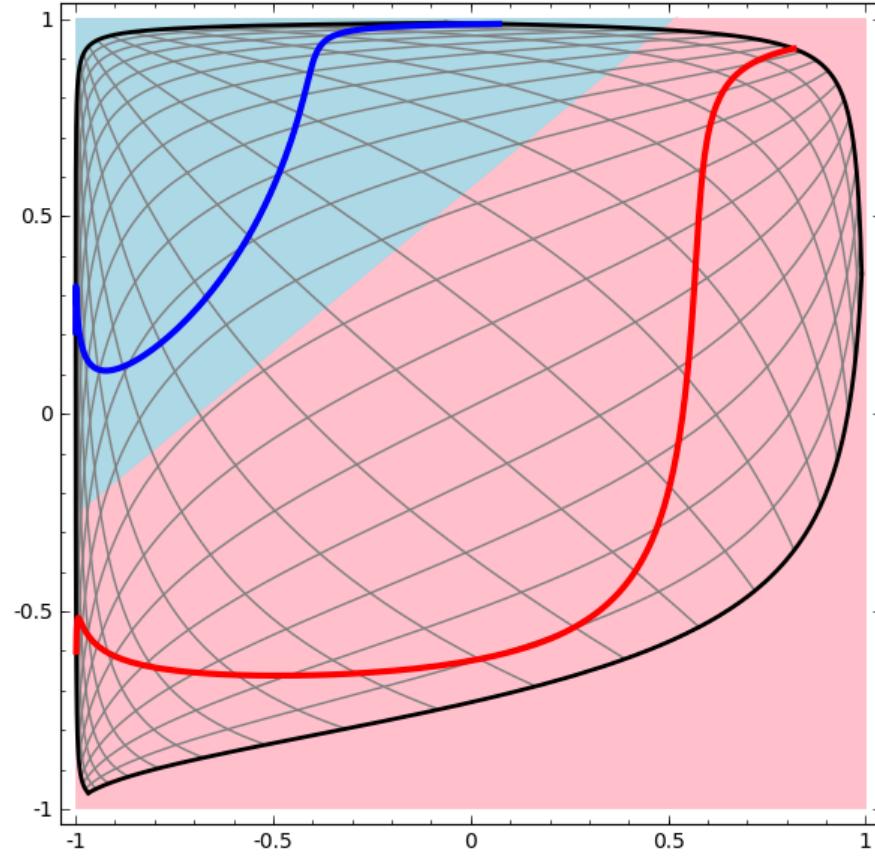
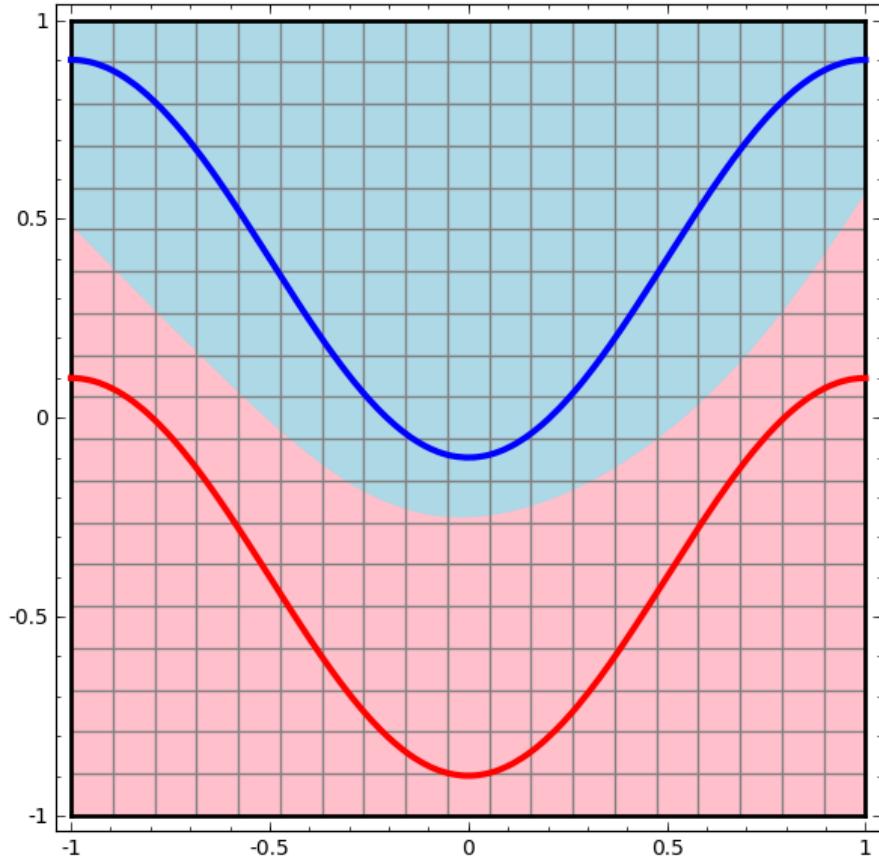


MNIST IN 2D

SPACE FOLDING

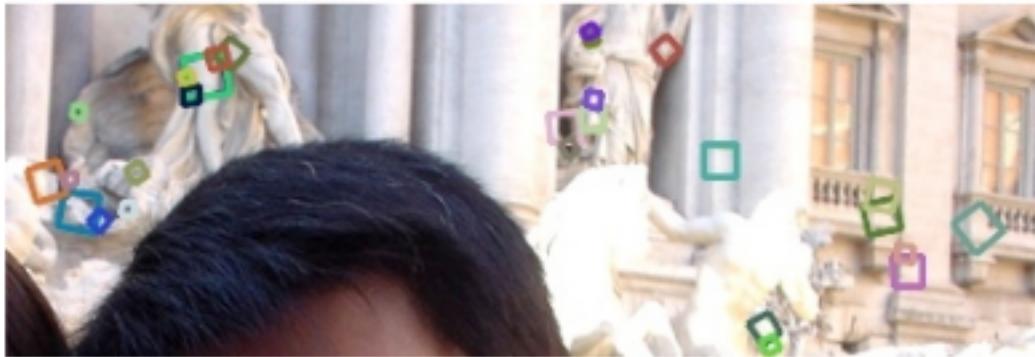


NONLINEAR TRANSFORMATIONS

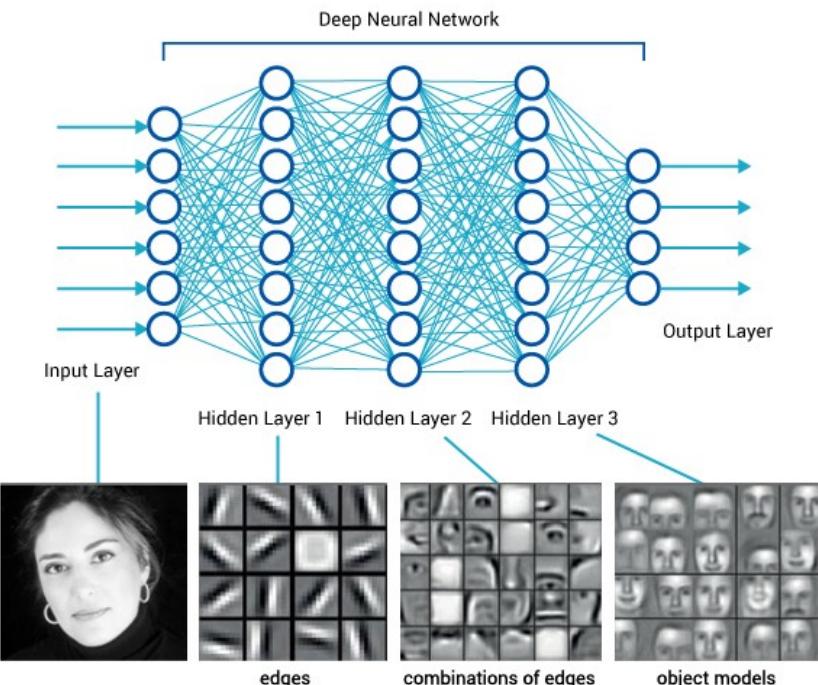


FROM FEATURES TO LAYERS

TRADITIONAL



DEEP LEARNING



③

WHAT DO SOLUTIONS
LOOK LIKE?

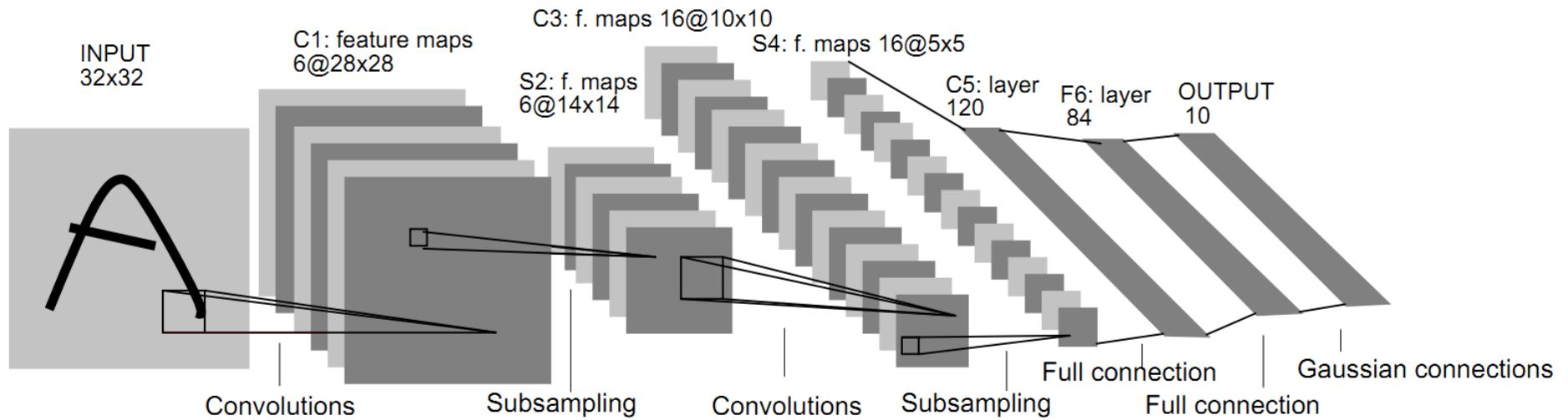
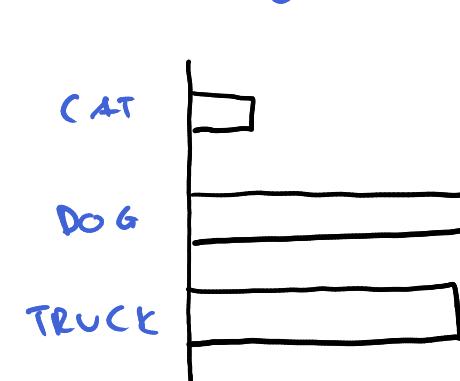
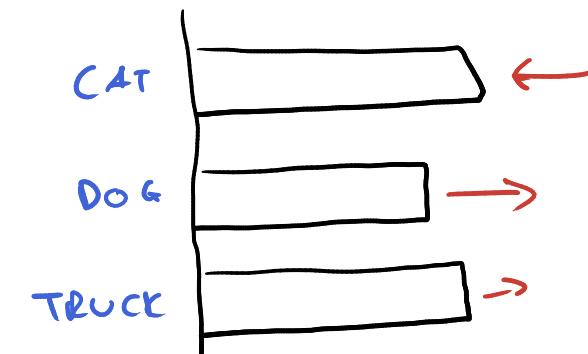
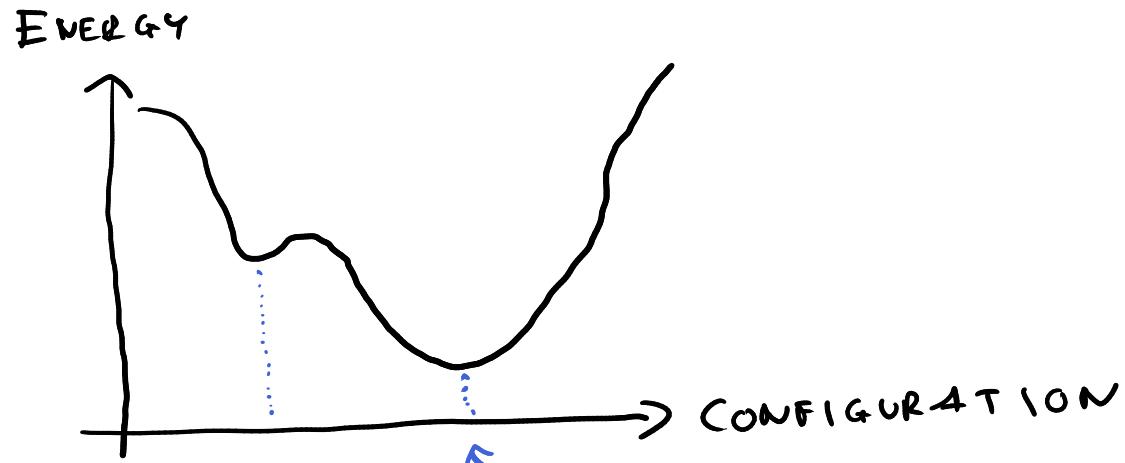
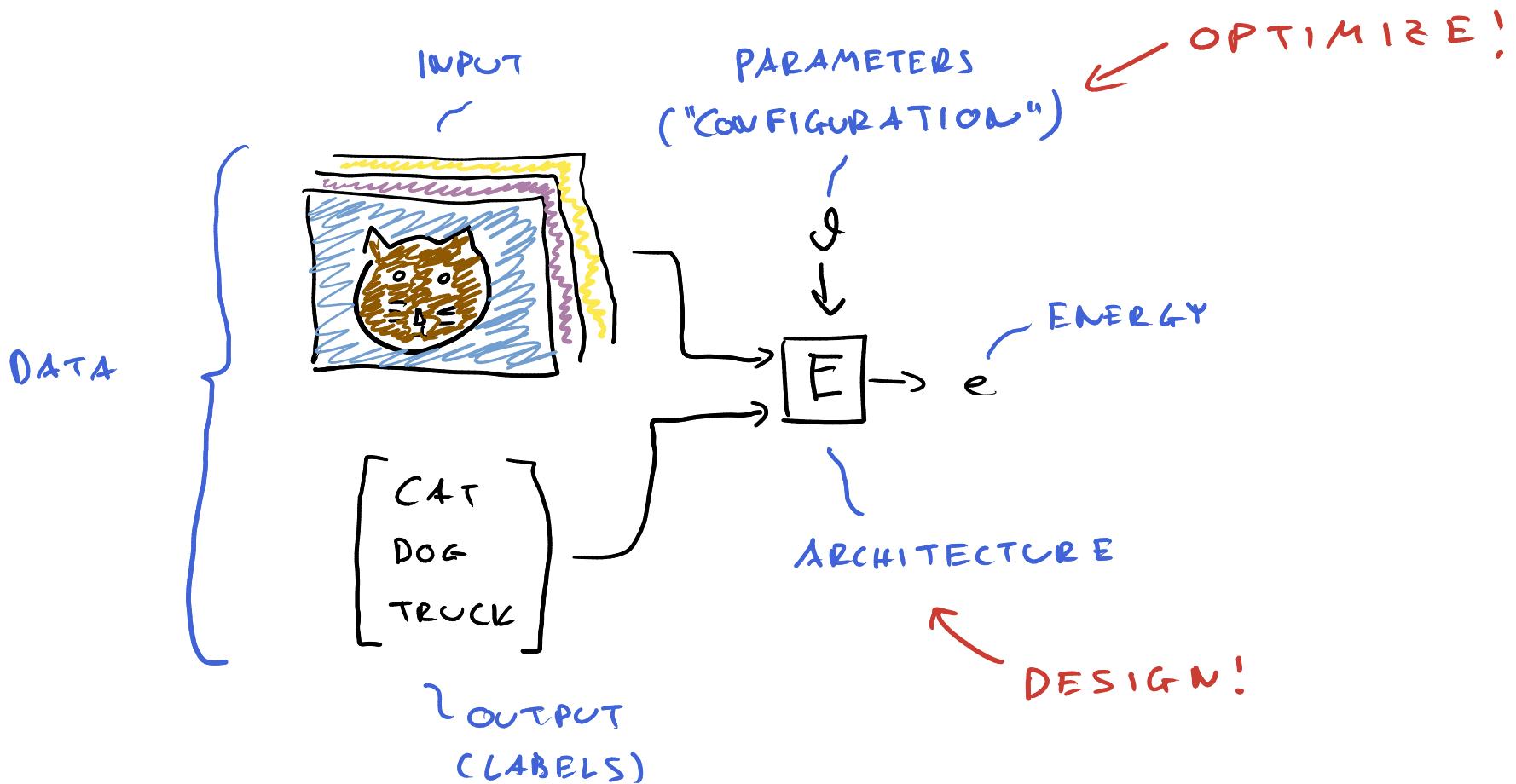


Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

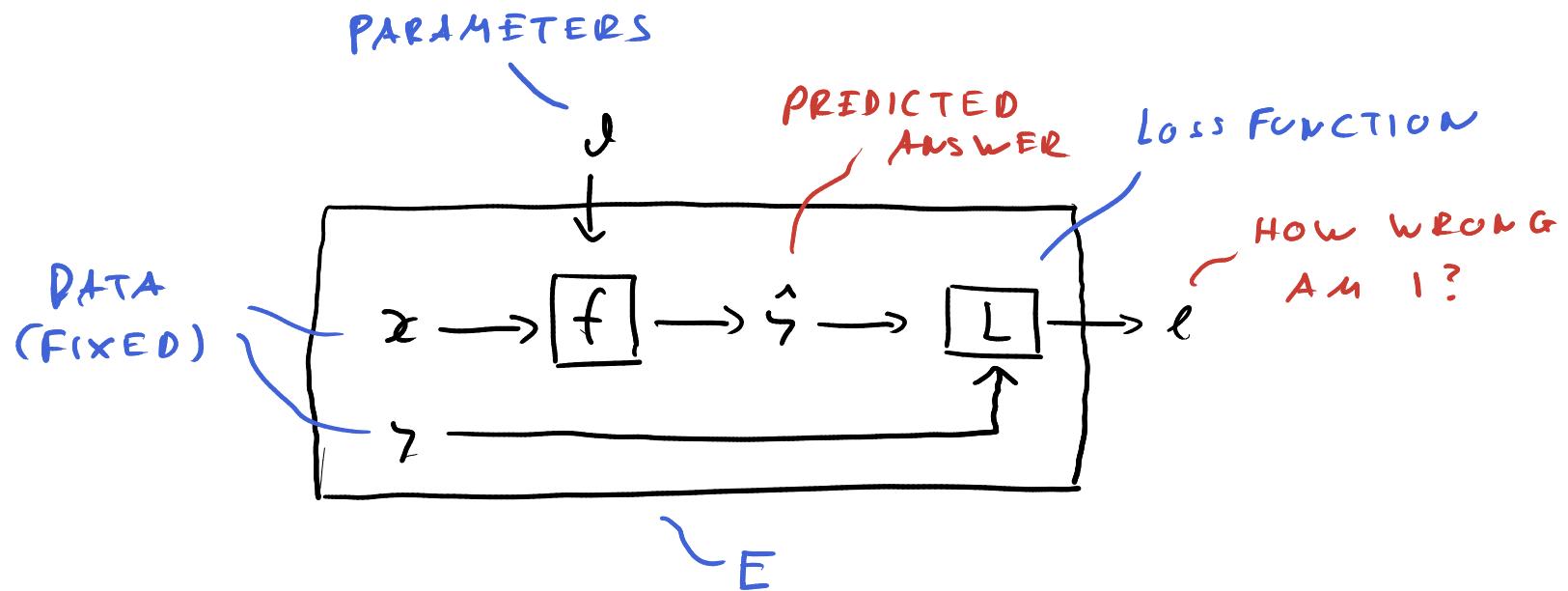
ENERGY MINIMIZATION



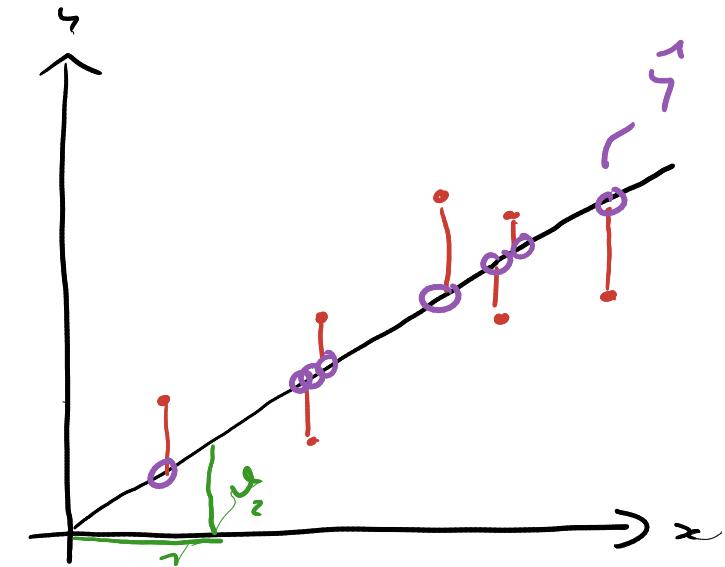
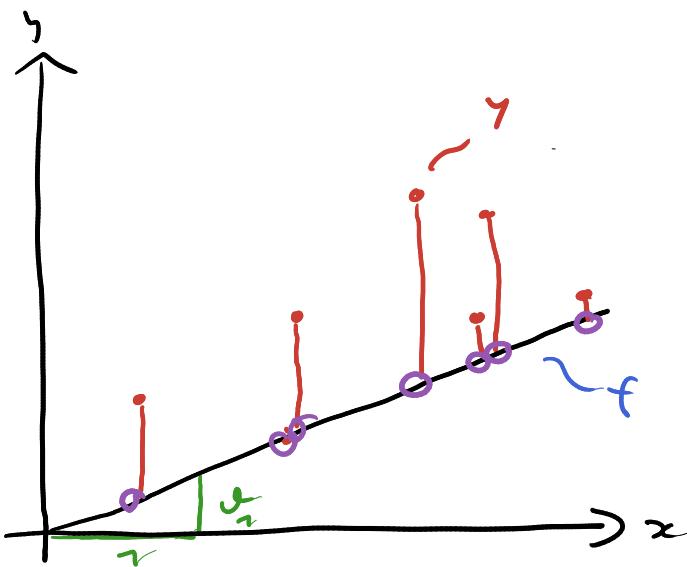
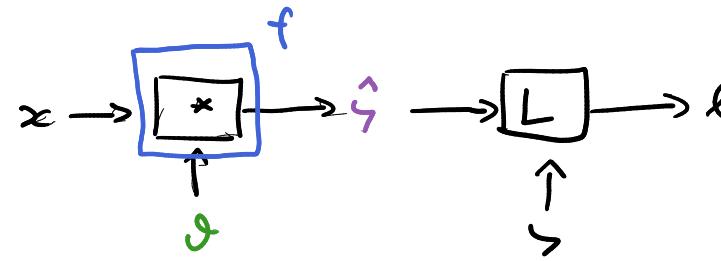
PARAMETRIZED FUNCTIONS



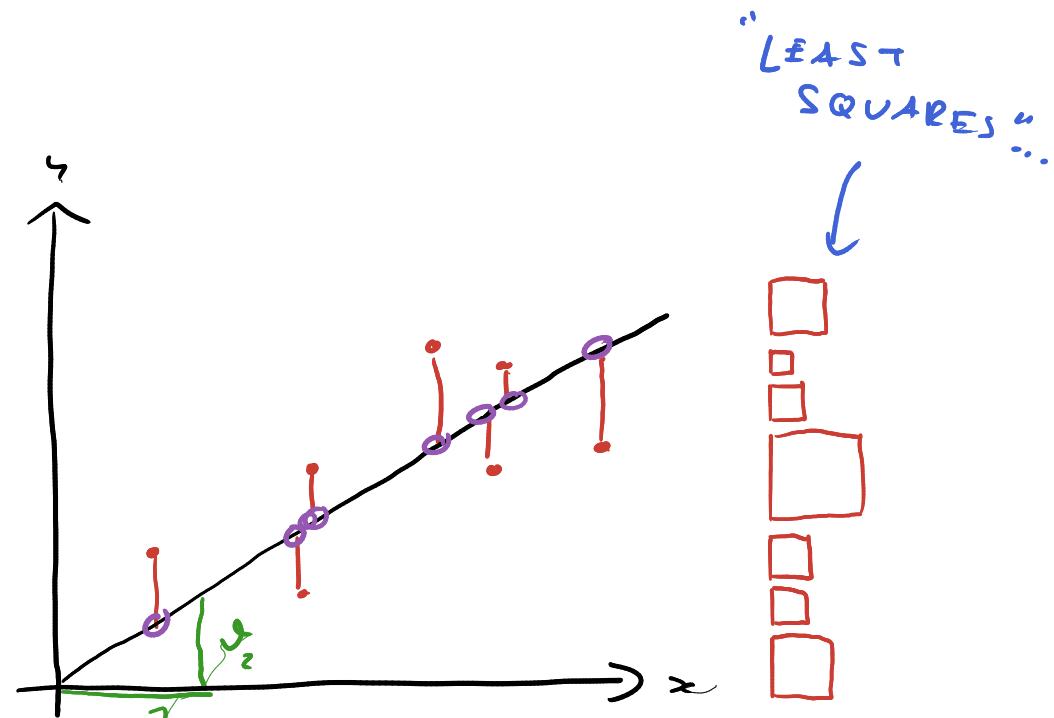
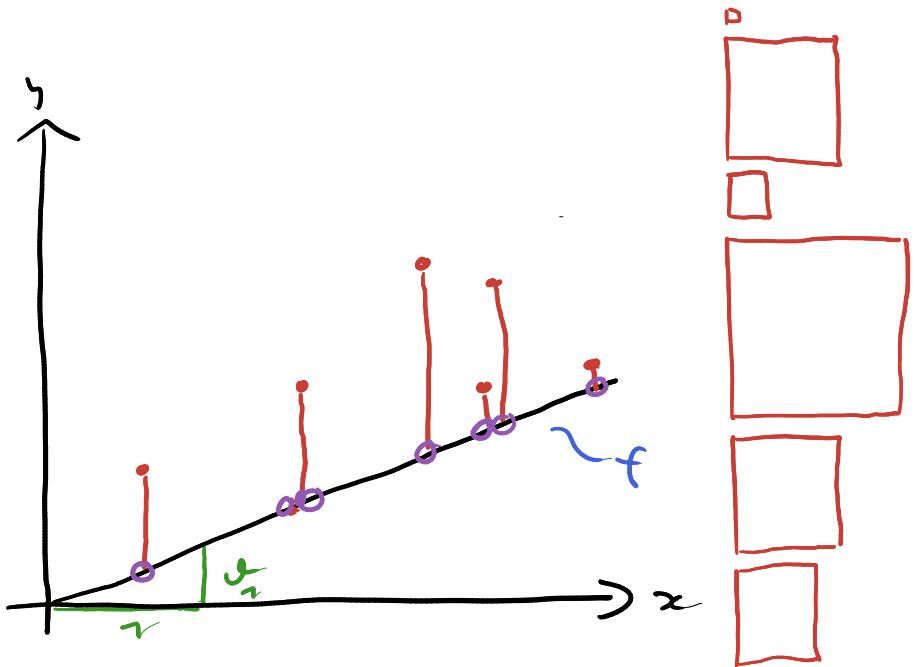
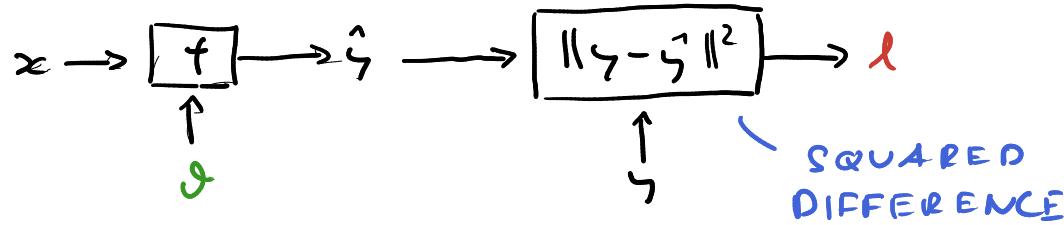
LOSS AS ENERGY



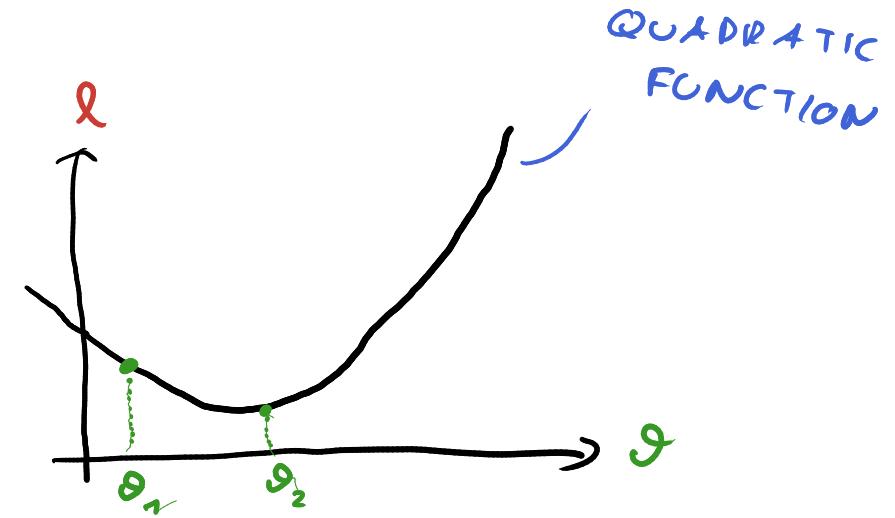
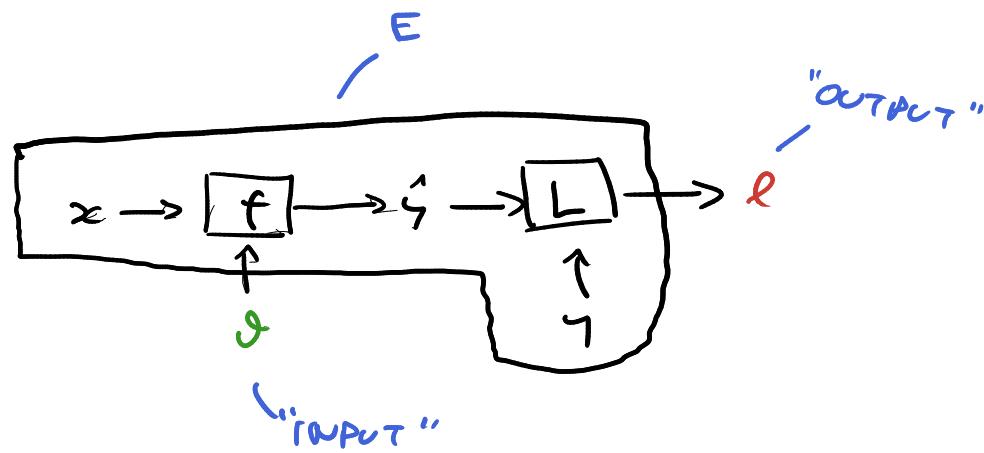
LINEAR REGRESSION



