

From Basic Machine Learning Models to Advanced Kernel Learning



Scott Pesme
Post-doc Inria Grenoble



Michaël Arbel
Researcher Inria Grenoble



Julien Mairal
Researcher Inria Grenoble

We work on the “theoretical aspects” of machine learning

Course structure



Scott Pesme
Post-doc Inria Grenoble



Michaël Arbel
Chercheur Inria Grenoble



Julien Mairal
Chercheur Inria Grenoble

**Basic Machine
Learning Models**

**Advanced Kernel
Learning**

2 x (~1h15 lecture) + 15 minutes break

Final grade: 50% mid-term + 50% final exam

What to expect?

- Understanding of basic ML methods:
 - *Linear regression, linear classification, unsupervised learning, neural networks*
 - *Optimisation methods (gradient descent)*
- A deeper dive into kernel methods

You will mostly need pre-requisites from linear algebra!

Introduction

A bit of AI history: the idea of thinking machines is not new!



Claude Shannon, 1961

youtube.com/watch?v=aygSMgK3BEM

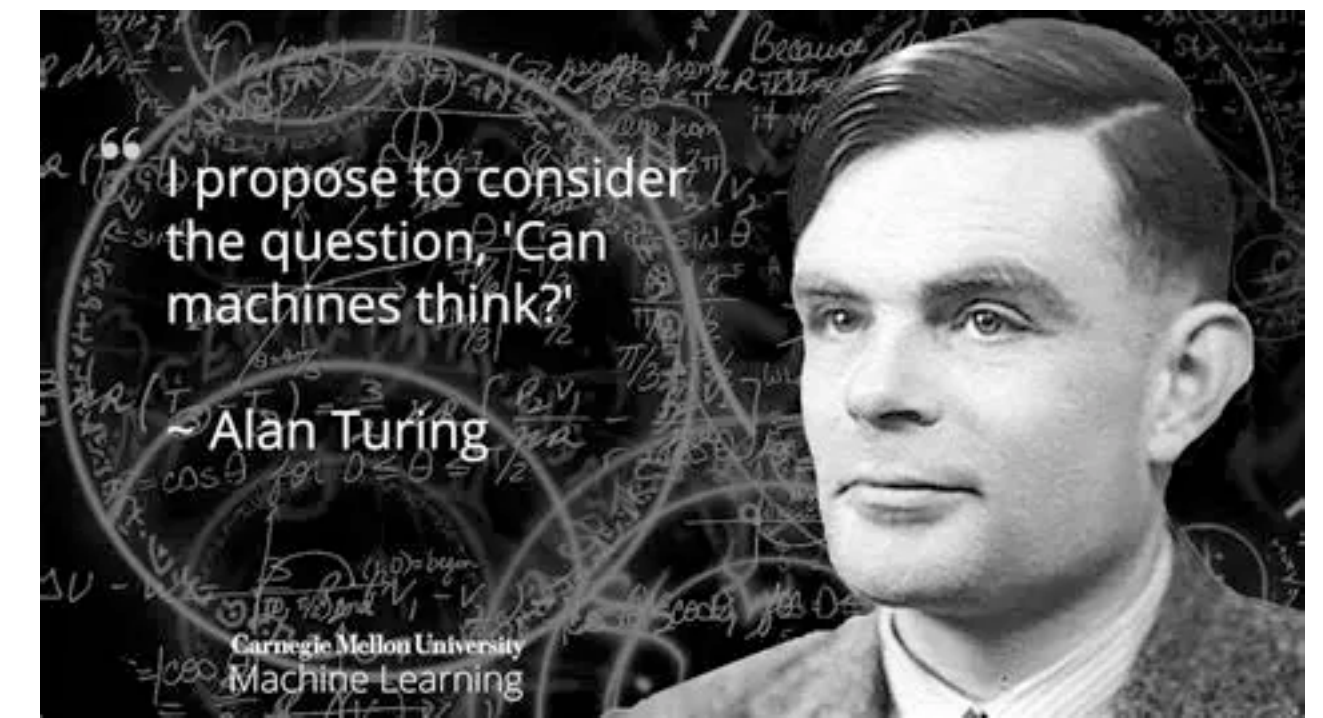
Journalist:

What's going to happen to us if machines can think? Can they?

Claude Shannon:

*I don't really know, but you come back in 4 or 5 years I'll probably say
"Sure they can think".*

A bit of AI history



Origins
Dartmouth Conference
Coined the term “Artificial intelligence”
“Birth of AI”

1940s - 1950s

1956

1960s - 1970s

1990s-2000s

2010s - 2020s

AI Winters

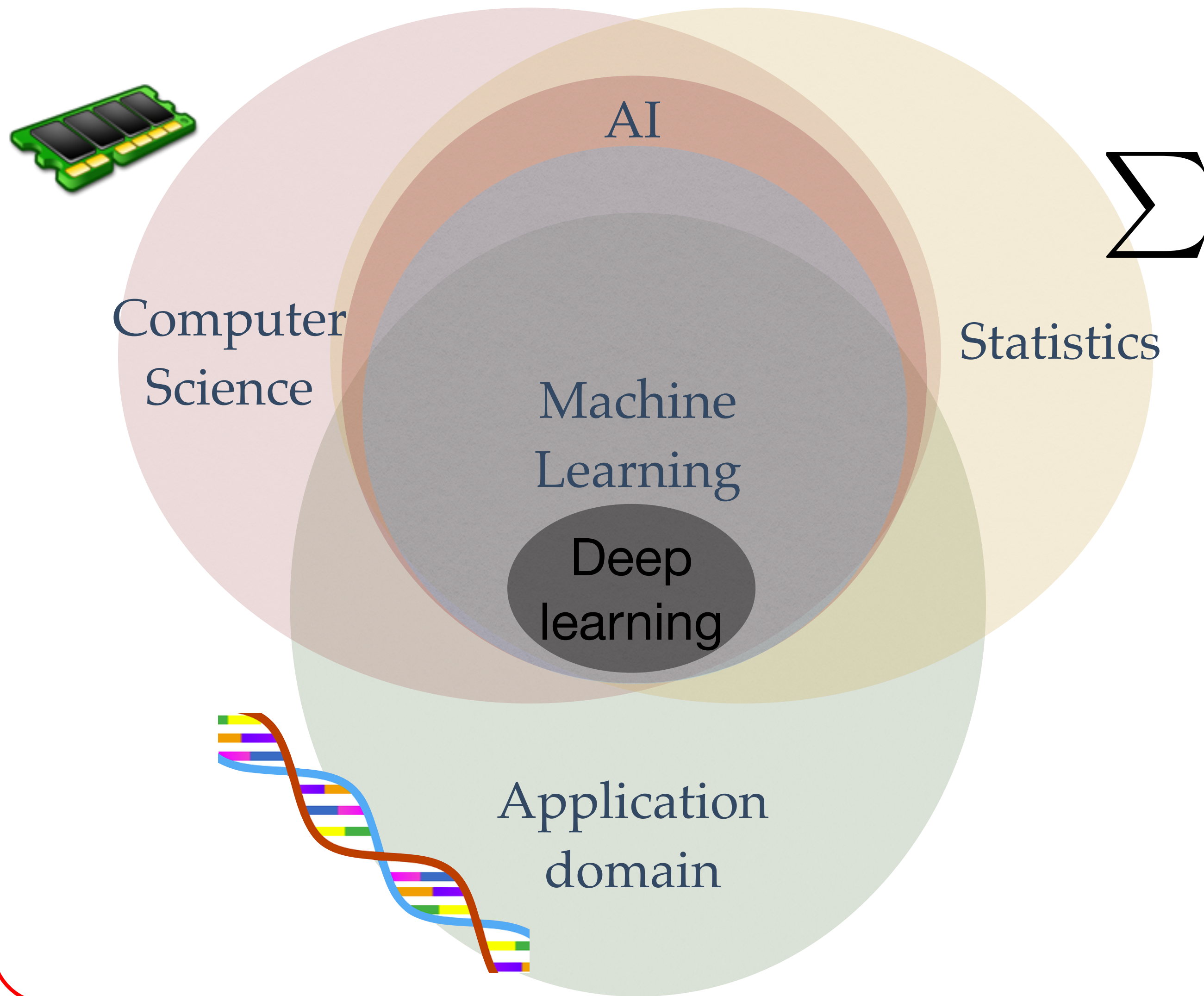
Theoretical foundations
Alan Turing: idea of “machine intelligence”
Turing test
Symbolic reasoning and neural networks

Expert systems

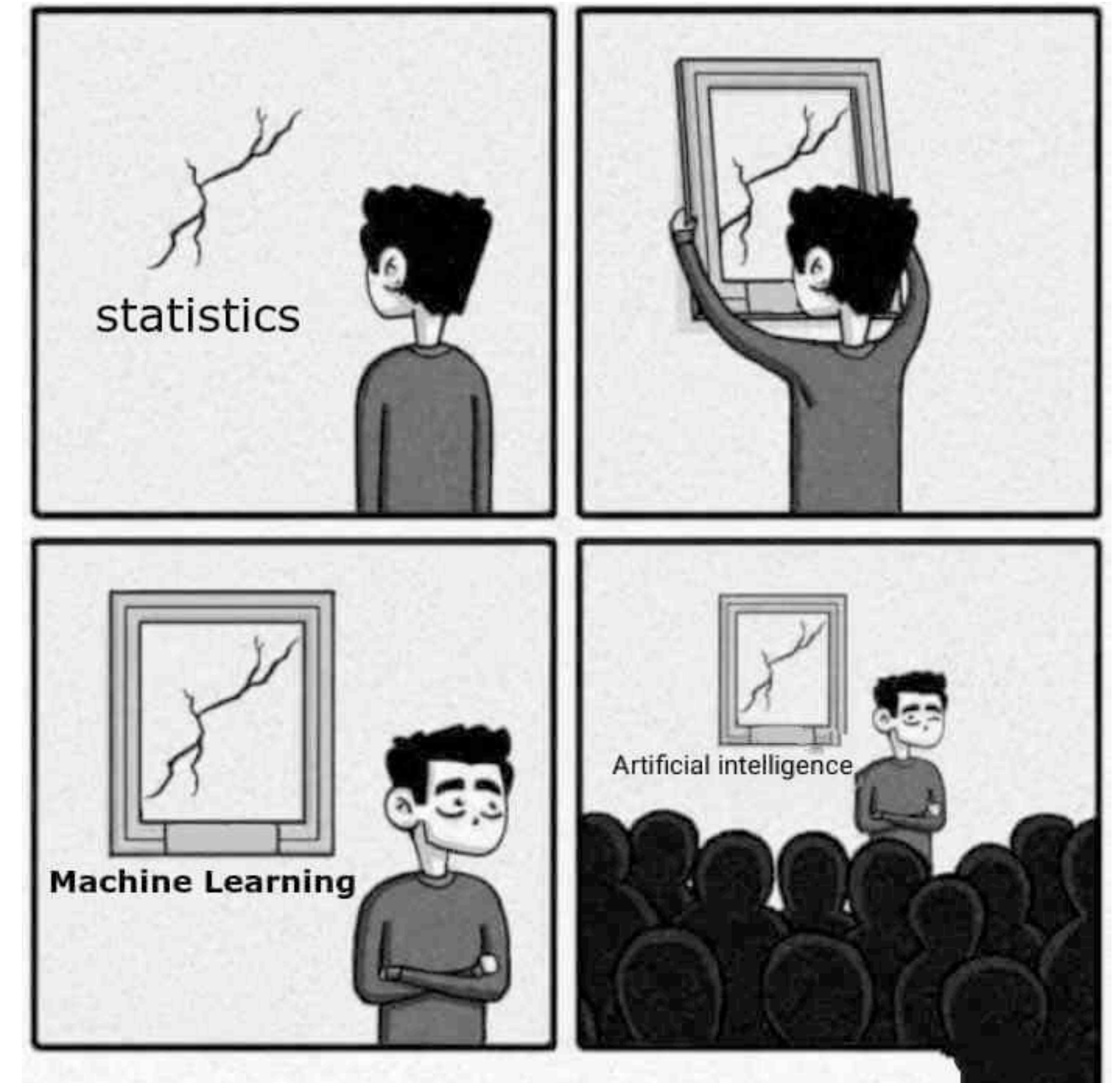
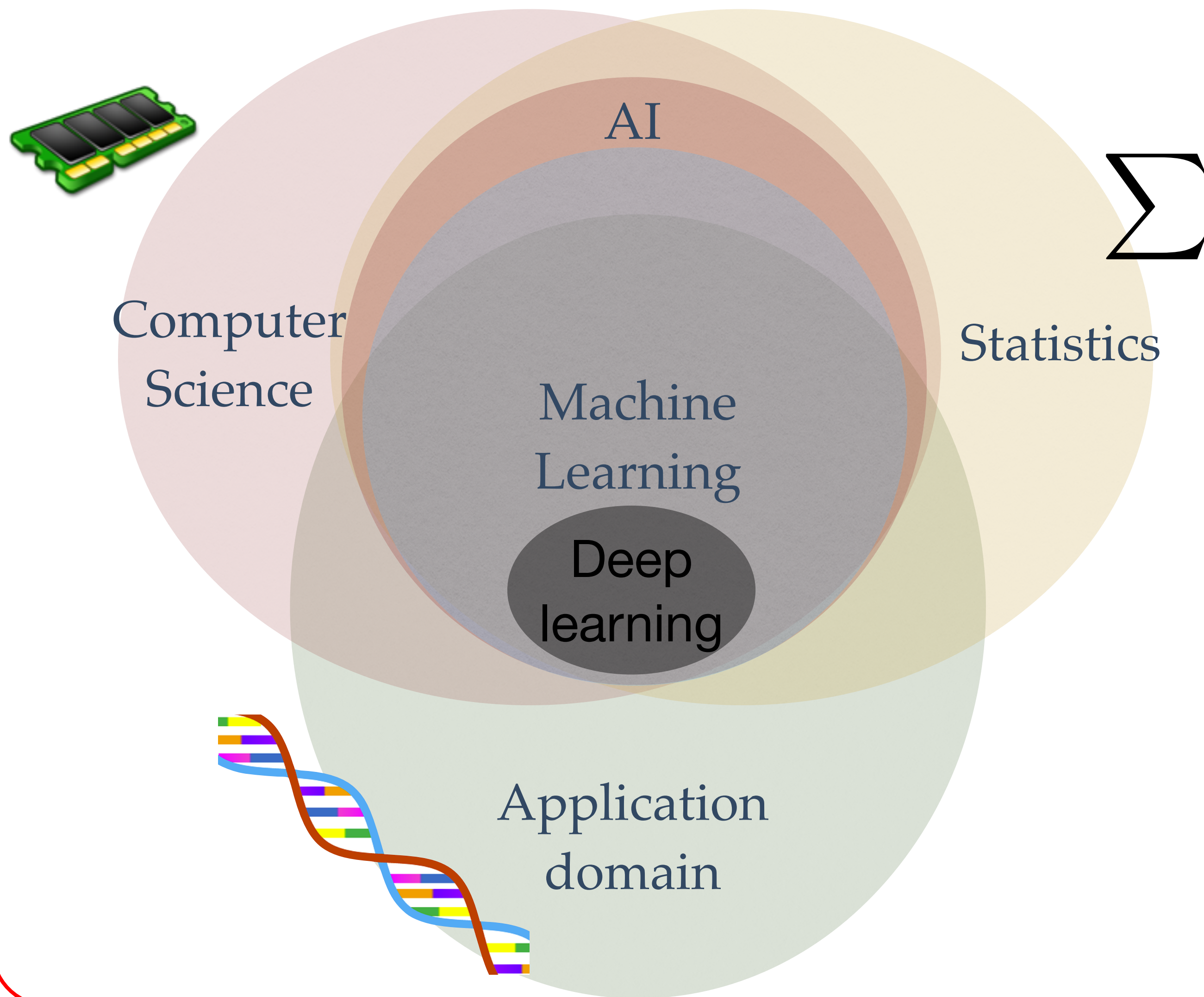
Practical breakthroughs
IBM’s Deep Blue (1996)
SVM, Reinforcement learning
Search engines, recommendation systems, logistics

Deep learning breakthrough
AlexNet (2012), AlphaGo (2016)
AlphaFold (2018), ChatGPT (2022)

Some vocabulary



Some vocabulary



Back to the introduction

What is Machine Learning?

Field which develops
“machines” / “algorithms” / “functions”
that learn from data

How does it work?

Training phase:

Neural network at school

Data



How does it work?

Training phase:

Neural network at school

Data



How does it work?

Training phase:

Neural network at school

Data



‘Cow’



How does it work?

Training phase:

Neural network at school

Data



How does it work?

Training phase:

Neural network at school

Data



‘Dog’



How does it work?

Training phase:

Neural network at school

Data



How does it work?

Training phase:

Neural network at school

Data



‘Chicken’ 

How does it work?

Training phase:

Neural network at school

Data



How does it work?

Training phase:

Neural network at school

Data



‘Donkey’



How does it work?

Training phase:

Neural network at school

Data



‘Donkey’



After training:

Proficient neural network



How does it work?

Training phase:

Neural network at school

Data



‘Donkey’



After training:

Proficient neural network



How does it work?

Training phase:

Neural network at school

Data



‘Donkey’



After training:

Proficient neural network



‘Lagopède
Alpin’

How does it work?

Training phase:

Neural network at school

Data



‘Donkey’



After training:

Proficient neural network



‘Lagopède
Alpin’

It learns ‘on its own’ to extract the important features

Machine learning vs expert systems

After training:



Proficient neural network



‘Lagopède
Alpin’

It learns ‘on its own’ to extract the important features

Machine learning vs expert systems

After training:

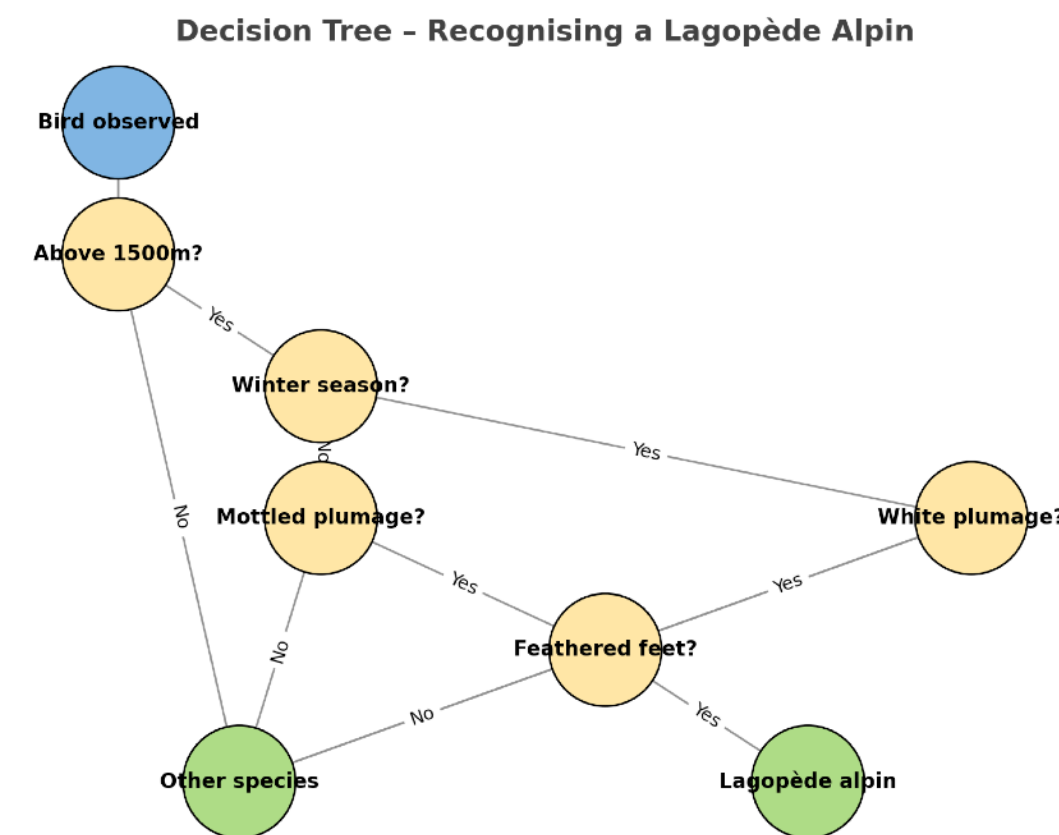


Proficient neural network



‘Lagopède
Alpin’

It learns ‘on its own’ to extract the important features



‘Lagopède
Alpin’

An “expert” gives a list of “if/then” rules

Pros and cons of two different “AI strategies”

Machine learning

Pros:

Cons:

Expert systems

Pros:

Cons:

Pros and cons of two different “AI strategies”

Machine learning

Pros:

- No need to find the rules
- Adapts to data
- Works very well!!