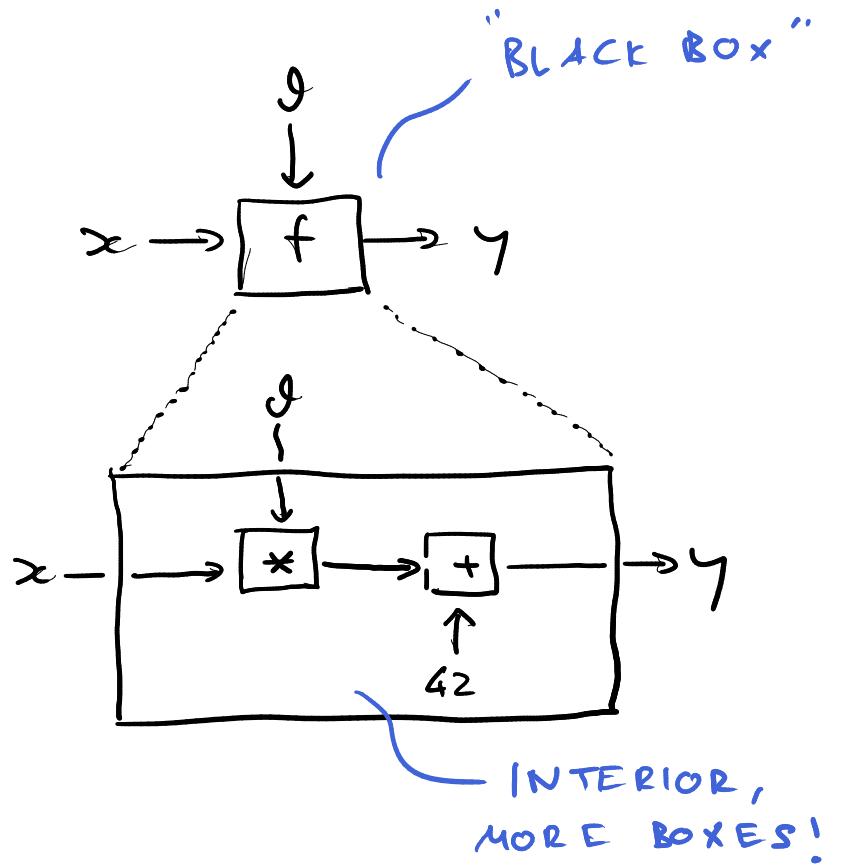
SQUARED DIFFERENCE LOSS



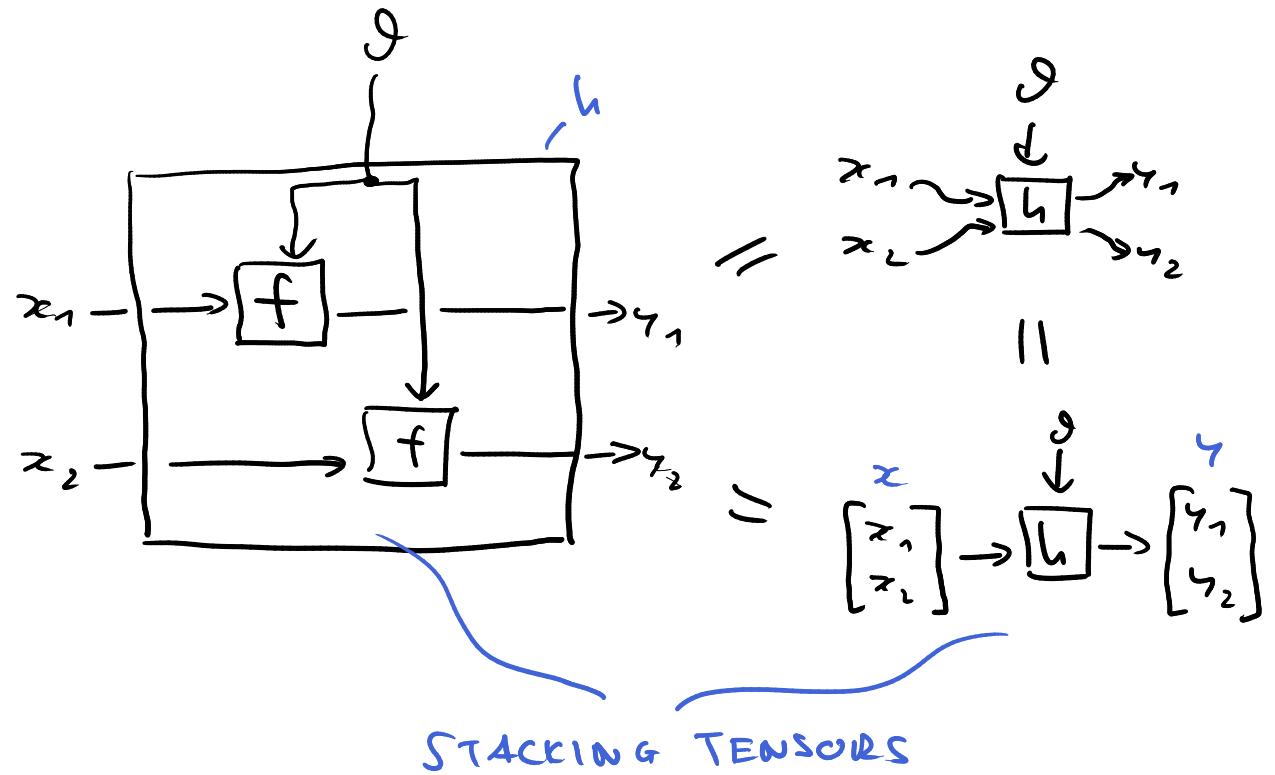
LEAST SQUARES OPTIMIZATION



LEGO FOR FUNCTIONS

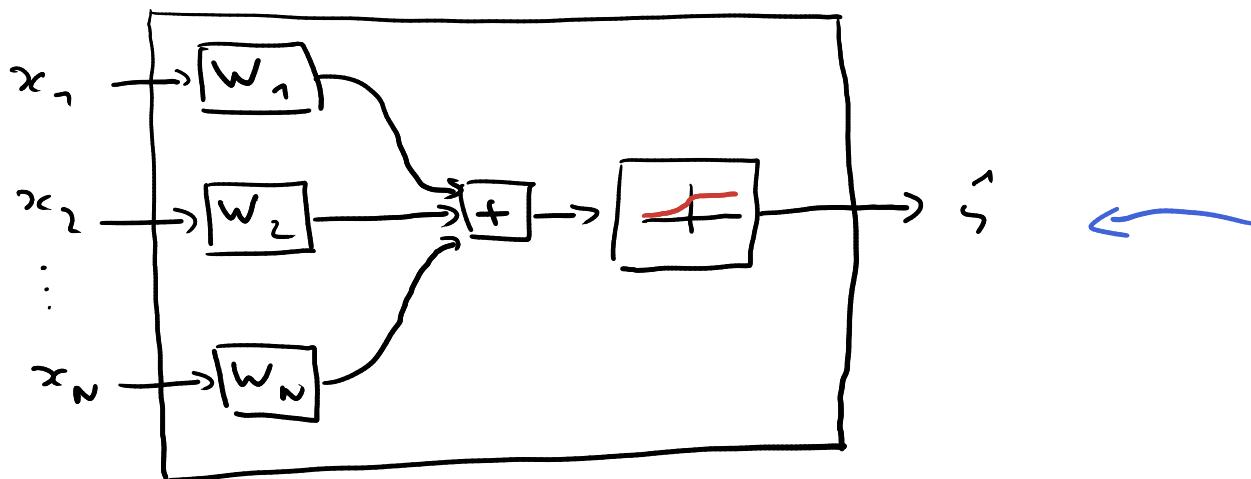
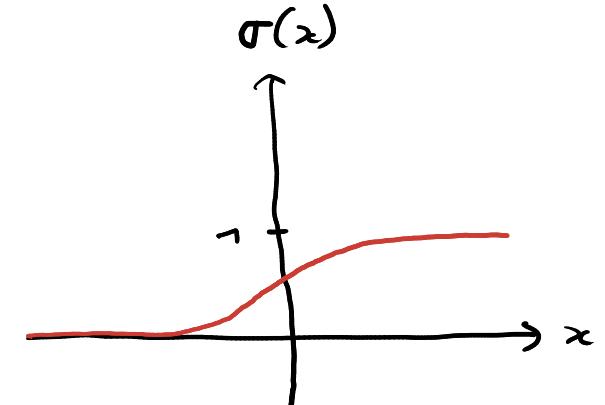
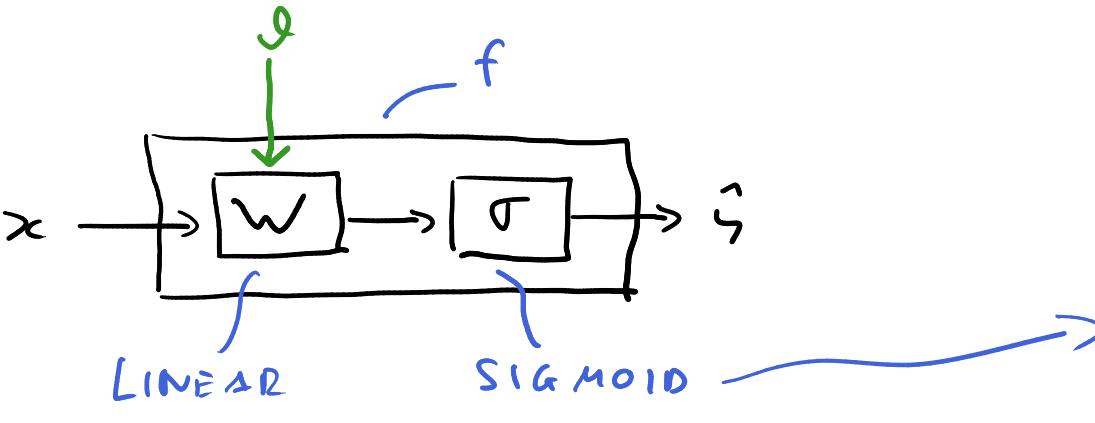


$$(x, \delta) \mapsto \delta x + 42$$



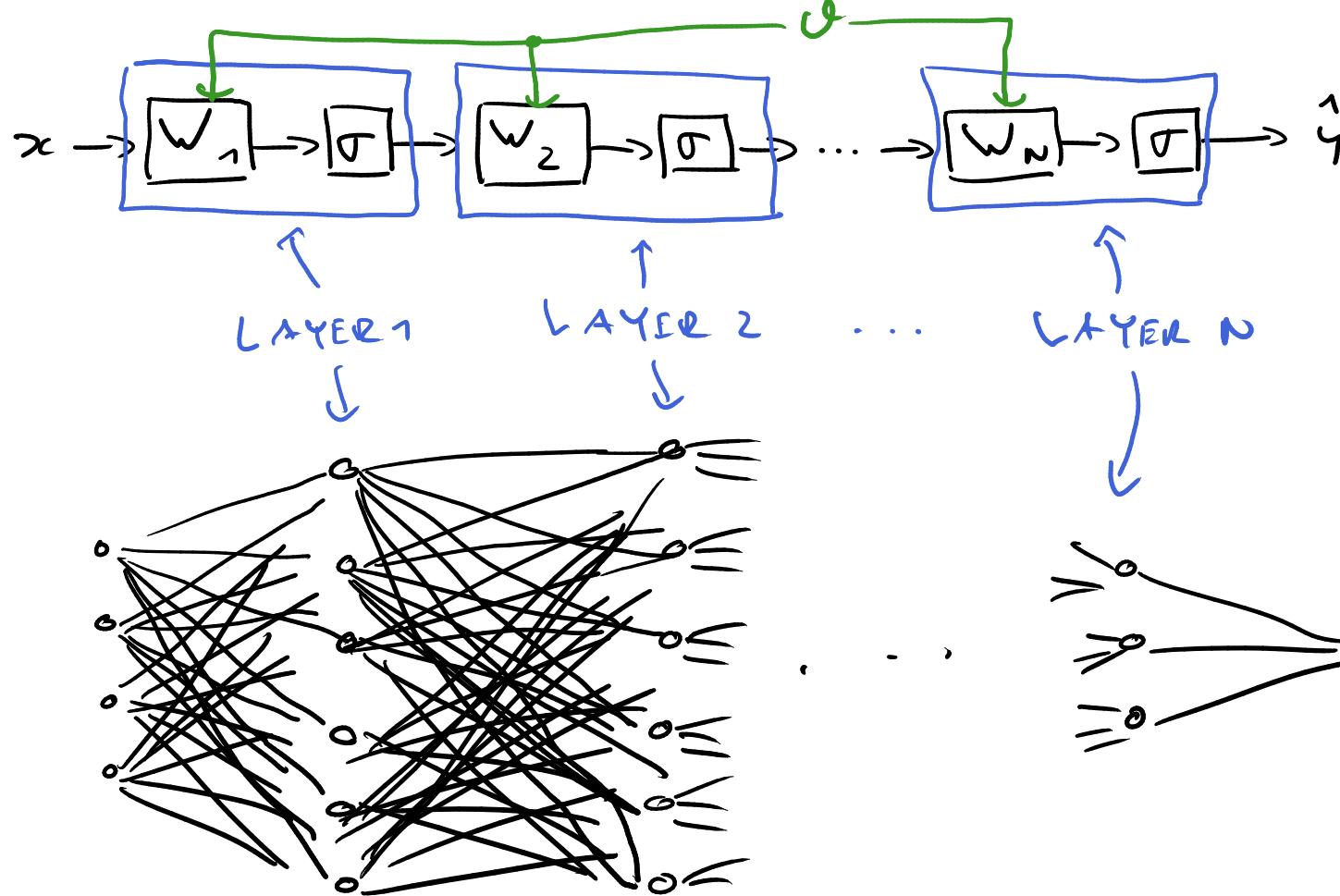
$$(x, \delta) \mapsto [f(x_1, \delta), f(x_2, \delta)]$$

LOGISTIC REGRESSION



REMEMBER THE
PERCEPTRON?

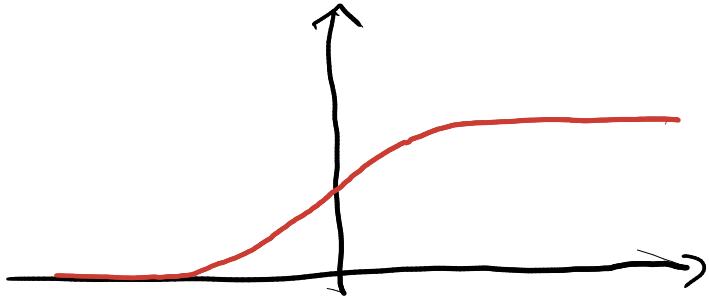
DEEP LEARNING = DEEP F



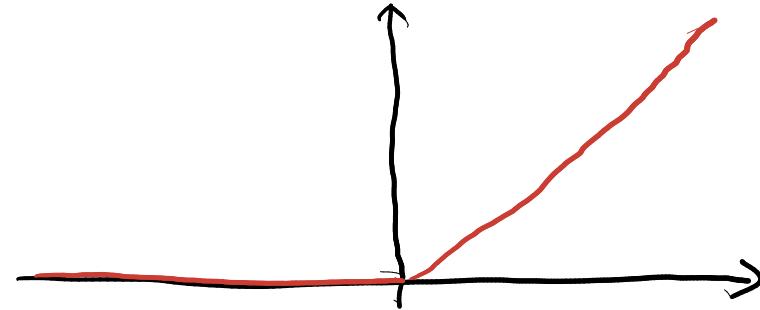
MANY
PARAMETERS!