Cons:

- Opacity (what the heck is going on??)
- Needs a “lot” of data

Expert systems

Pros:

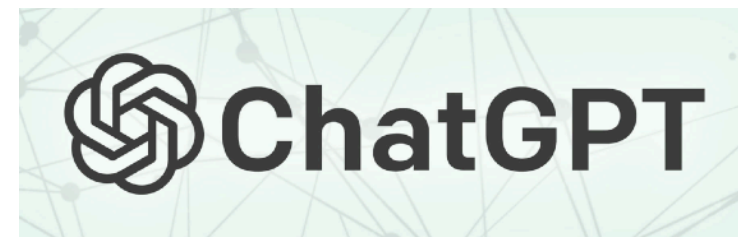
- Transparent and understandable

Cons:

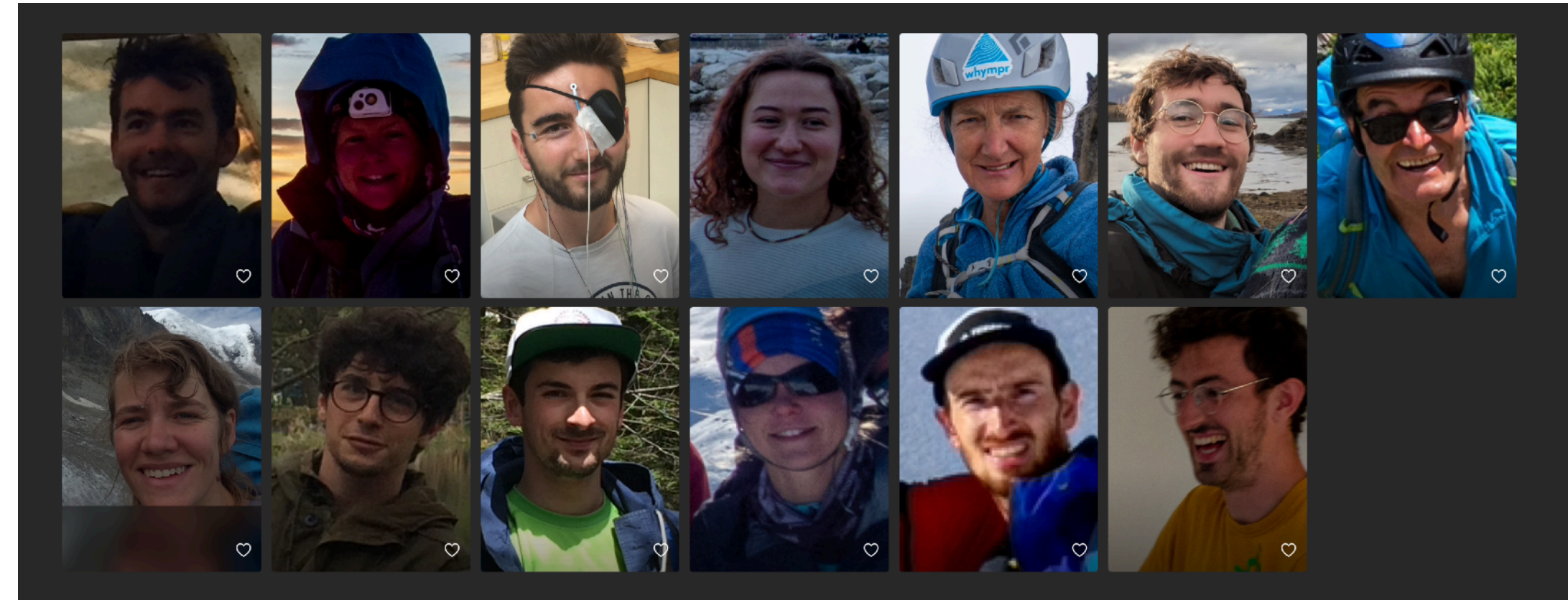
- Finding the rules is hard (impossible?)
- Doesn't work very well

Where is machine learning?

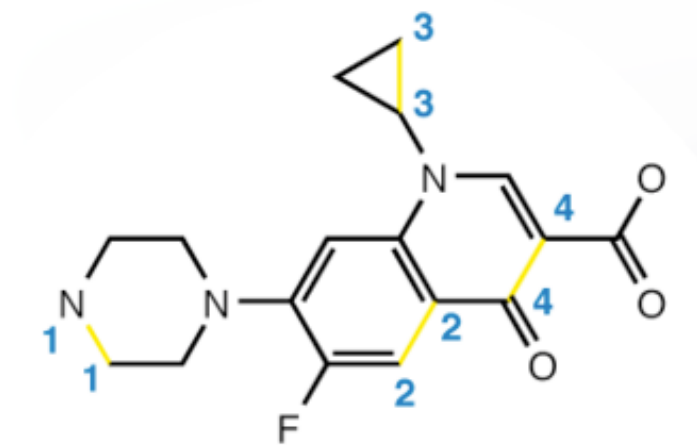
Text generation:



Apple's photo app:



Drug discovery:



PlantNet:



16 mai 2024



Robinia viscosa Michx. ex Vent.

Robinier visqueux

Fabaceae

DETAILS

↑ PARTAGER

Image generation:



But also:

Deep fakes

Scams

Mass surveillance



Learning a mapping / function from input to output

Input $x \in \mathbb{R}^d$

Output $y \in \mathbb{R}^p$

*Trainable
function*

$$f_w(x) = \hat{y}$$



Learning a mapping / function from input to output

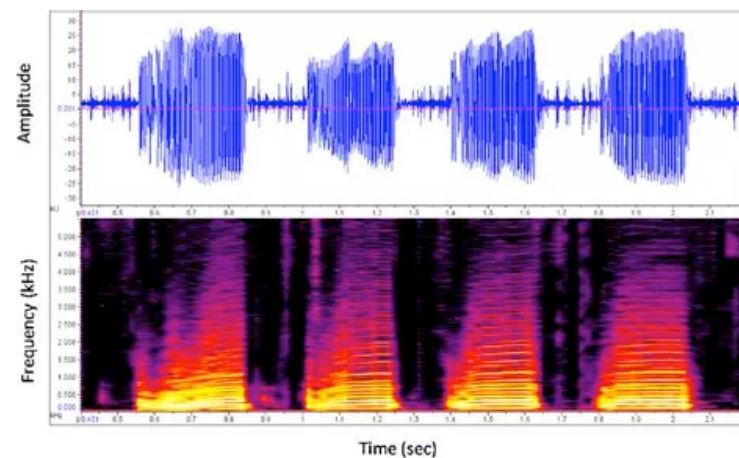
Input $x \in \mathbb{R}^d$

Text *What is the capital of France?*

Image



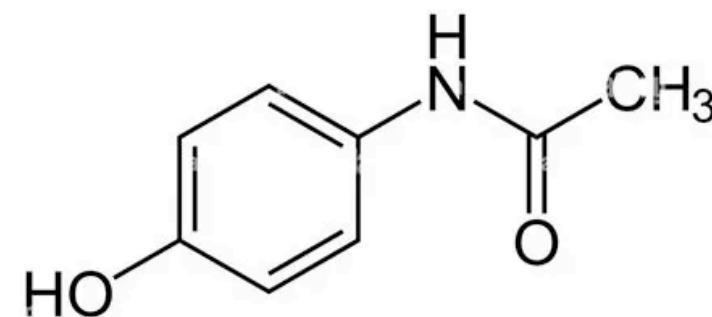
Audio



Video



Graph



“Feature” vector (Height, age, income)

Trainable function

$$\downarrow$$
$$f_w(x) = \hat{y}$$



Output $y \in \mathbb{R}^p$

Text

Image

Audio

Video

Graph

“Feature” vector

How is the input represented “inside the machine”?

Input

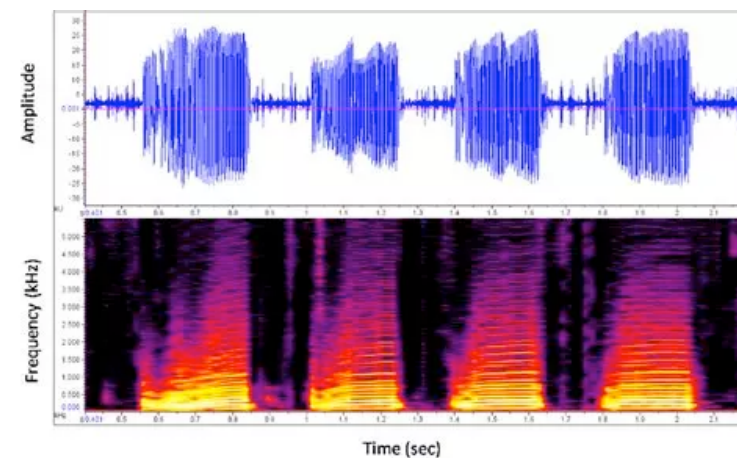
Text

*What is the capital
of France?*

Image



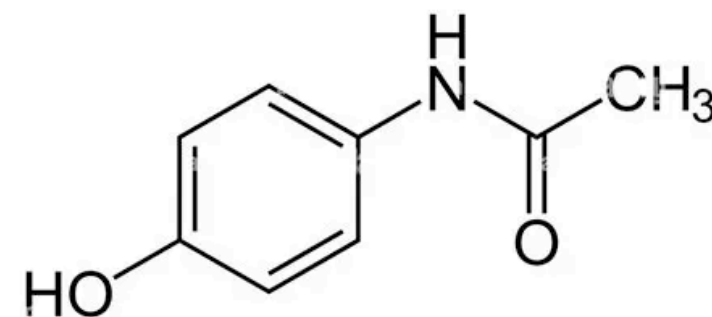
Audio



Video



Graph



“Feature” vector (Height, age, income)