"MULTI-LAYER PERCEPTRON" / FEEDFORWARD NETWORK

DIFFERENT NONLINEARITIES



SIGMOID

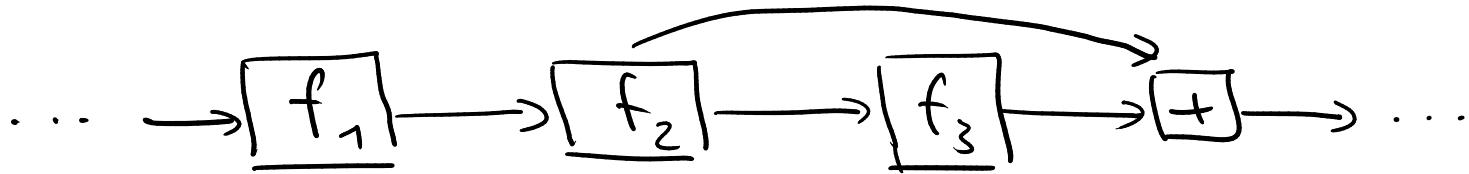


RELU

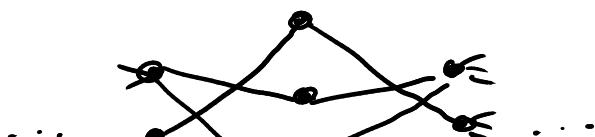


LINEAR OUTPUT

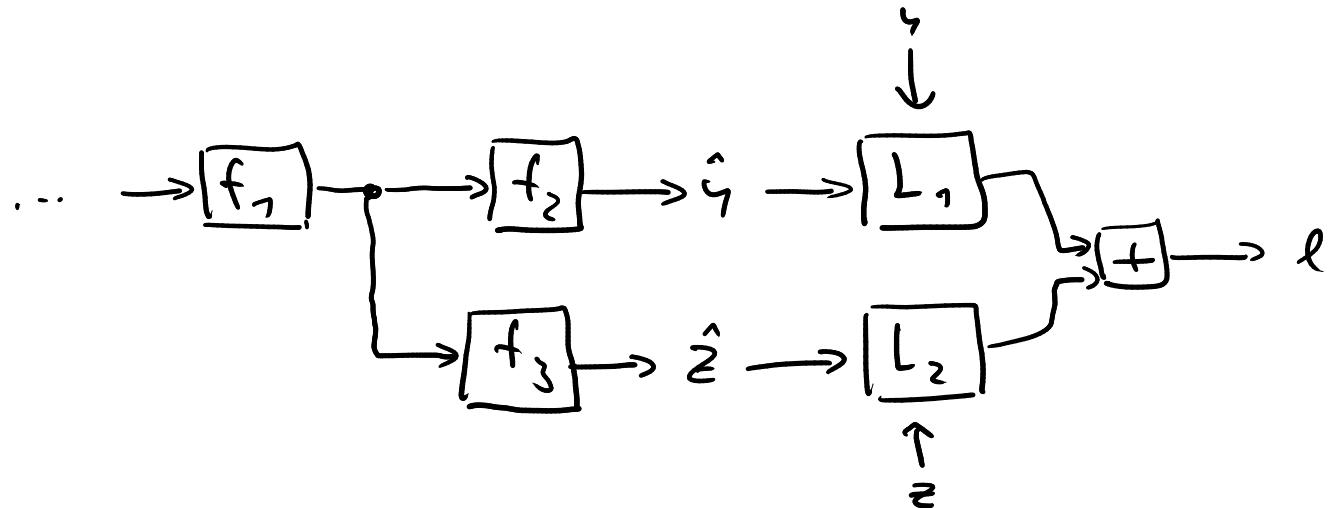
MORE VARIATION



SKIP CONNECTIONS

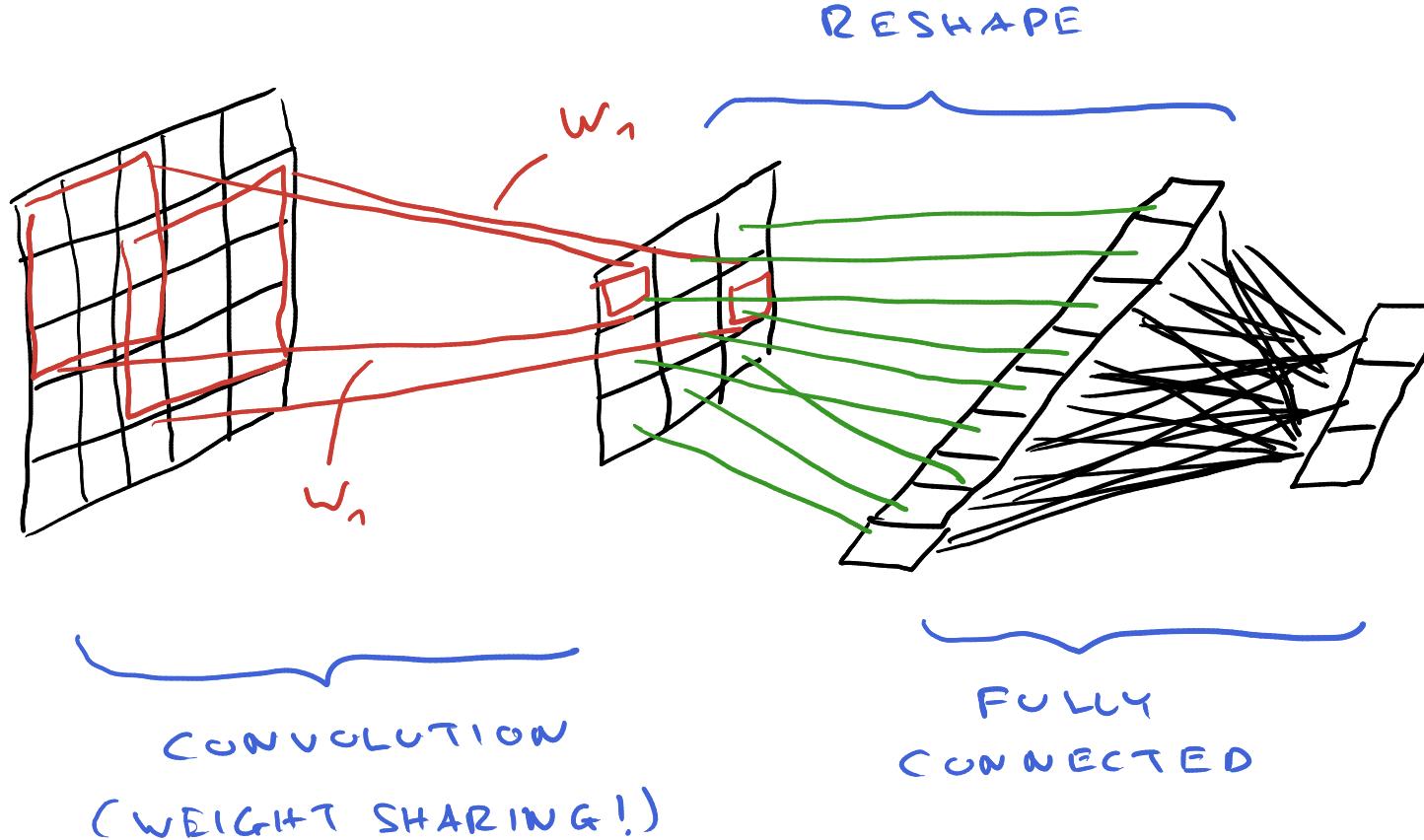


DROPOUT

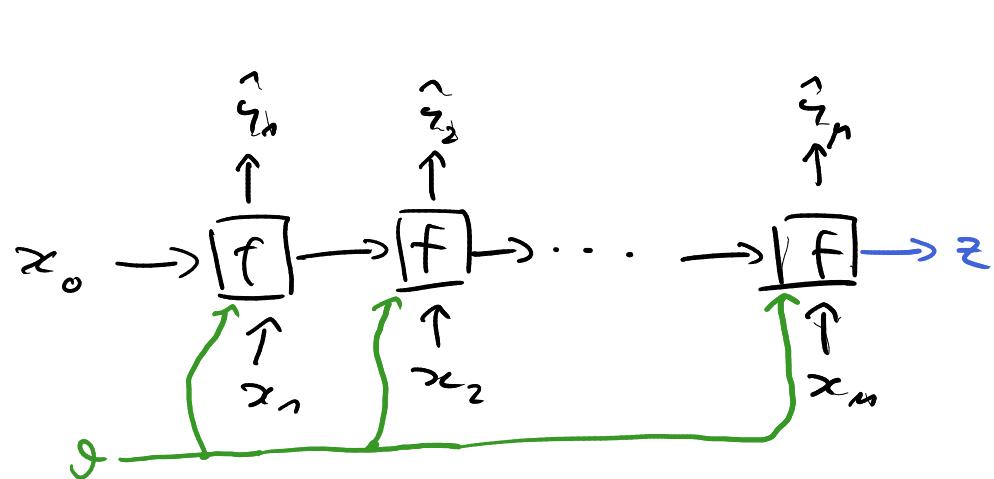


SPLIT LOSS

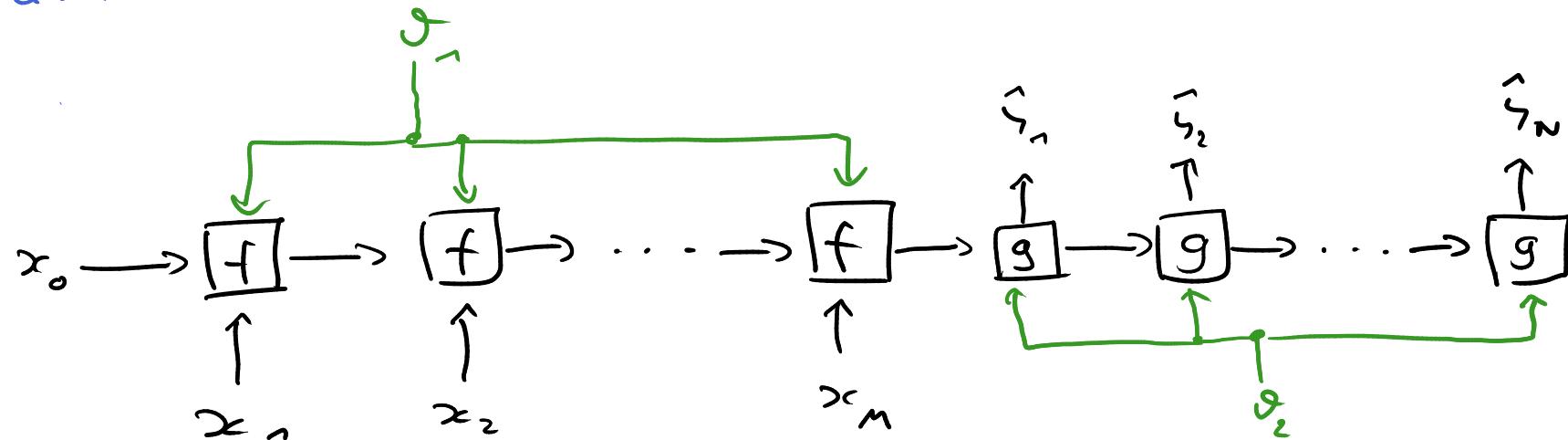
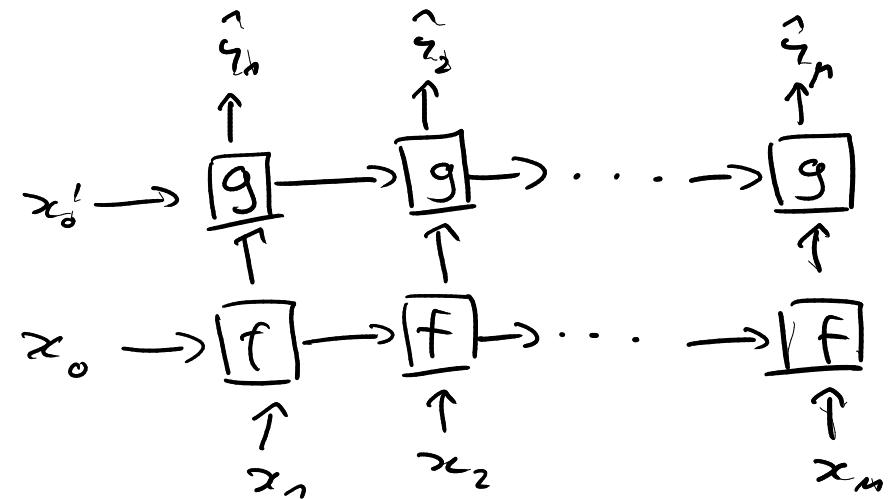
CONVOLUTIONAL NETWORKS



RECURRENT NETWORKS

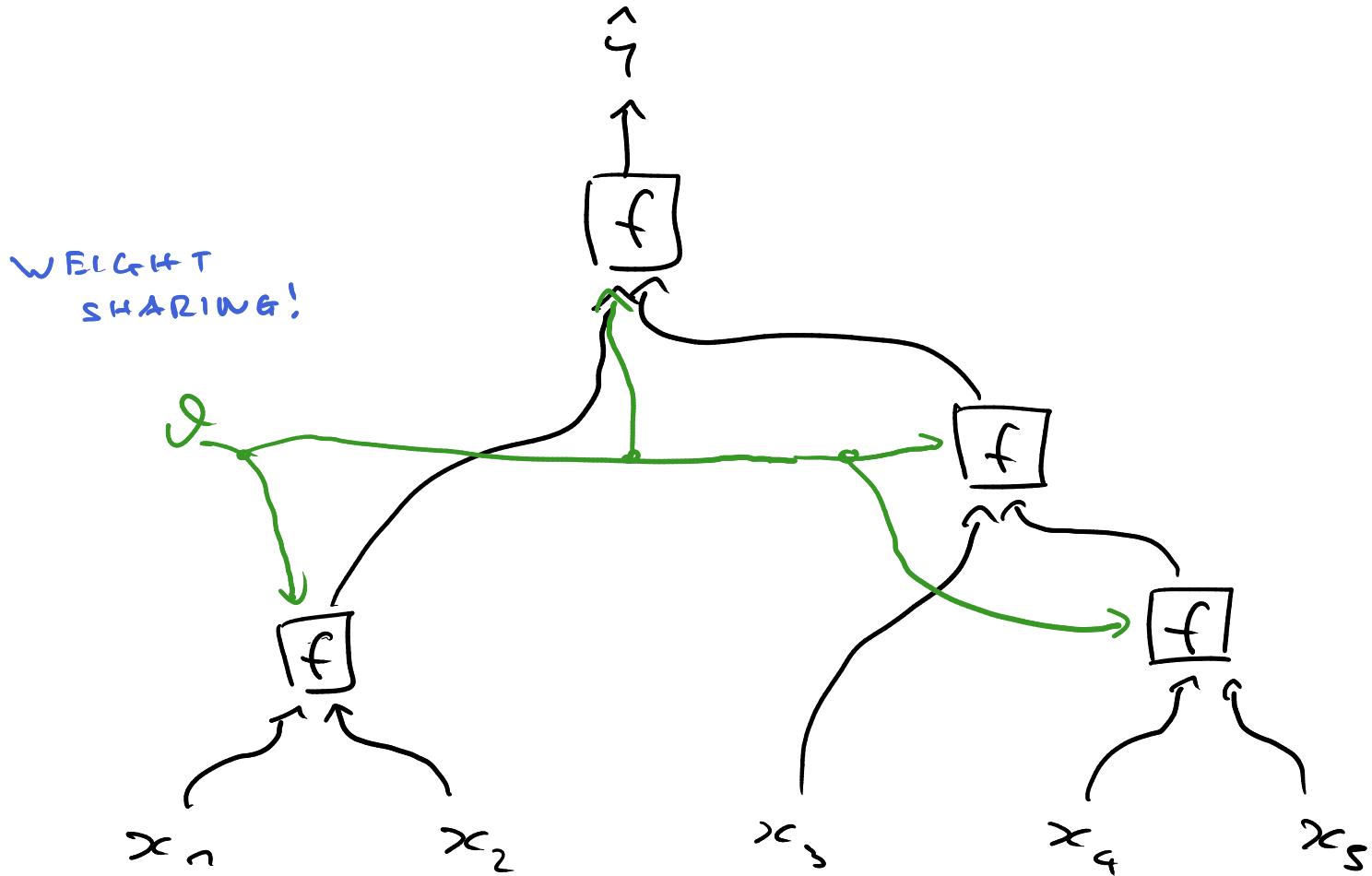


WEIGHT SHARING!