“Digital” representation

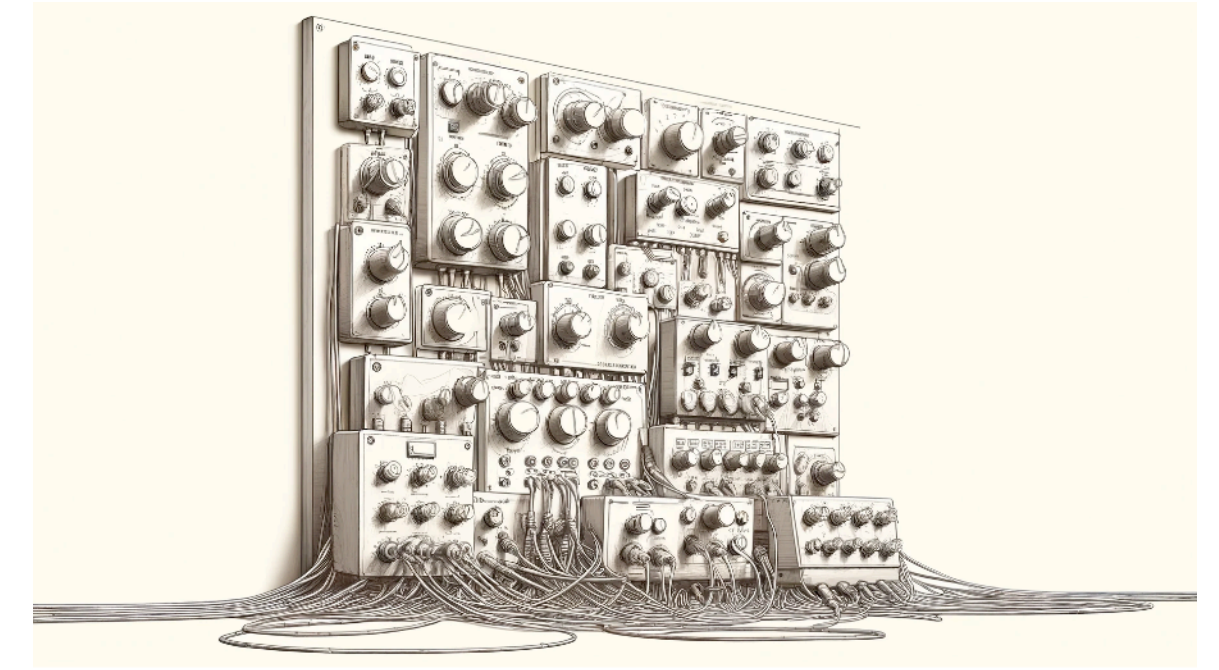
$$x = ???$$

How do we get a machine to learn?

The neural network



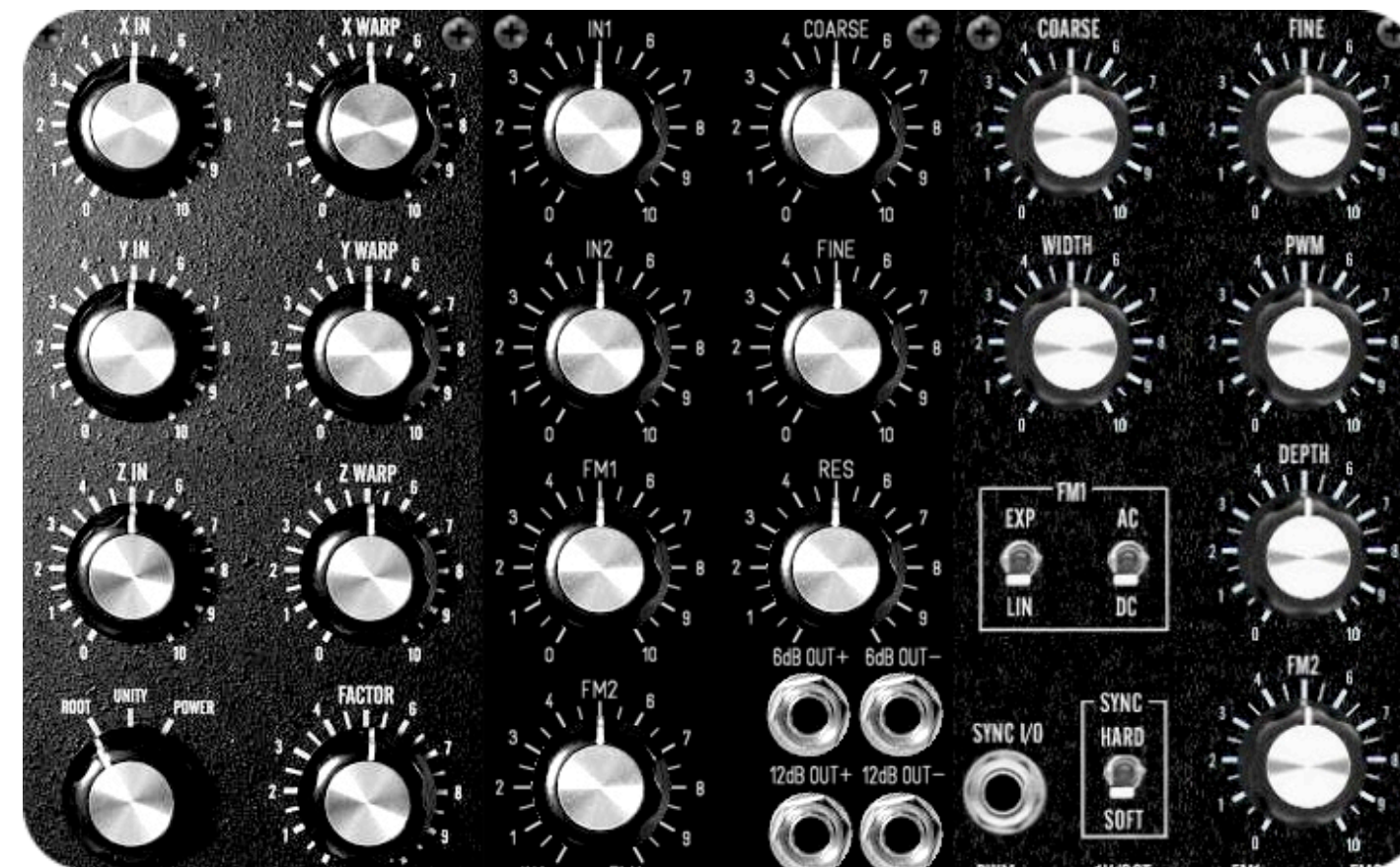
looks rather like:



Training:

Neural network at school

Data

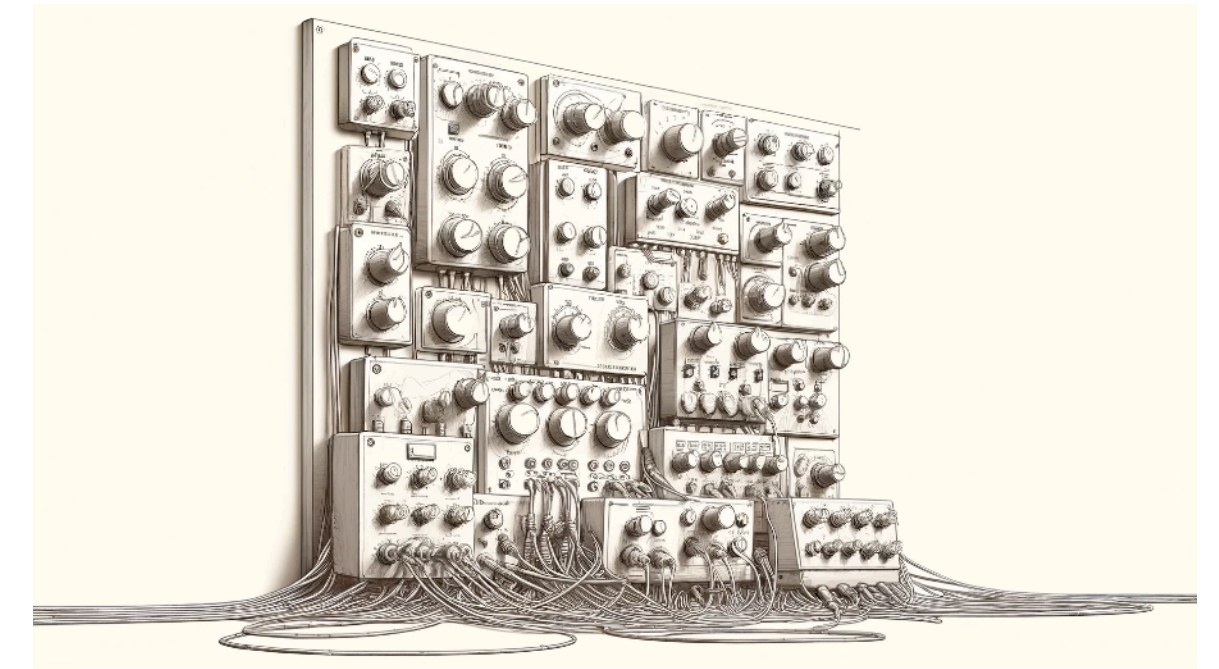


How do we get a machine to learn?