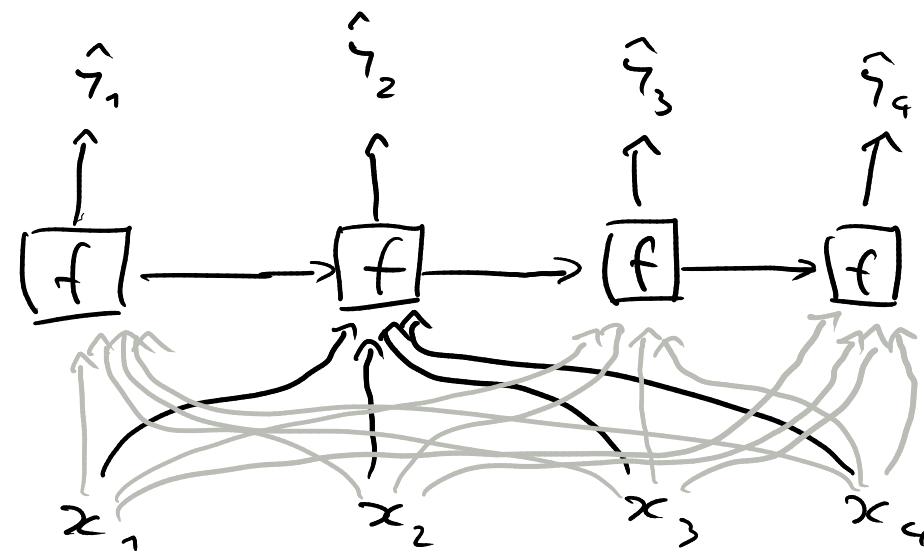


SEQ2SEQ

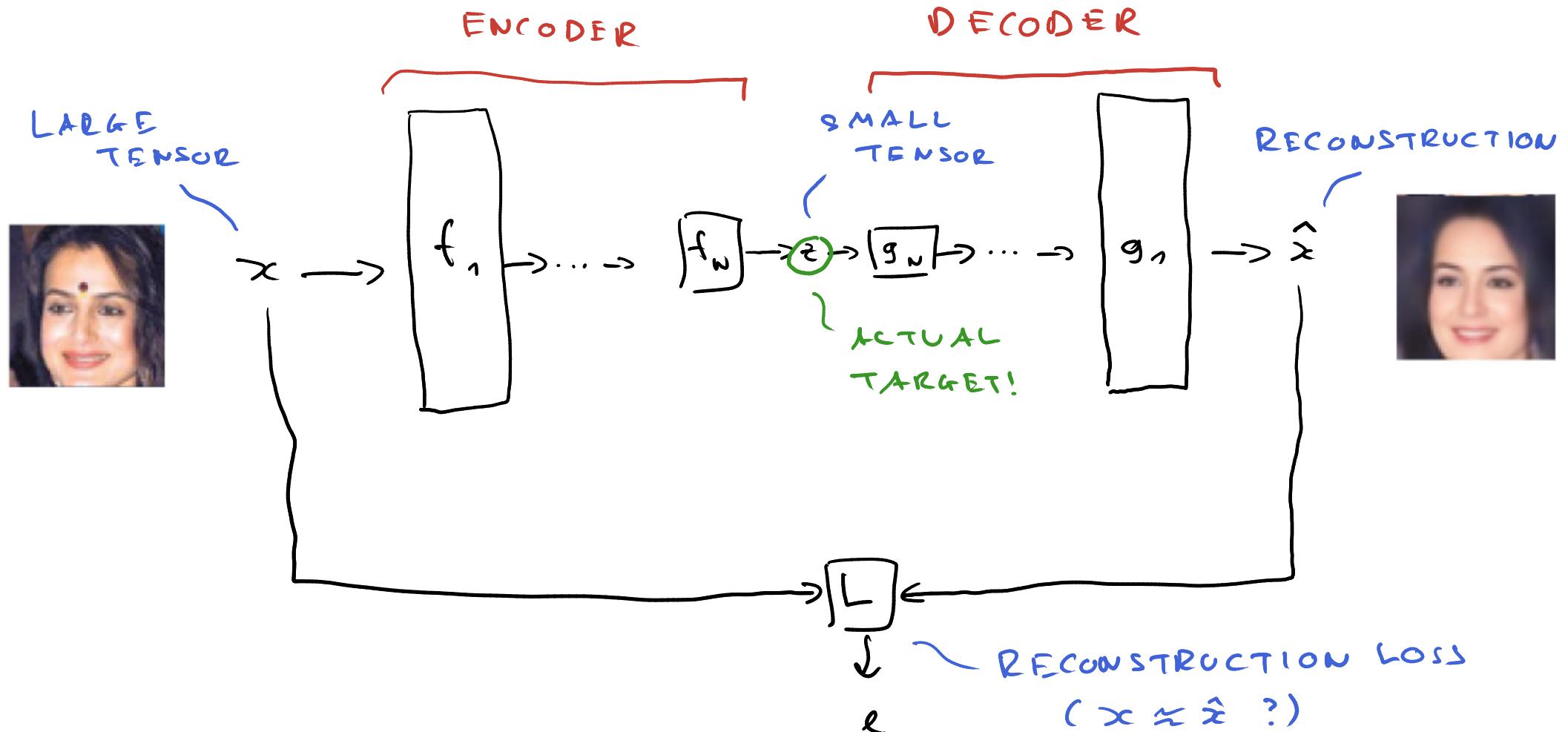
RECURSIVE NETWORKS



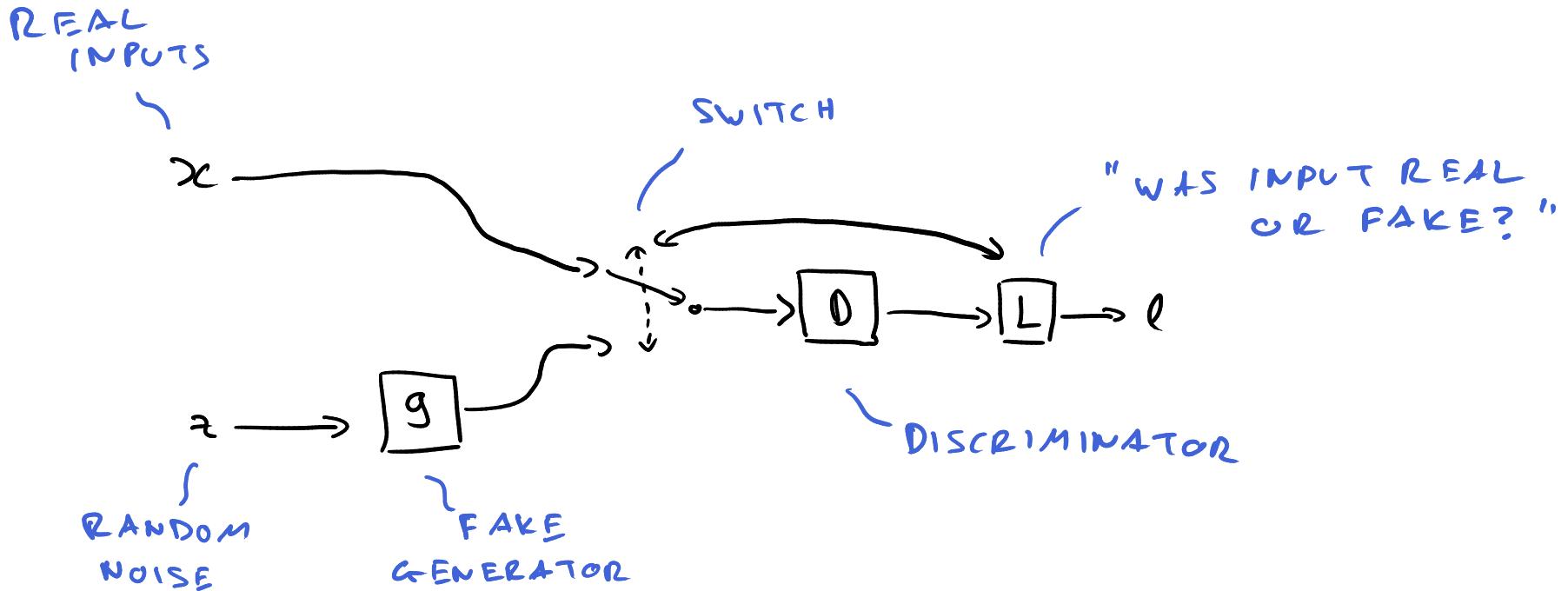
ATTENTION!



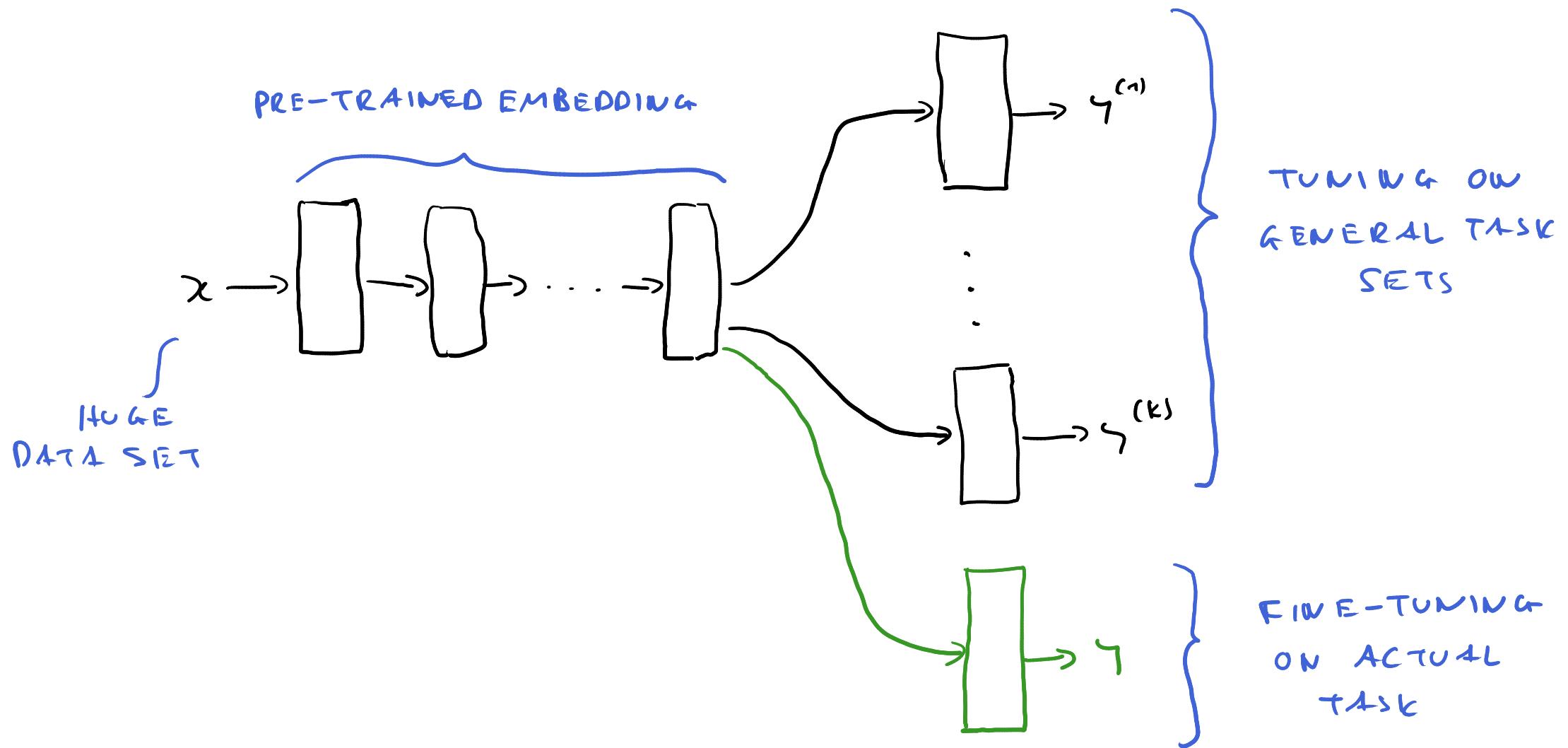
AUTO ENCODER



GENERATIVE ADVERSERIAL NETWORKS



TRANSFER LEARNING



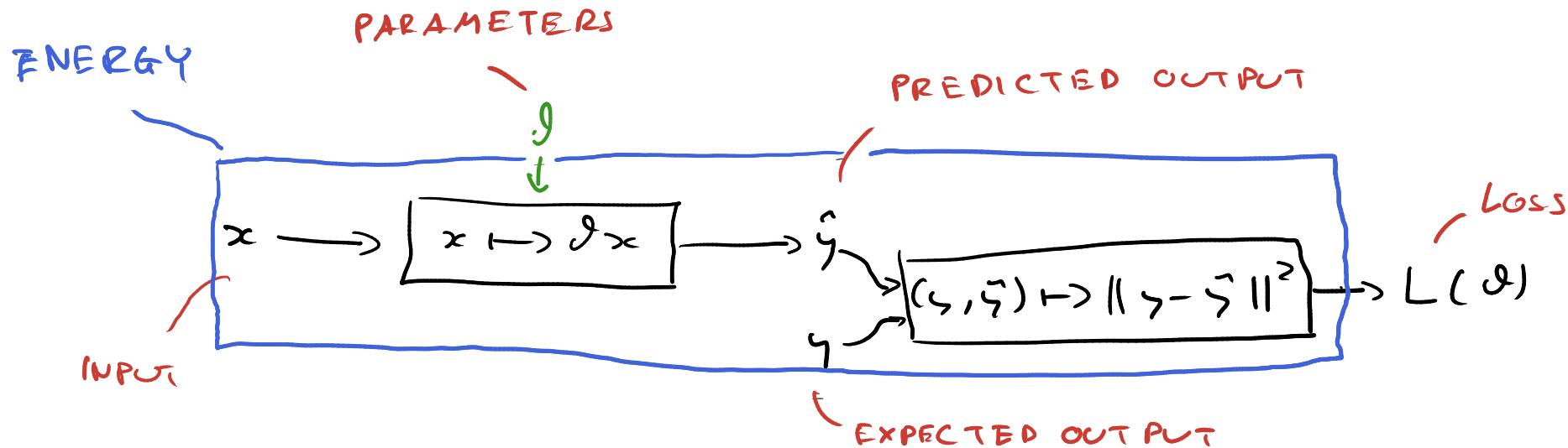
④

HOW ARE SOLUTIONS
IMPLEMENTED?



<https://pxhere.com/en/photo/942155>

OPTIMIZATION IN OLS



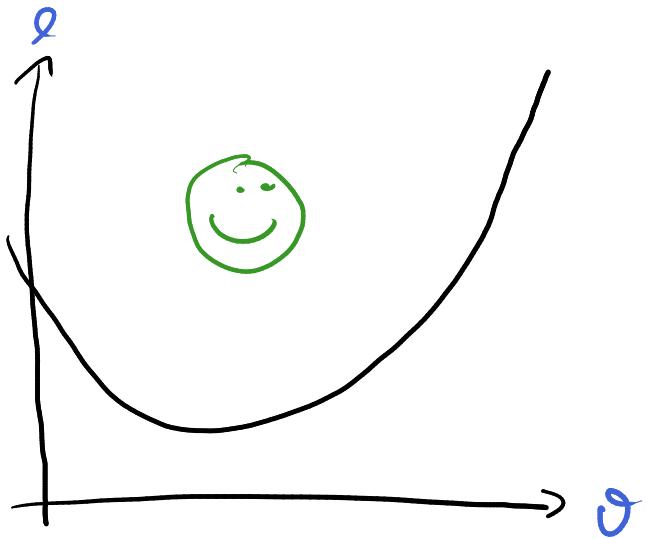
$$\frac{dL}{d\theta} = \phi$$

"ABLEITEN, NULL SETZEN"

$$\theta^* = (X^\top X)^{-1} X^\top y$$

Complexity of Loss Functions

LEAST SQUARES: QUADRATIC



- CONVEX

- SIMPLE MATH

AN: "LANDSCAPE"



- NON-CONVEX

- NO ANALYTICAL
SOLUTION

REAL LOSS LANDSCAPES



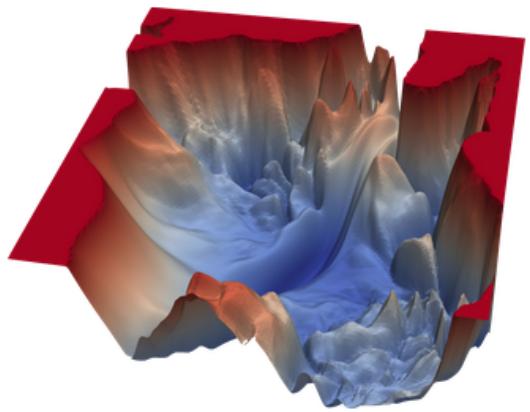
<https://losslandscape.com/gallery/>

THERE'S HOPE!

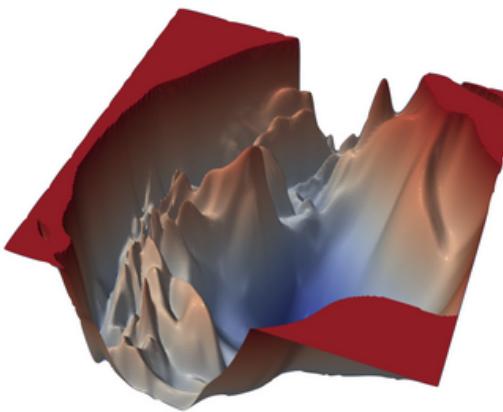


INFLUENCE OF ARCHITECTURES

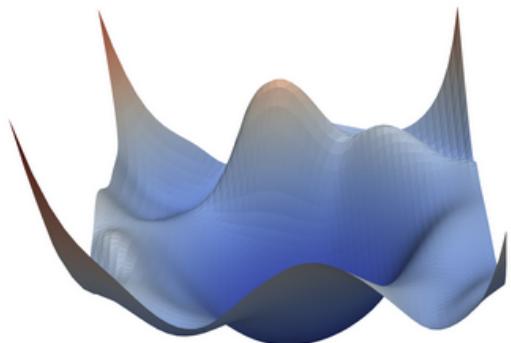
VGG-56



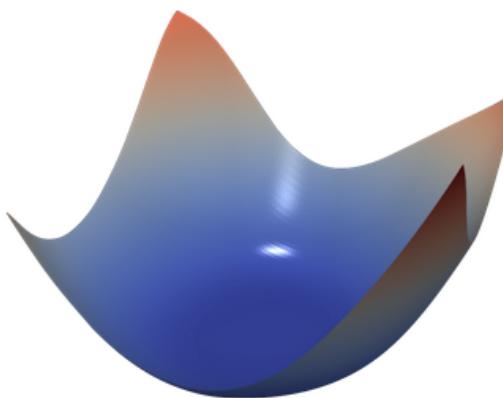
VGG-110



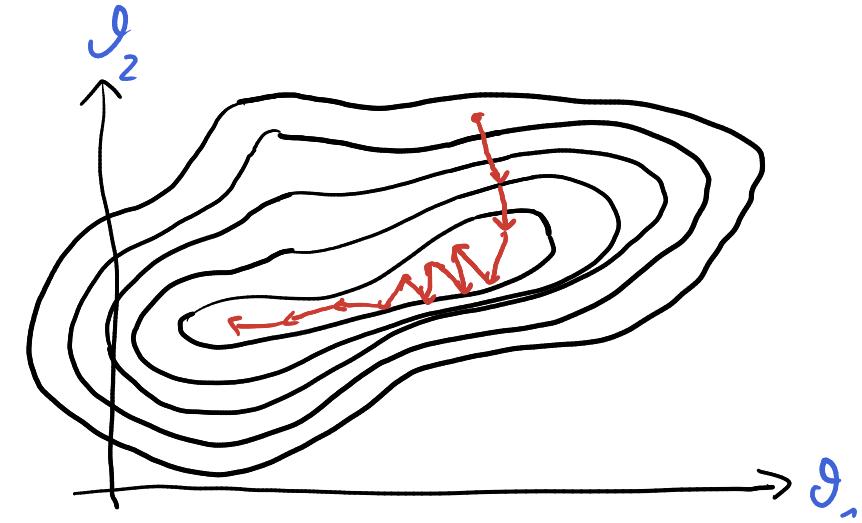
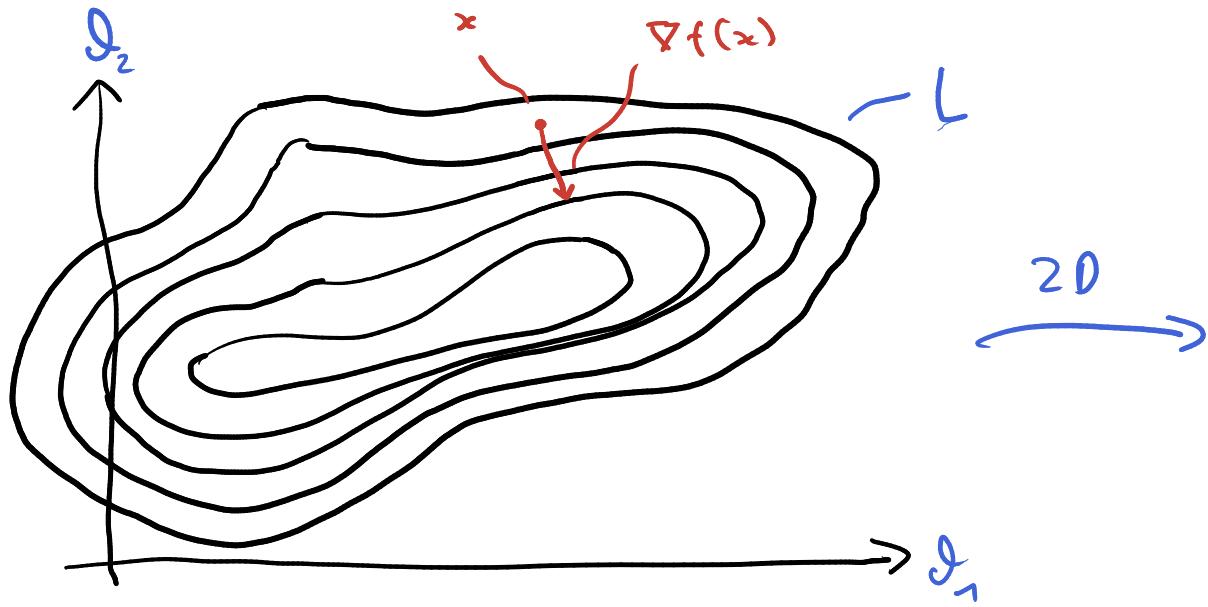
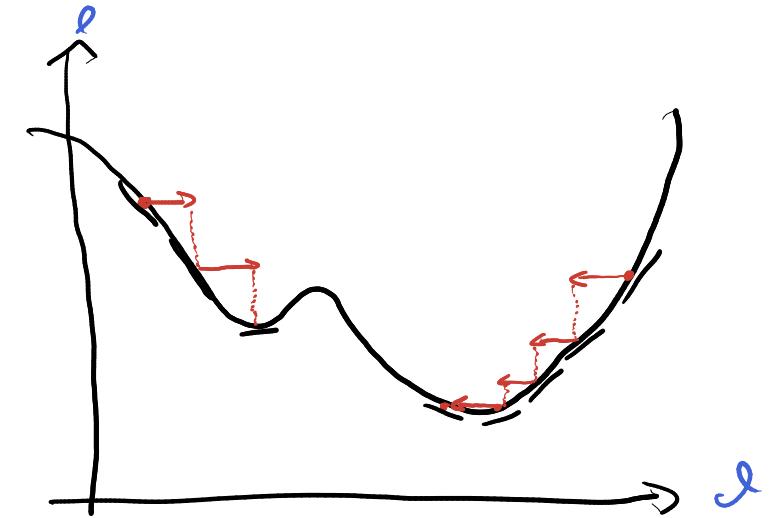
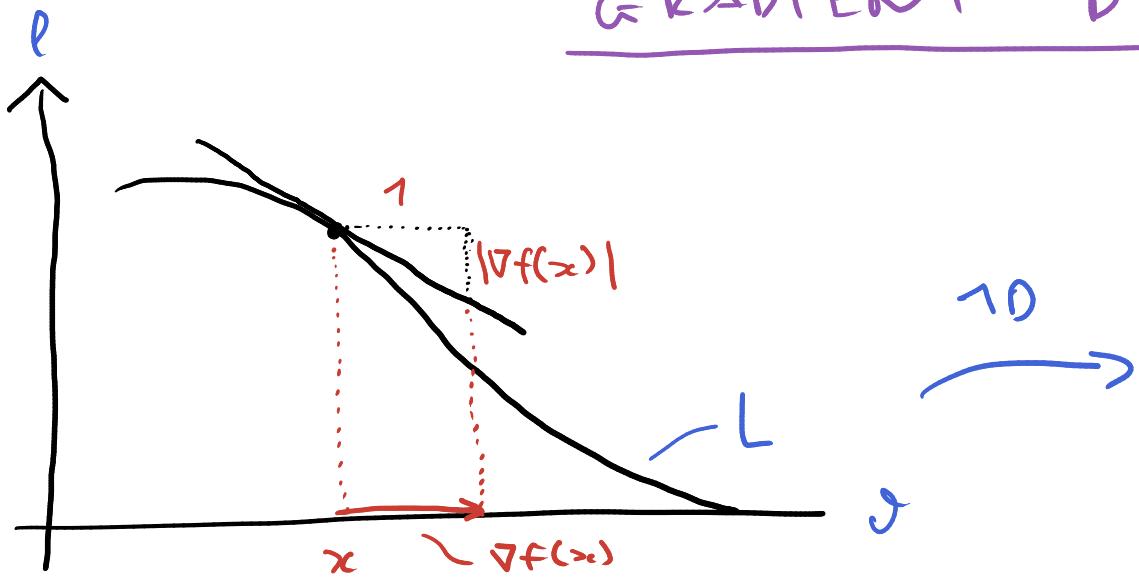
Renset-56



Densenet-121



GRADIENT DESCENT



VISUALIZATION OF LEARNING



After 1 epoch



After 250 epochs

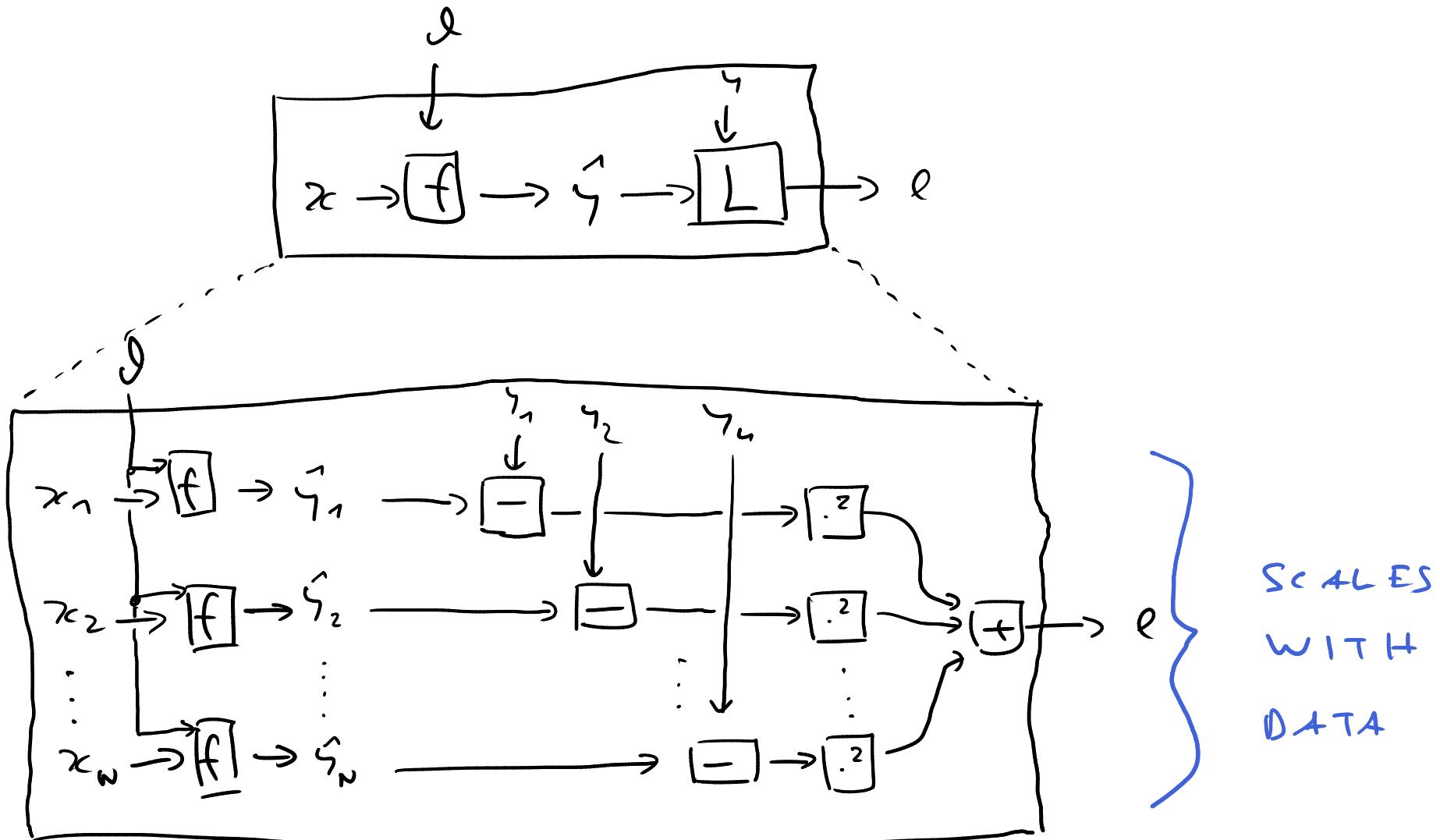


After 500 epochs

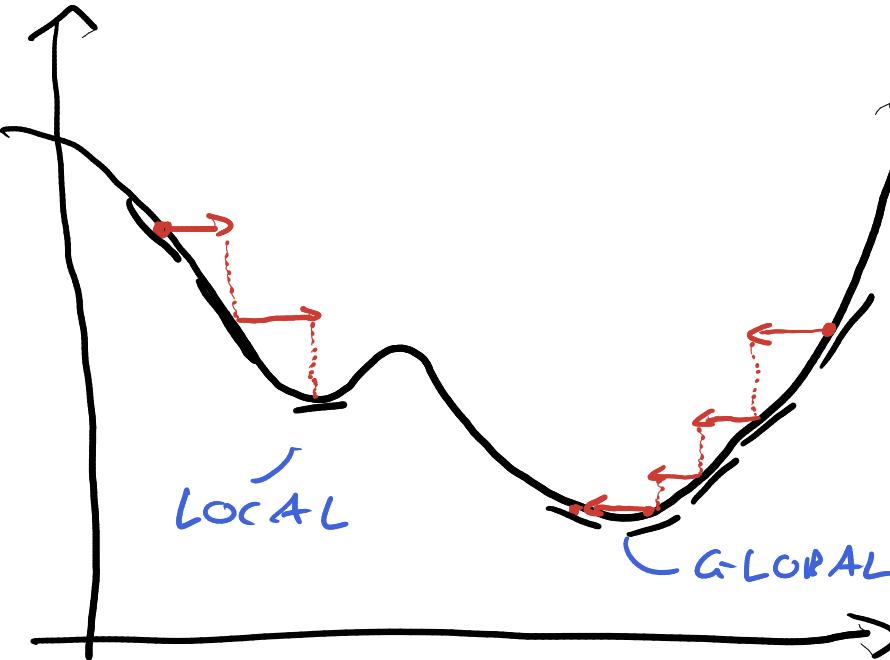


Using separate test-set

NOT SO SIMPLE...



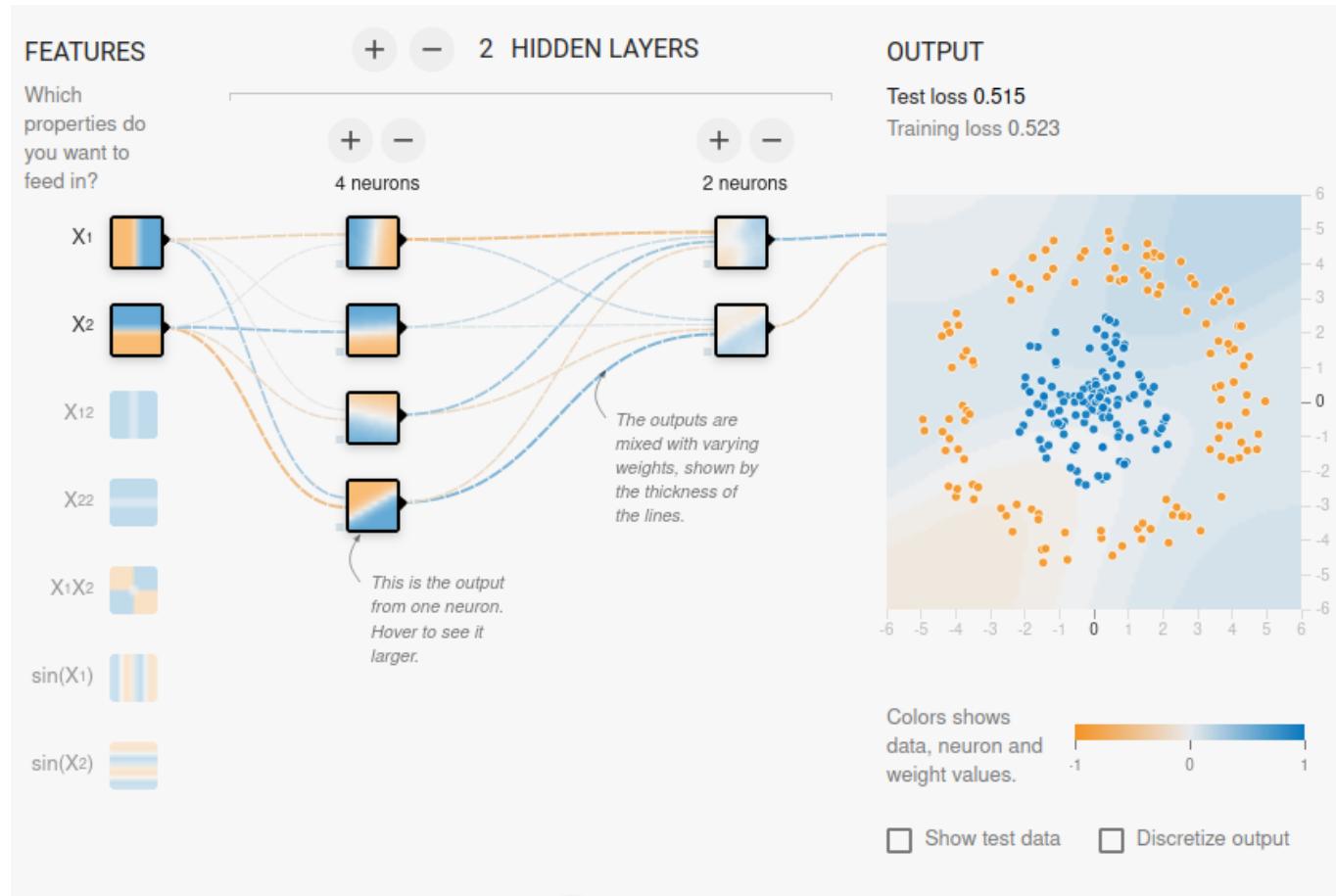
STUCK IN LOCAL OPTIMA



STOCHASTIC GRADIENT DESCENT!

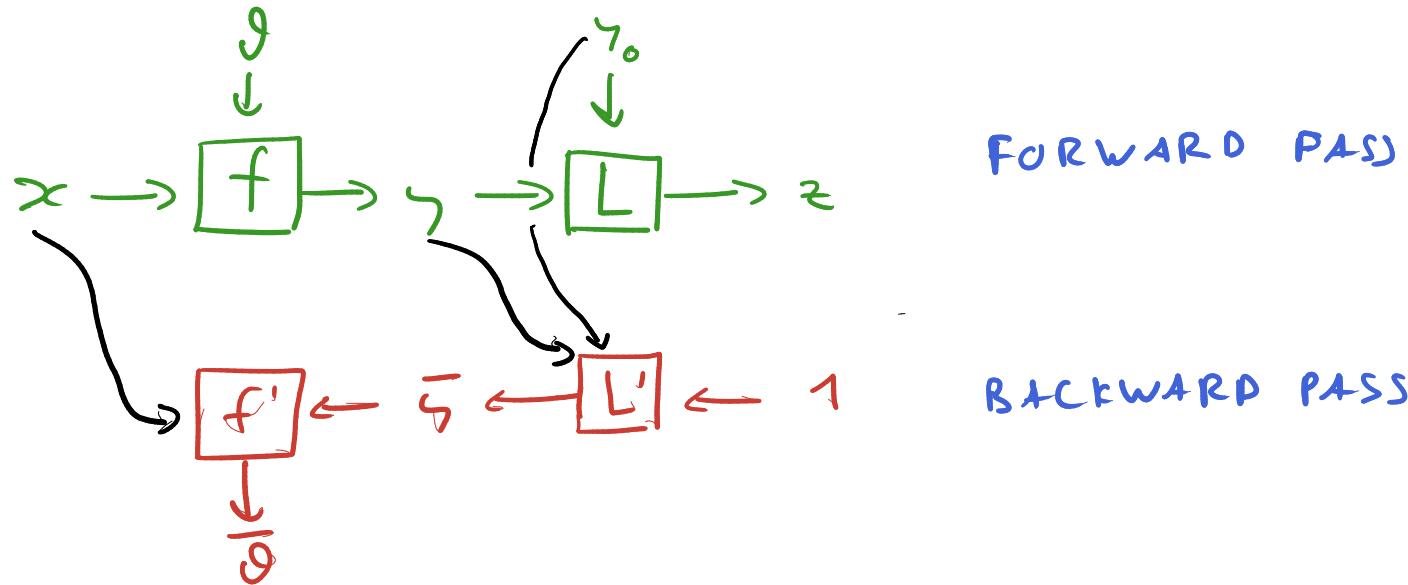
- 1) SAMPLE (\tilde{x}, \tilde{y}) FROM DATA (SMALL)
 - 2) CALCULATE $\nabla L(\theta)$ ON SAMPLE ~
FIXED
SIZE!
 - 3) DO GRADIENT STEP IN NOISY
DIRECTION!
~ WIGGLE OUT OF LOCAL
OPTIMA!
- ALSO: MORE ROBUST LEARNING