The neural network



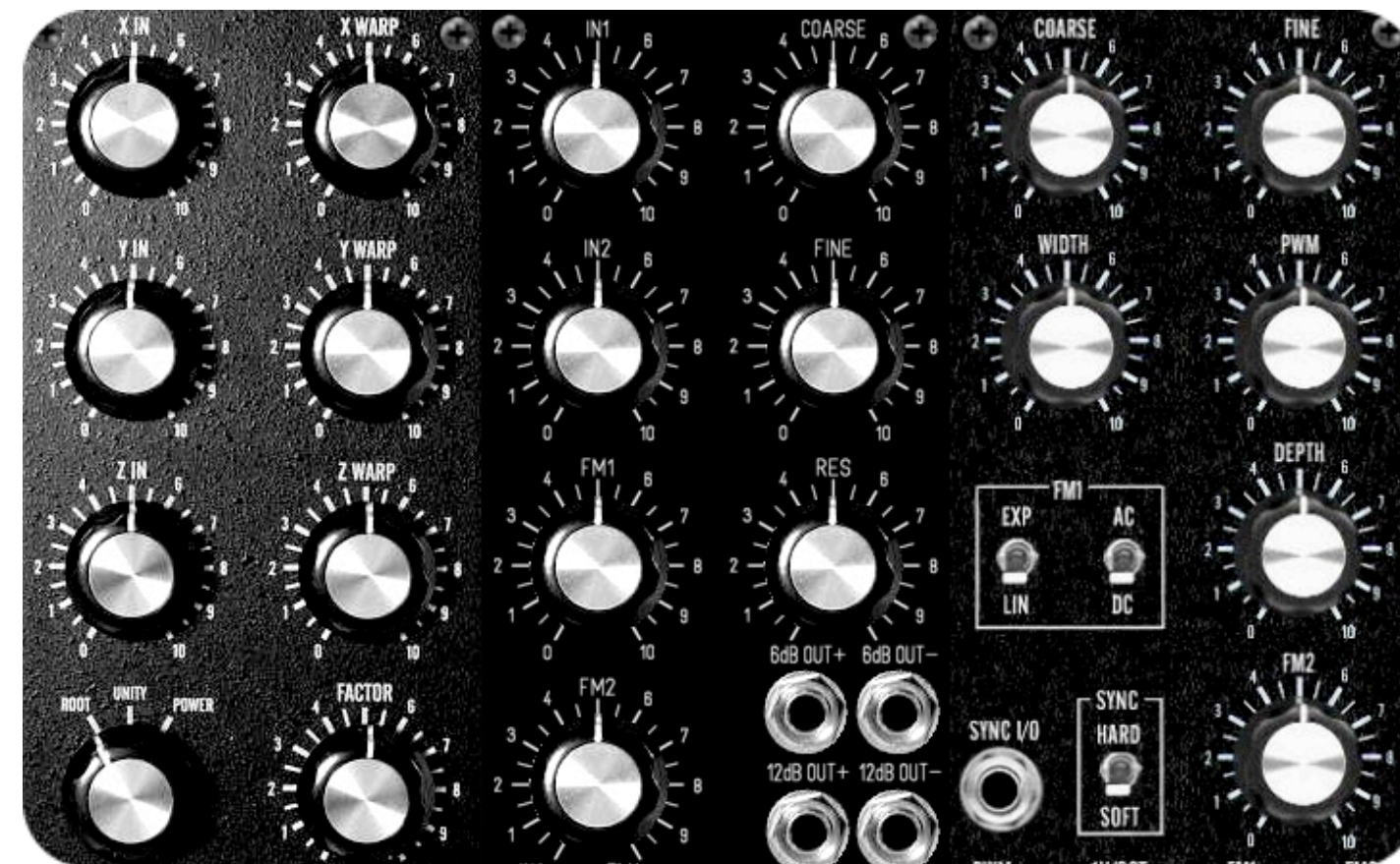
looks rather like:



Training:

Neural network at school

Data

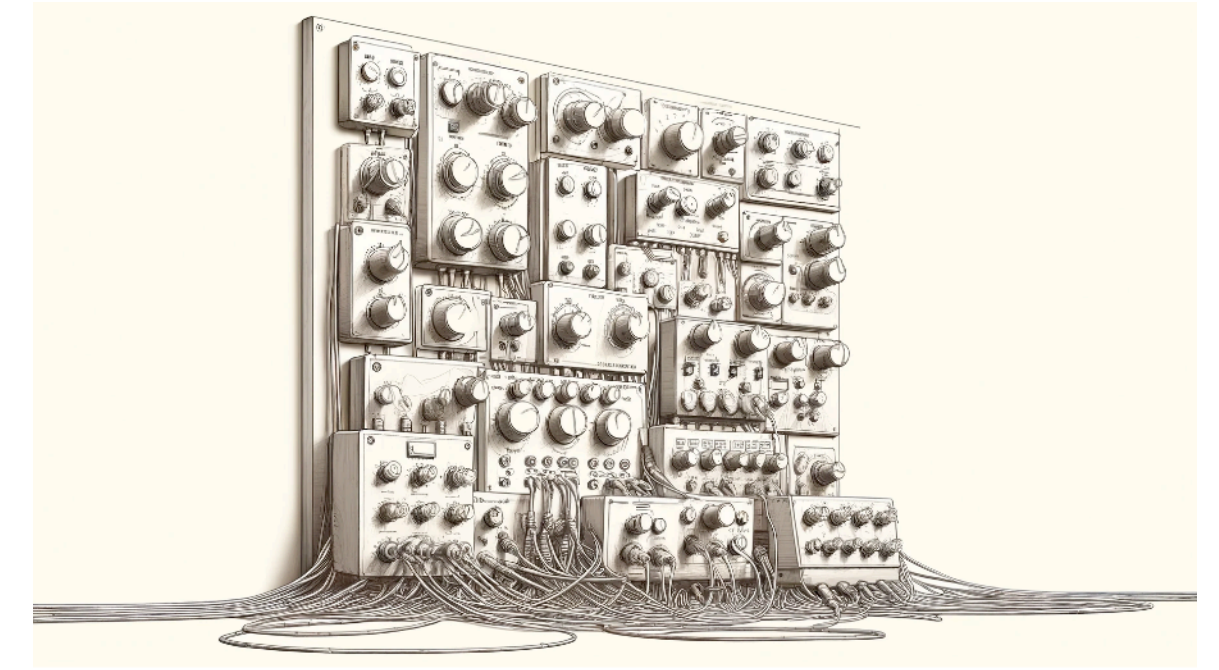


How do we get a machine to learn?

The neural network



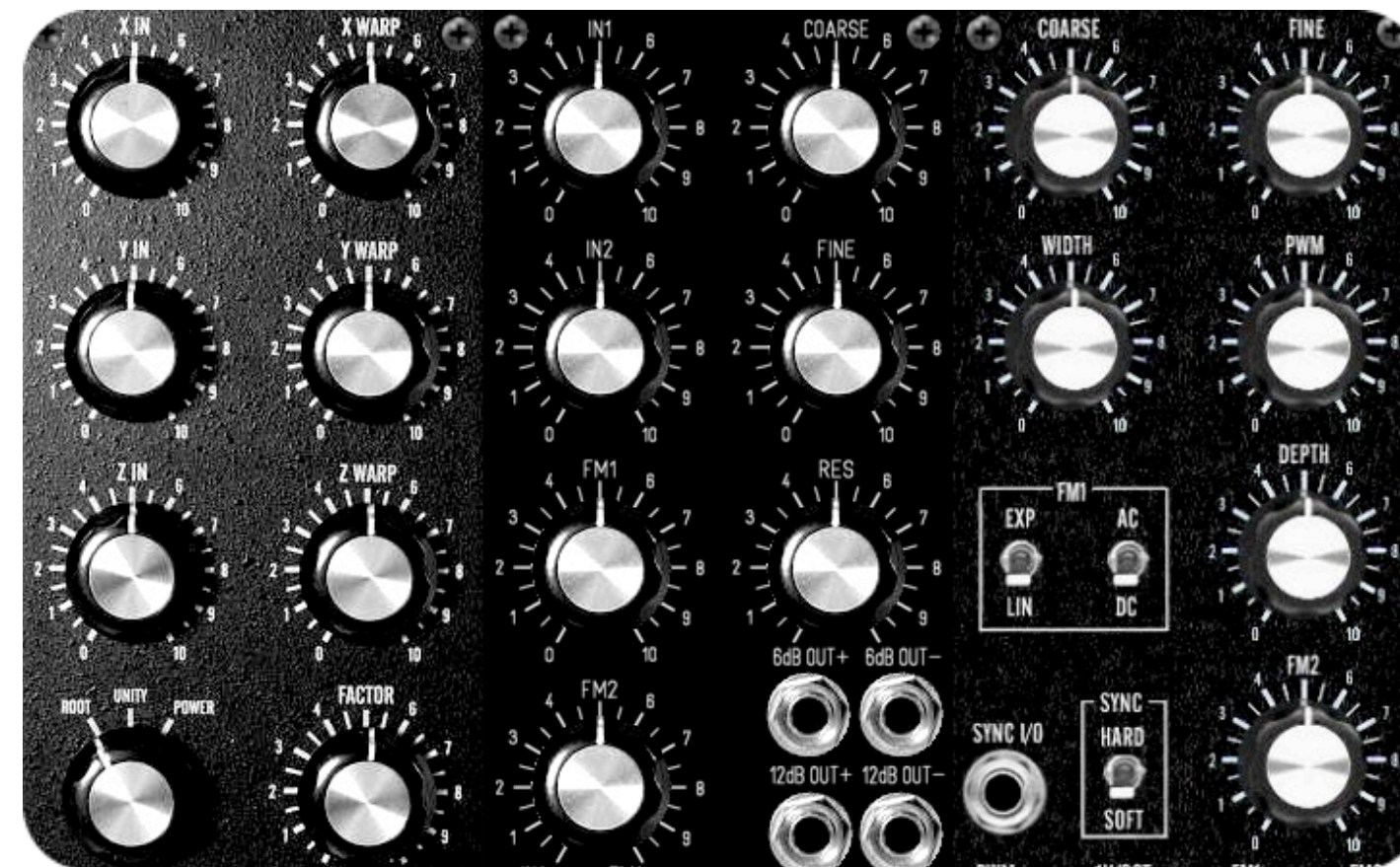
looks rather like:



Training:

Neural network at school

Data



‘Vache’

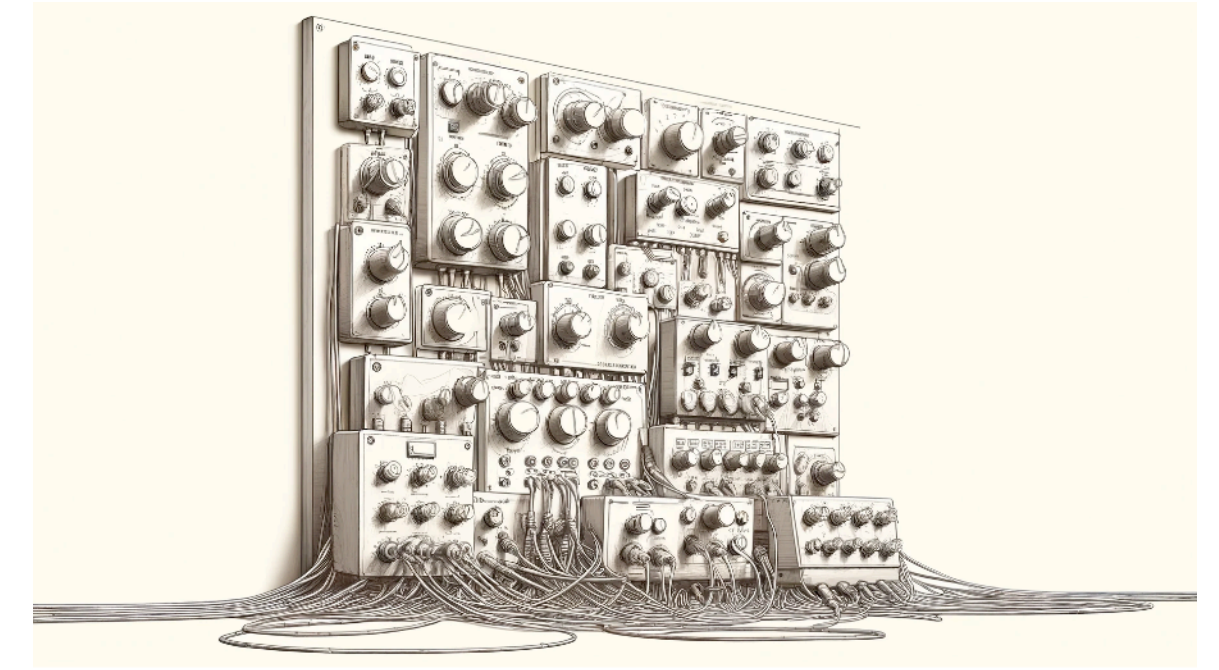


How do we get a machine to learn?