PLAY GROUND



<https://playground.tensorflow.org>

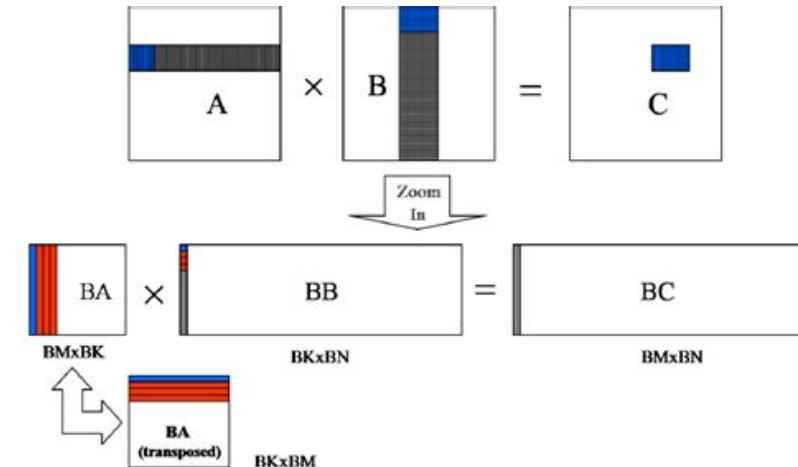
MAY THE GRAPH BE WITH YOU



$$\bar{e} = \nabla_{\theta} L(\overbrace{f(x, \theta)}^h, \gamma_0) = \underbrace{L'(\overbrace{f(x, \theta)}^h, \gamma_0)}_{\bar{h}} f'(x, \theta)$$

LINEAR ALGEBRA'S DIRTY SECRETS

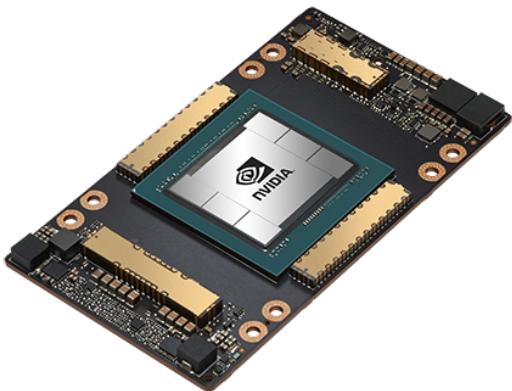
```
1 void Gemm(int m, int n, int k, double *A, double *B, double *C) {
2     // Declarations: mc, nc, kc, ...
3     for ( jc = 0; jc < n; jc += nc ) {                                // Loop 1
4         for ( pc = 0; pc < k; pc += kc ) {                            // Loop 2
5             // B(pc : pc + kc - 1, jc : jc + nc - 1) → Bc
6             Pack_buffer_B(kc, nc, &B(pc, jc), &Bc);
7             for ( ic = 0; ic < m; ic += mc ) {                          // Loop 3
8                 // A(ic : ic + mc - 1, pc : pc + kc - 1) → Ac
9                 Pack_buffer_A(mc, kc, &A(ic, pc), &Ac);
10                // Macro-kernel:
11                for ( jr = 0; jr < nc; jr += nr ) {                      // Loop 4
12                    for ( ir = 0; ir < mc; ir += mr ) {                  // Loop 5
13                        // Micro-kernel:
14                        //   Cc(ir : ir + mr - 1, jr : jr + nr - 1) +=
15                        //   Ac(ir : ir + mr - 1, 1 : 1 + kc - 1) ·
16                        //   Bc(j, 1 : 1 + kc - 1, r : jr + nr - 1)
17                        Gemm_mkernnel( mr, nr, kc, &Ac(ir, 1), &Bc(1, jr),
18                                         &Cc(ir, jr) );
19                }
20            }
21        }
22    }
```



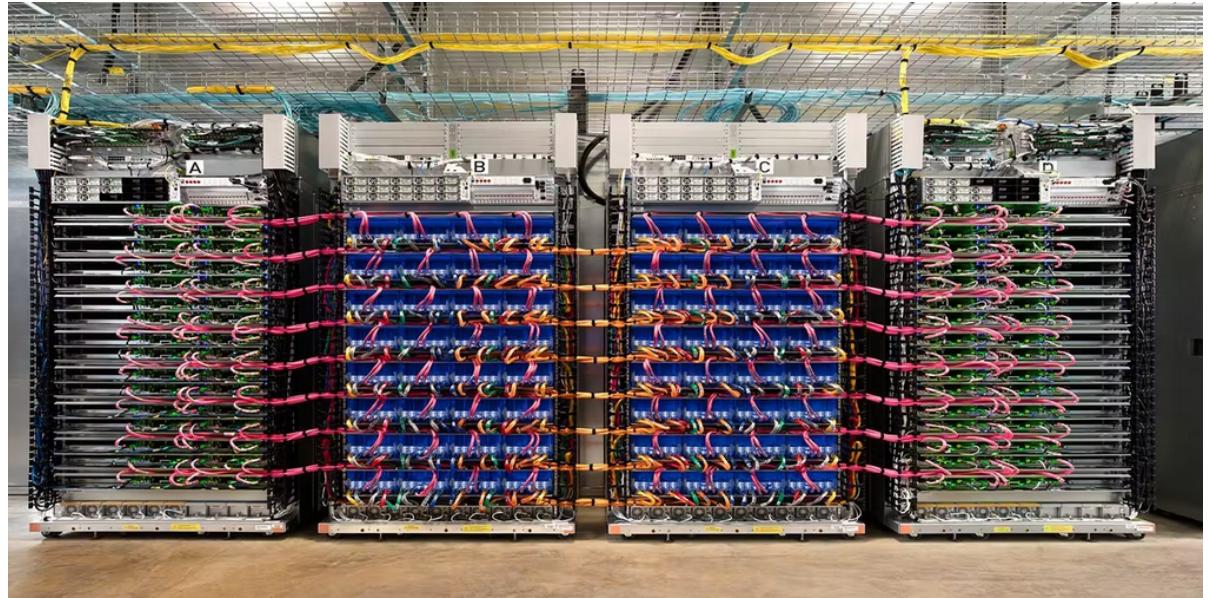
<http://dx.doi.org/10.1007/s10586-019-02927-z>

http://dx.doi.org/10.1007/978-3-642-01970-8_89

MAKE THE HARDWARE GREAT AGAIN

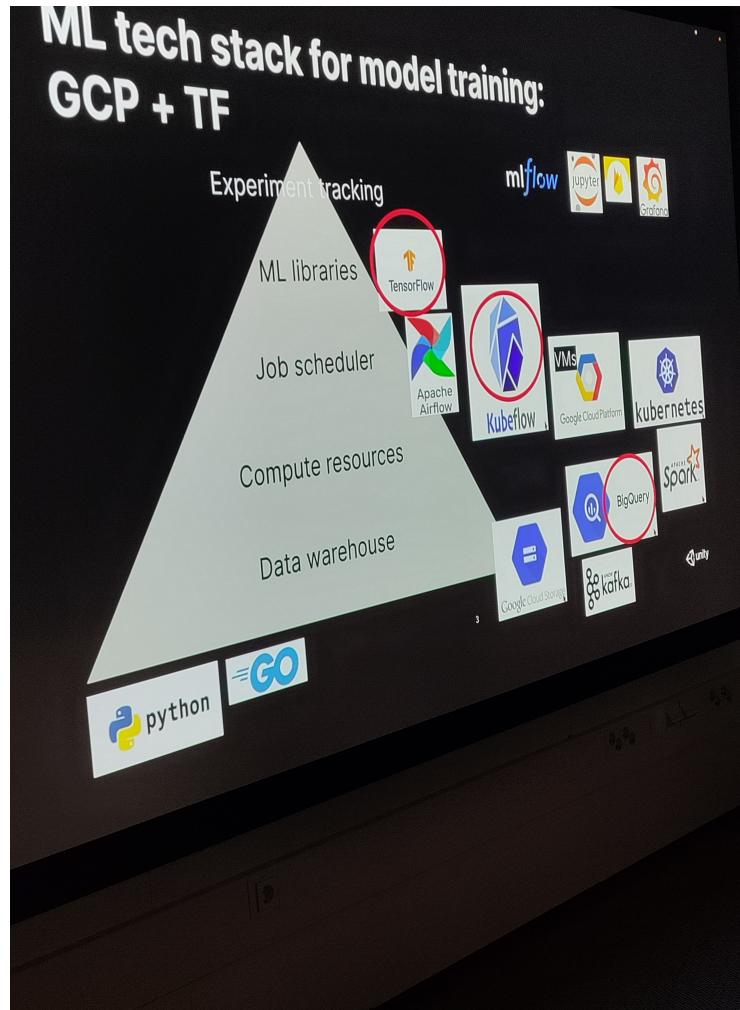


<https://www.gigabyte.com/id/Solutions/nvidia-a100>



<https://www.inverse.com/article/31745-google-tensor-processing-unit>

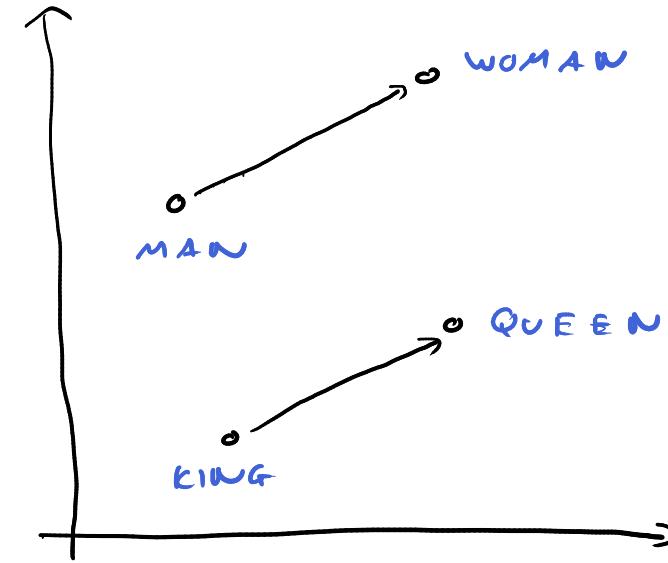
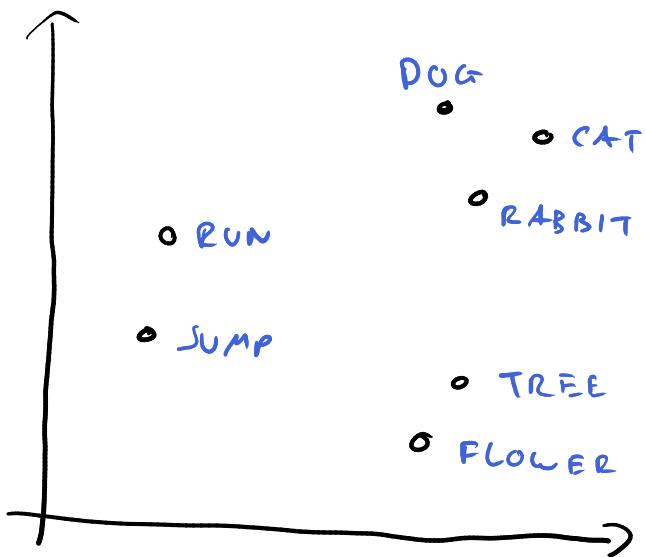
THE POWER OF AN ECOSYSTEM