The neural network



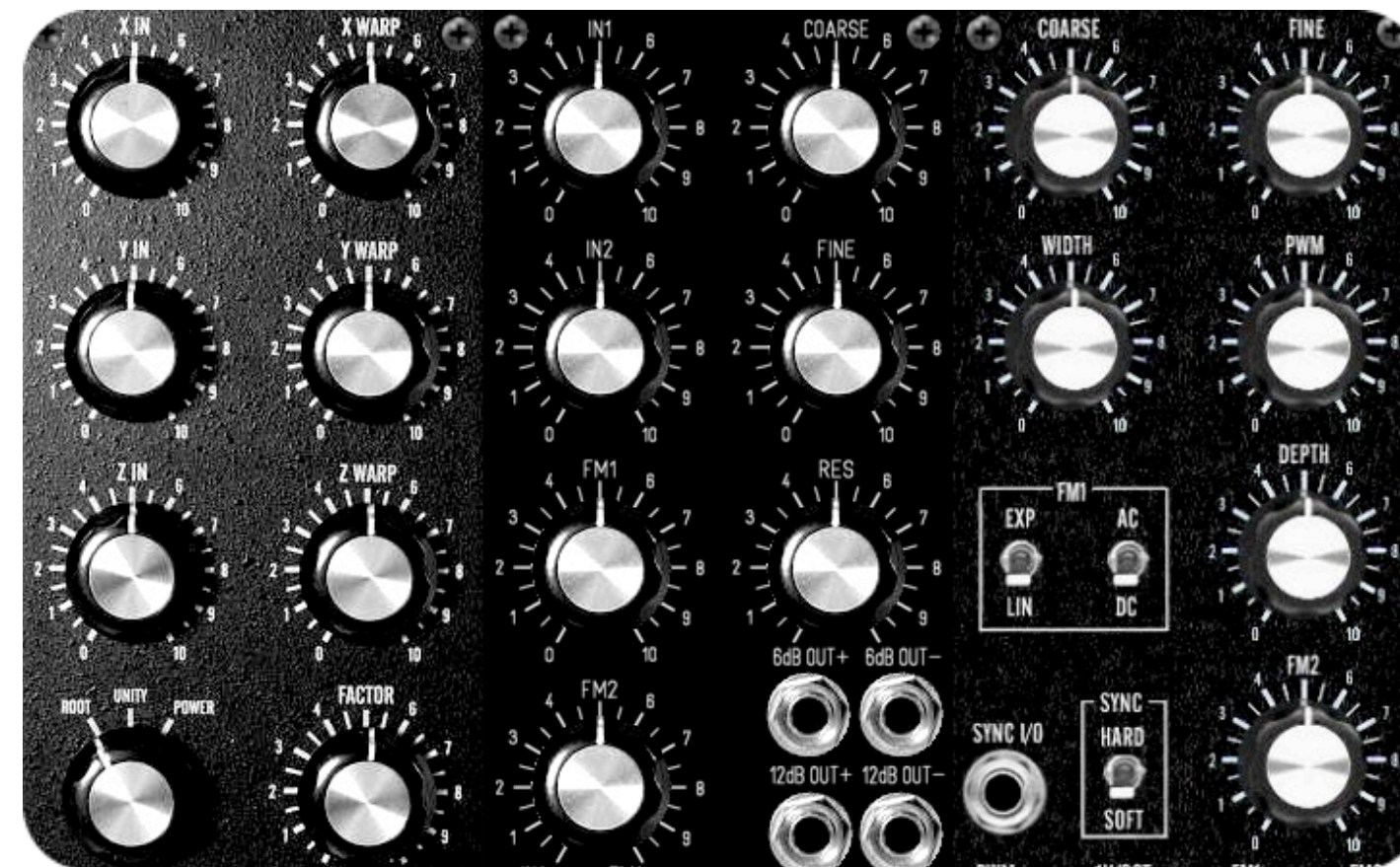
looks rather like:



Training:

Neural network at school

Data

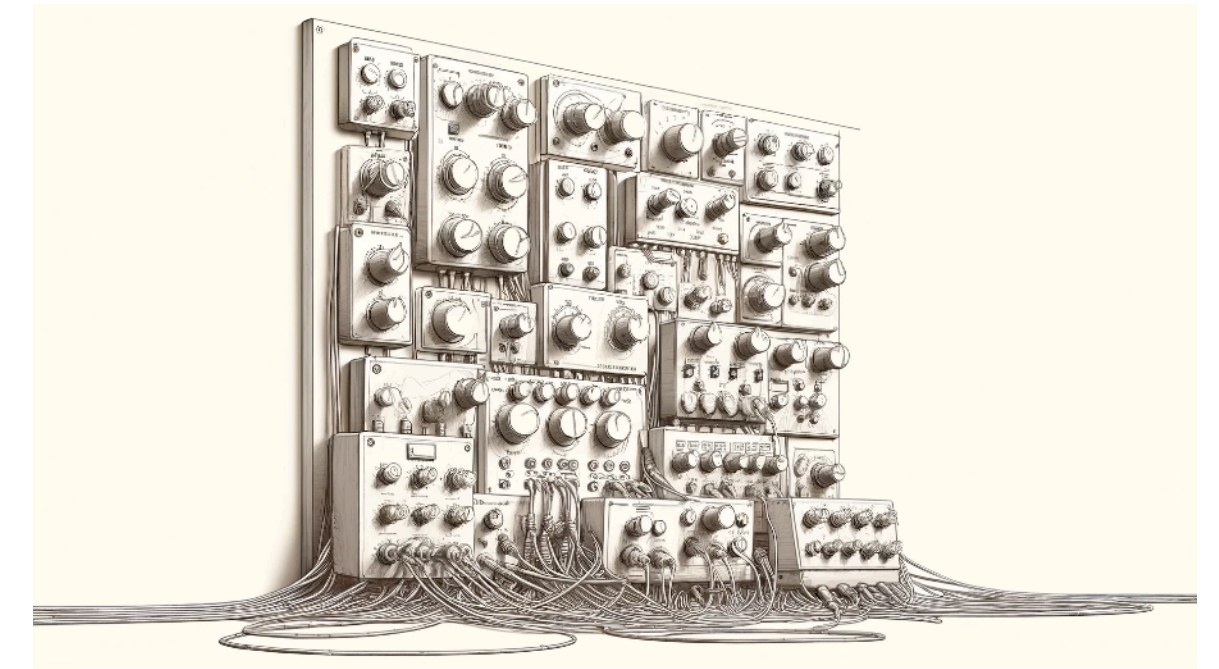


How do we get a machine to learn?

The neural network



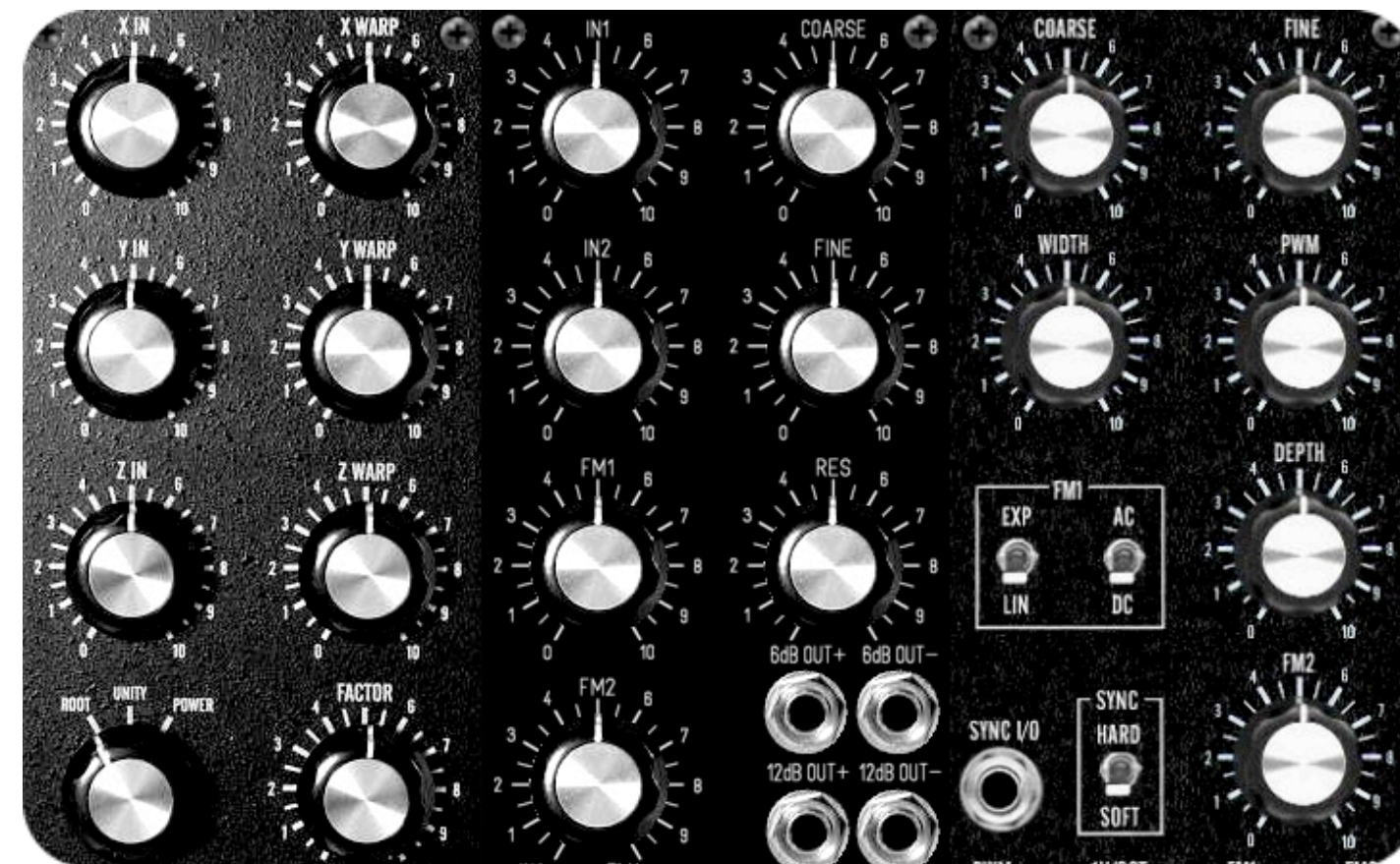
looks rather like:



Training:

Neural network at school

Data



‘Chien’

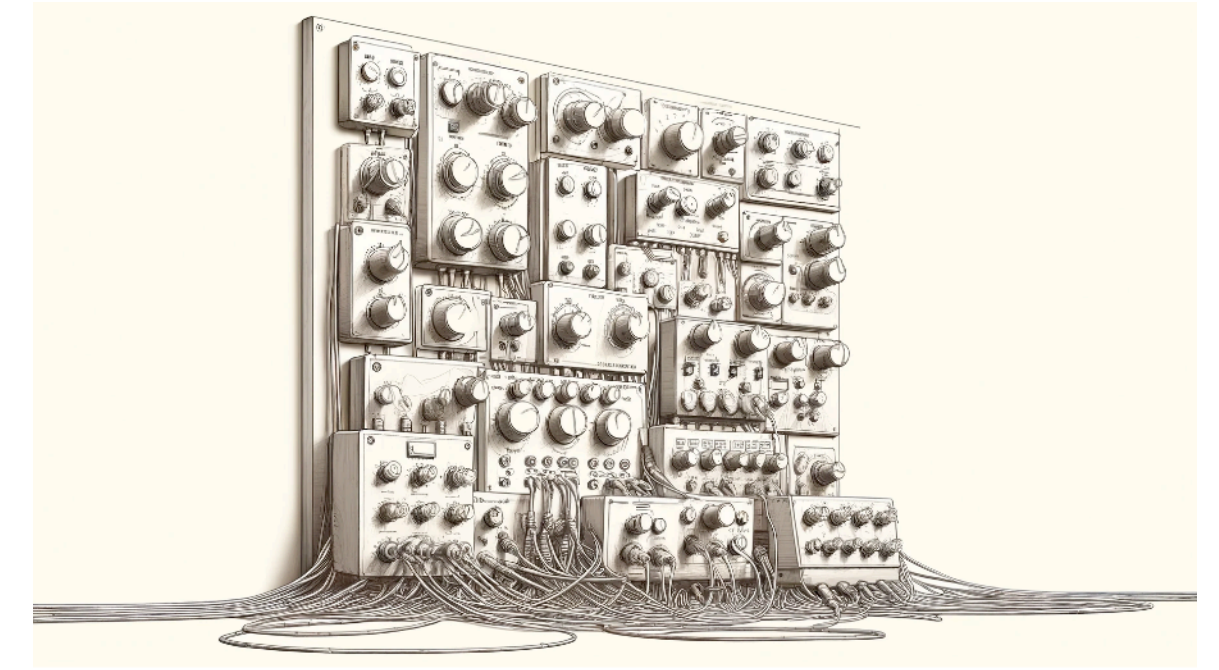


How do we get a machine to learn?

The neural network



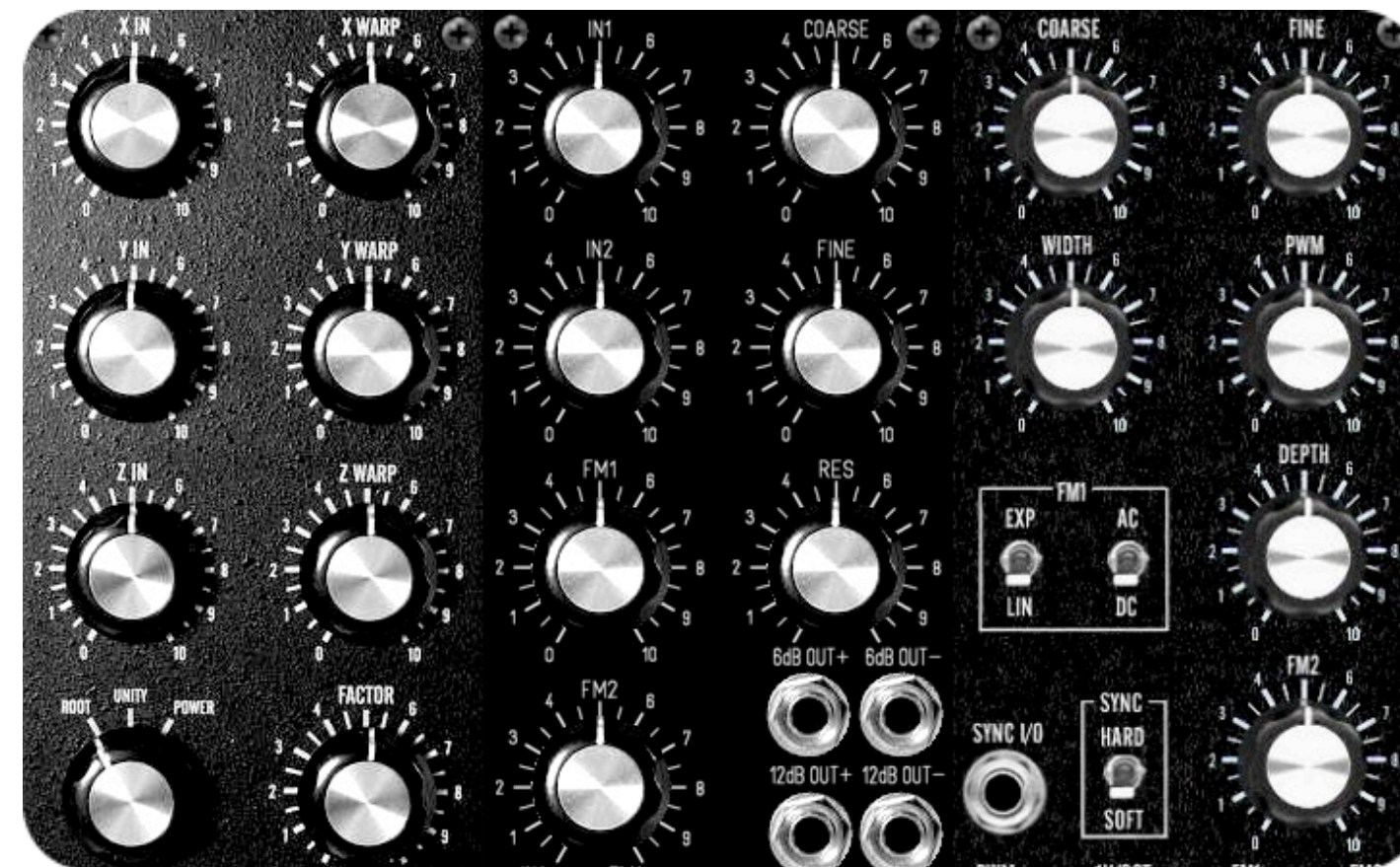
looks rather like:



Training:

Neural network at school

Data

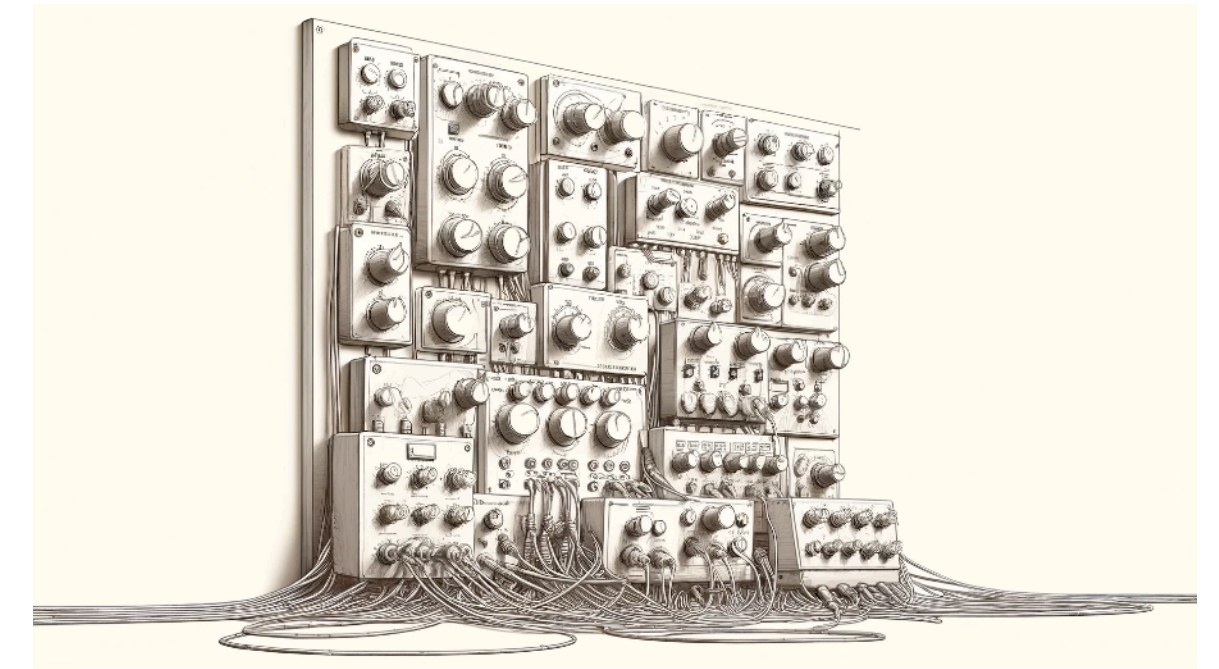


How do we get a machine to learn?