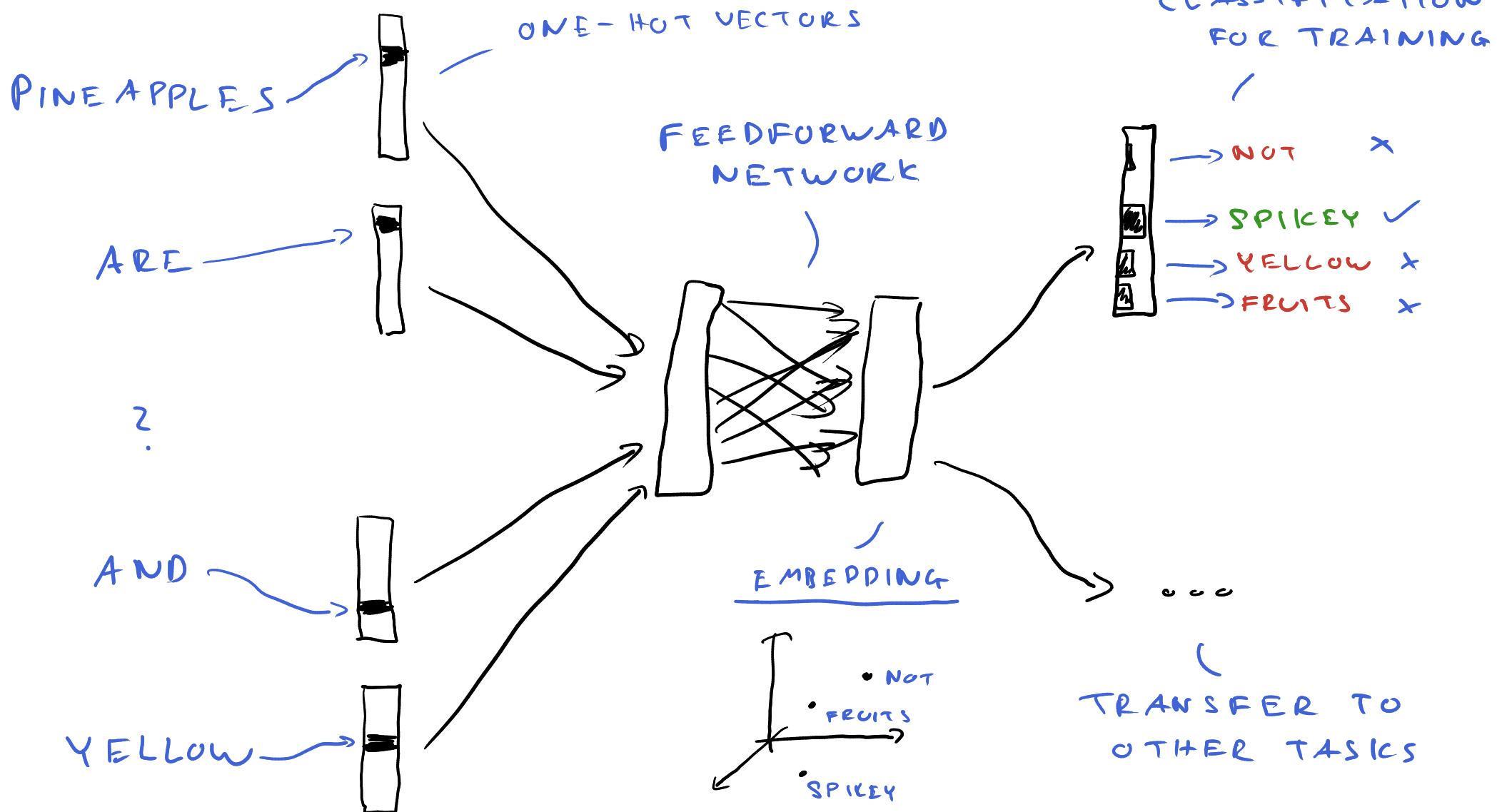
⑤

SOME LANGUAGE - RELATED
USE CASES

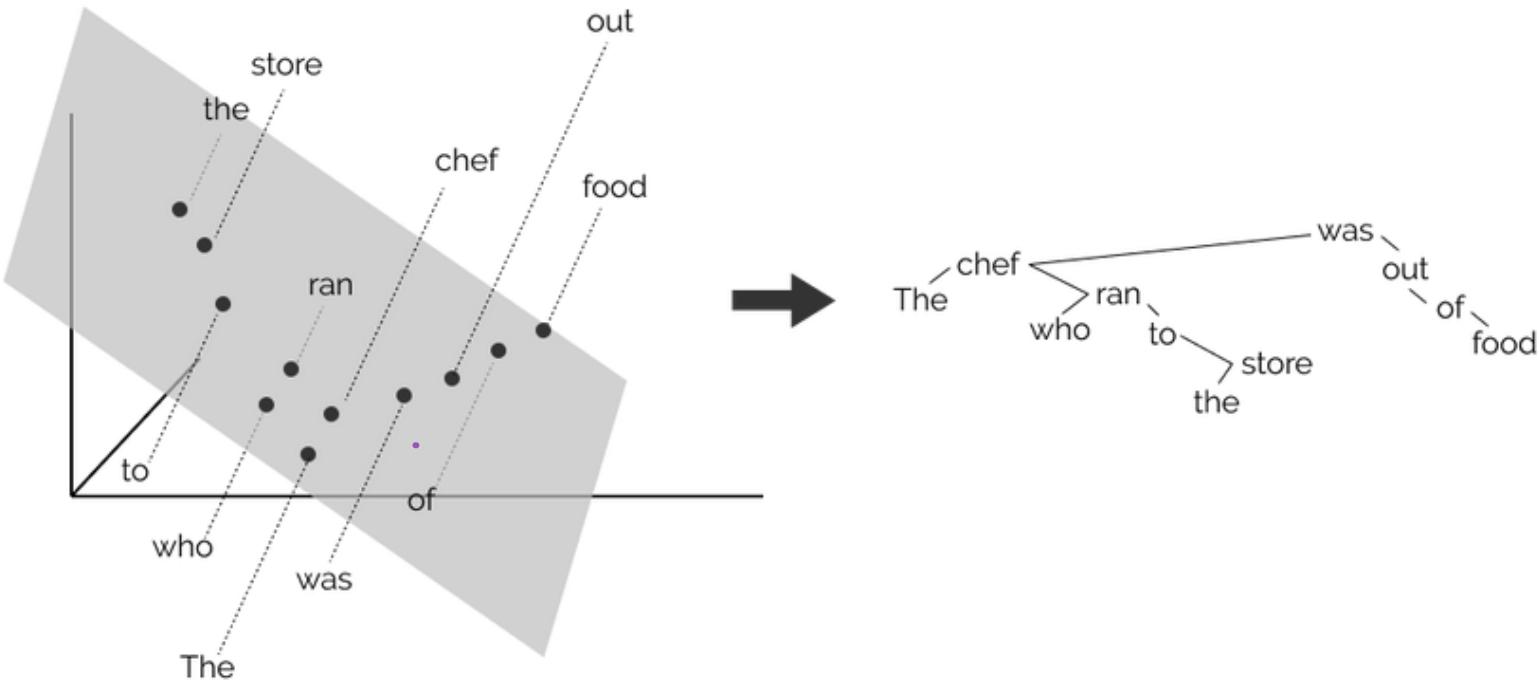
WORD EMBEDDINGS



CONTINUOUS BAG-OF-WORDS



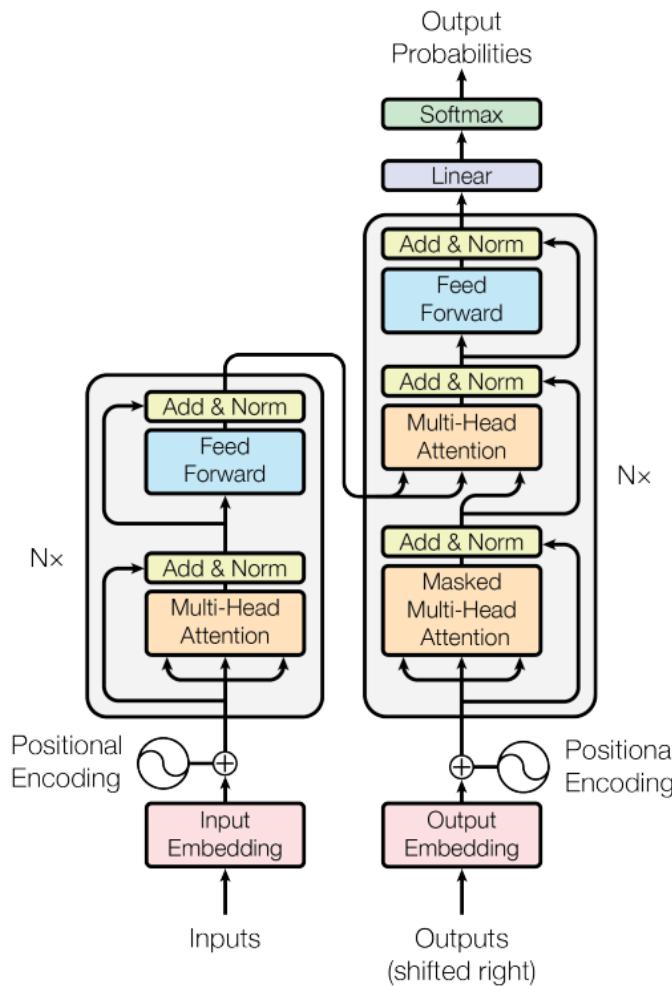
SYNTAX FROM EMBEDDINGS



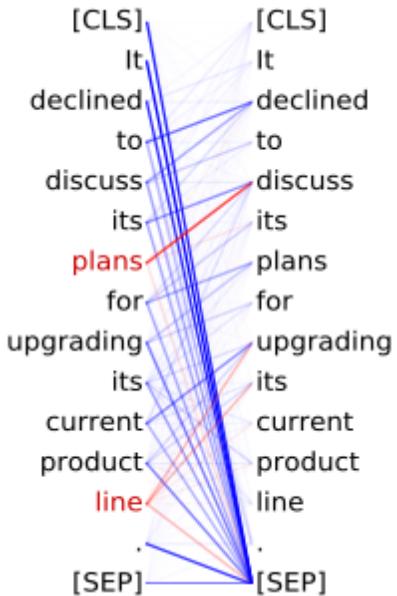
The complex financing plan in the S+L bailout law includes raising \$ 30 billion from debt issued by the newly created RTC .

ATTENTION IS ALL YOU NEED

TRANSFORMER:

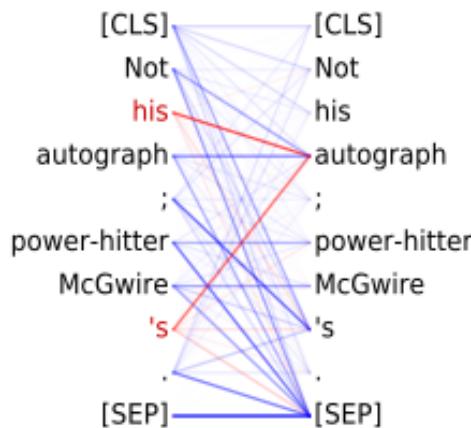


WHAT DOES ATTENTION LOOK AT?



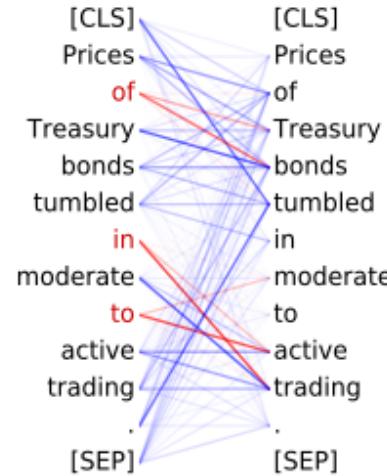
DOBJ

86.8%



POSS

80.5%



POBJ

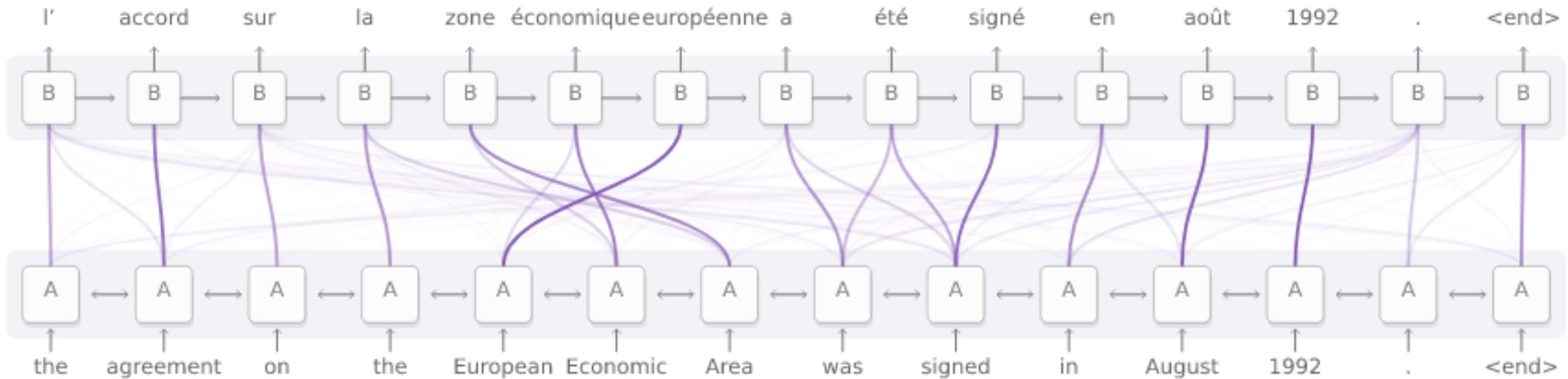
76.3%



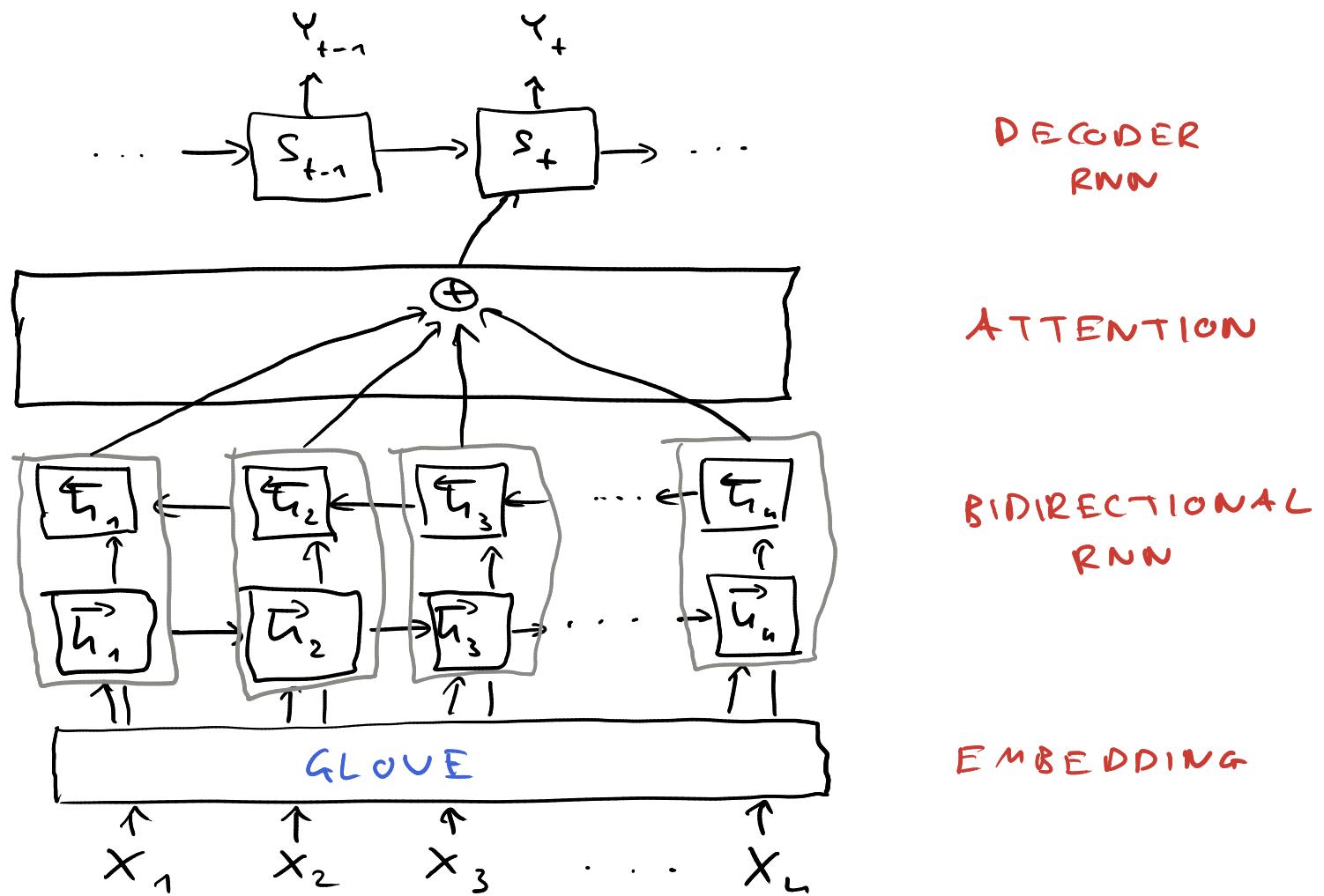
CORPUS

65.1%

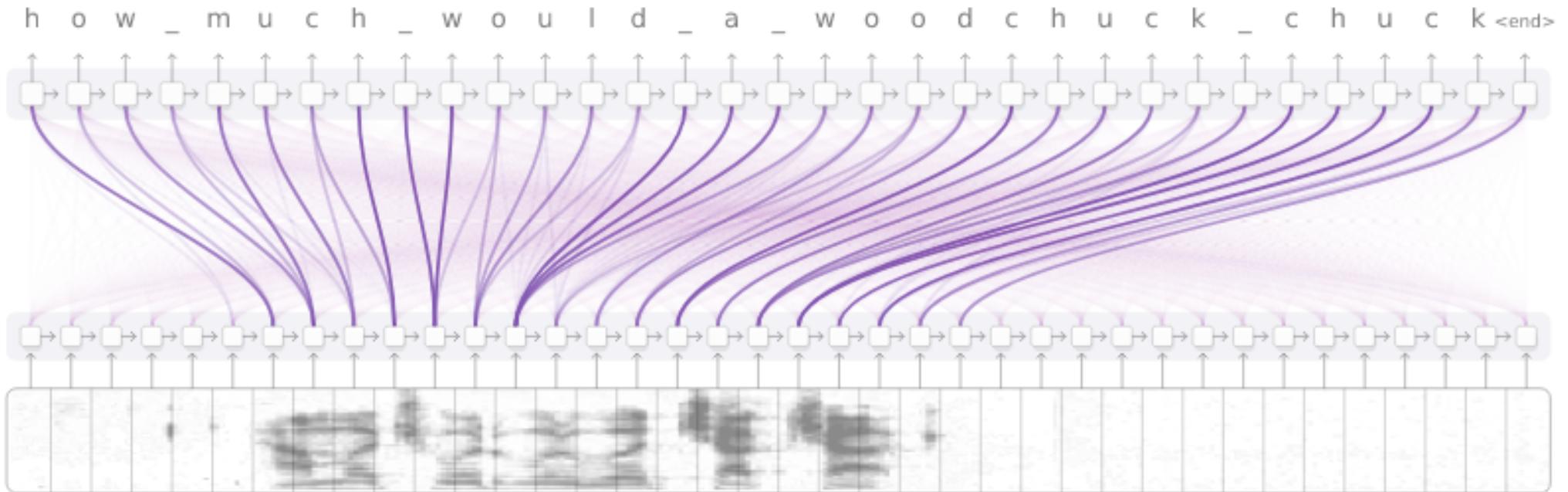
ATTENTION FOR SEQUENCE PREDICTION



TYPICAL ARCHITECTURE



WAV2VEC



DIALOGUE SYSTEMS

User: How are things?

Meena: They're good. How about you?

User: I'm trying to decide what to watch

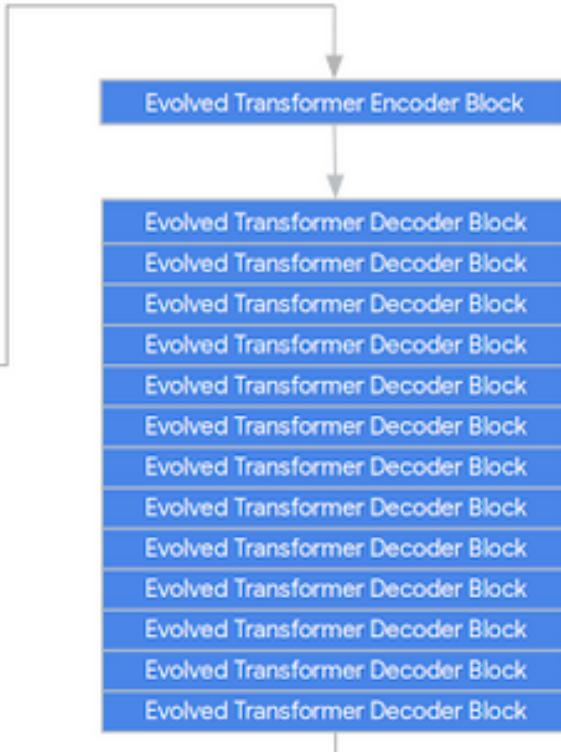
Meena: What are some of your favorite shows?

User: I like the Good Place

Meena: I haven't watched that. My favorite shows are Supernatural, and Star Trek: TNG.

User: What is TNG?

Meena: The Next Generation



REFERENCES

INTRODUCTORY:

- Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, 14(10), e12625. <https://doi.org/10.1111/phc3.12625>
- Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, 15(4), 233.
- Liquet, B., Moka, S., Nazarathy, Y. (2022). The Mathematical Engineering of Deep Learning (draft): Introduction. <https://deeplearningmath.org/drafts/chap1.pdf>

GENERAL:

- Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.
- Murphy, K. P. (2012). Machine learning: A probabilistic perspective. MIT Press.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
<http://www.deeplearningbook.org>

REFERENCES

OPTIMIZATION & LEARNING:

- Kochenderfer, M. J., & Wheeler, T. A. (2019). Algorithms for Optimization. MIT Press.
<https://algorithmsbook.com/optimization/>
- LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., & Huang, F. J. (2006). A Tutorial on Energy-Based Learning. <http://yann.lecun.org/exdb/publis/pdf/lecun-06.pdf>