The neural network



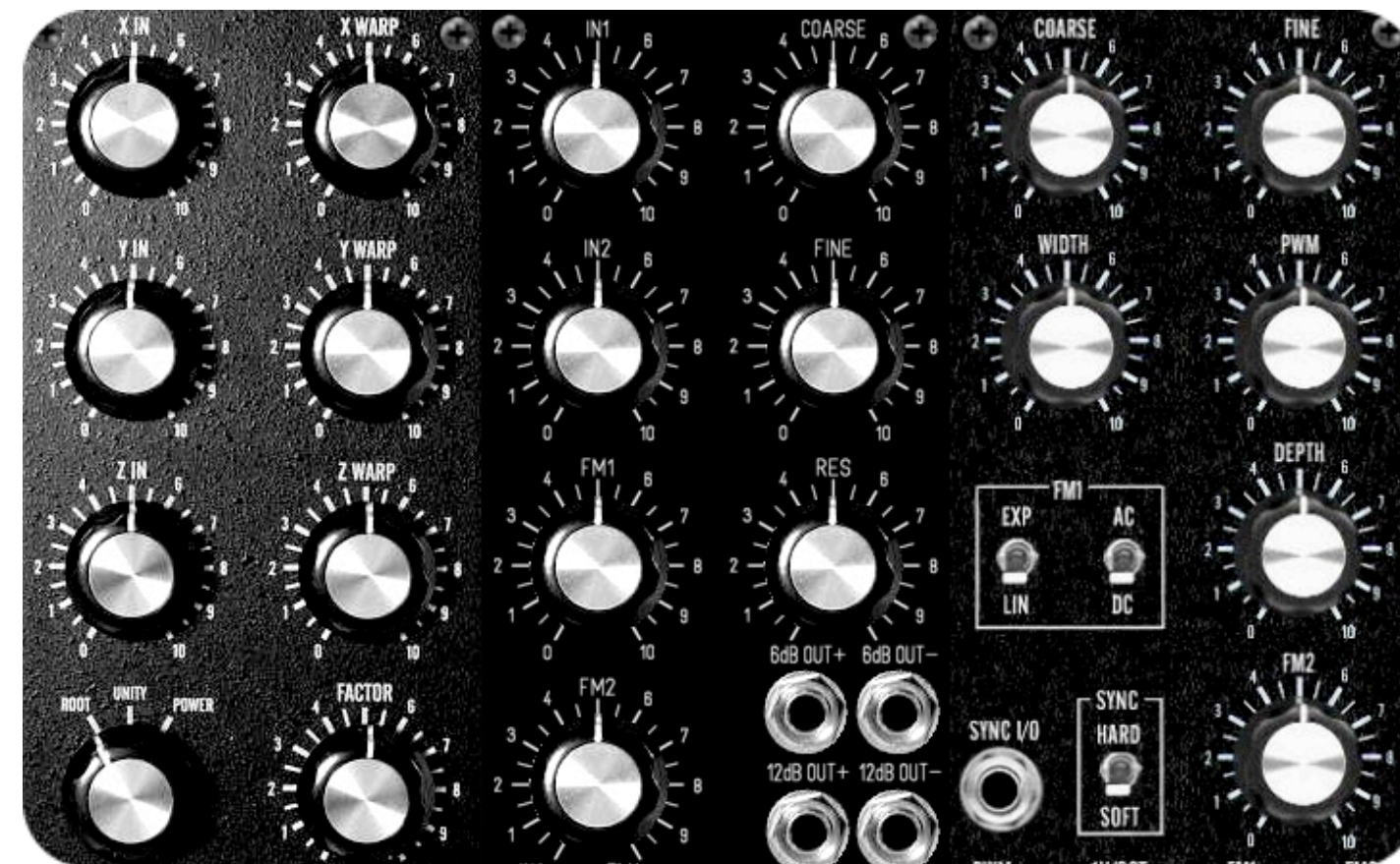
looks rather like:



Training:

Neural network at school

Data



‘Poule’

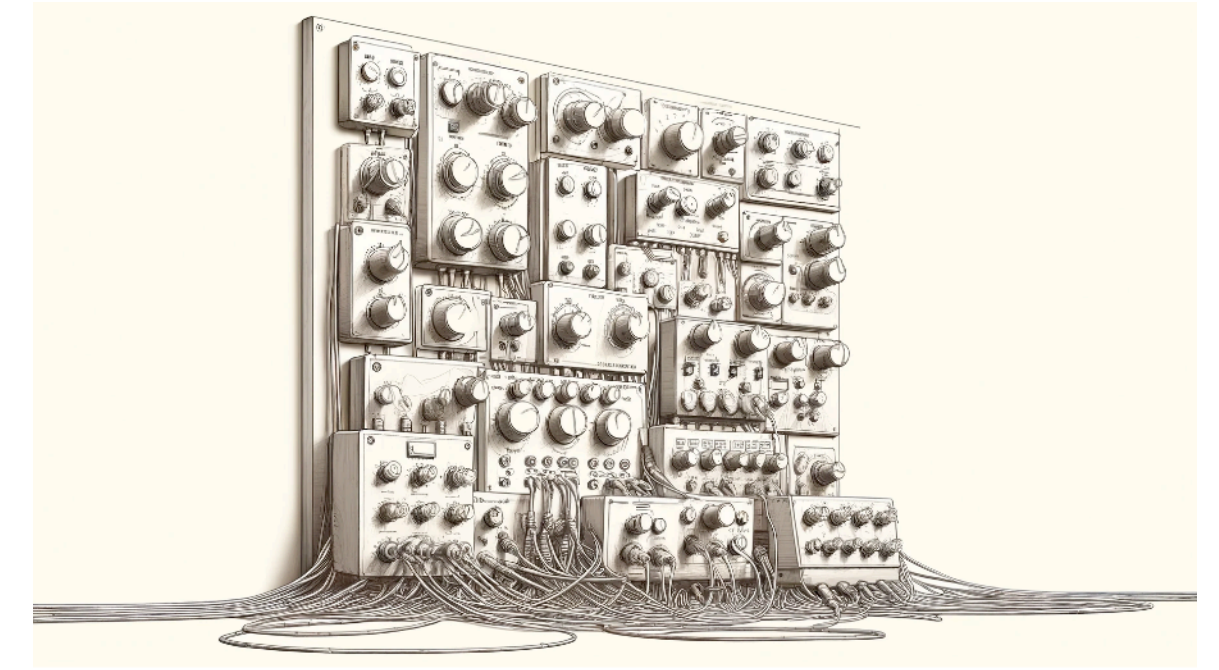


How do we get a machine to learn?

The neural network



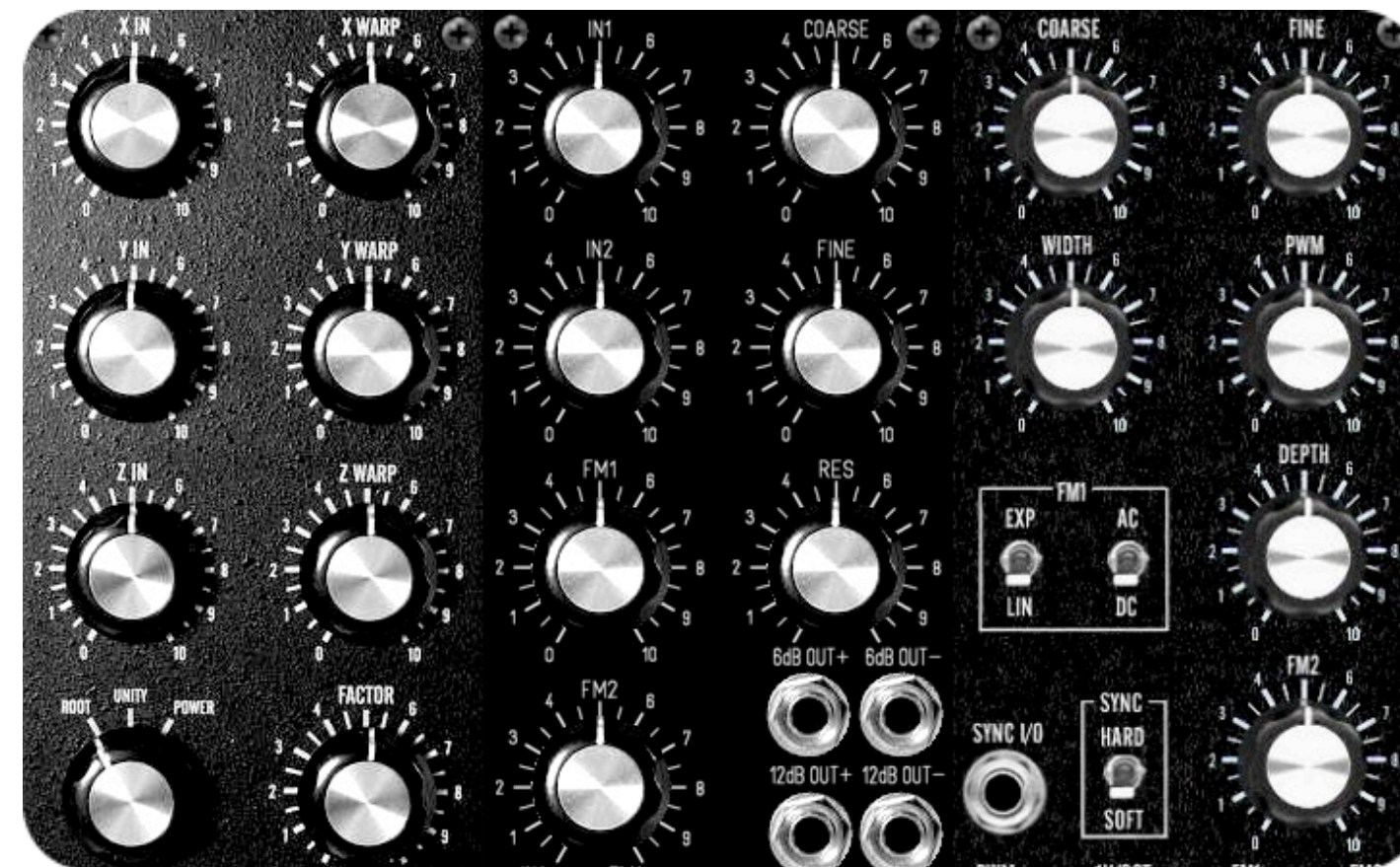
looks rather like:



Training:

Neural network at school

Data

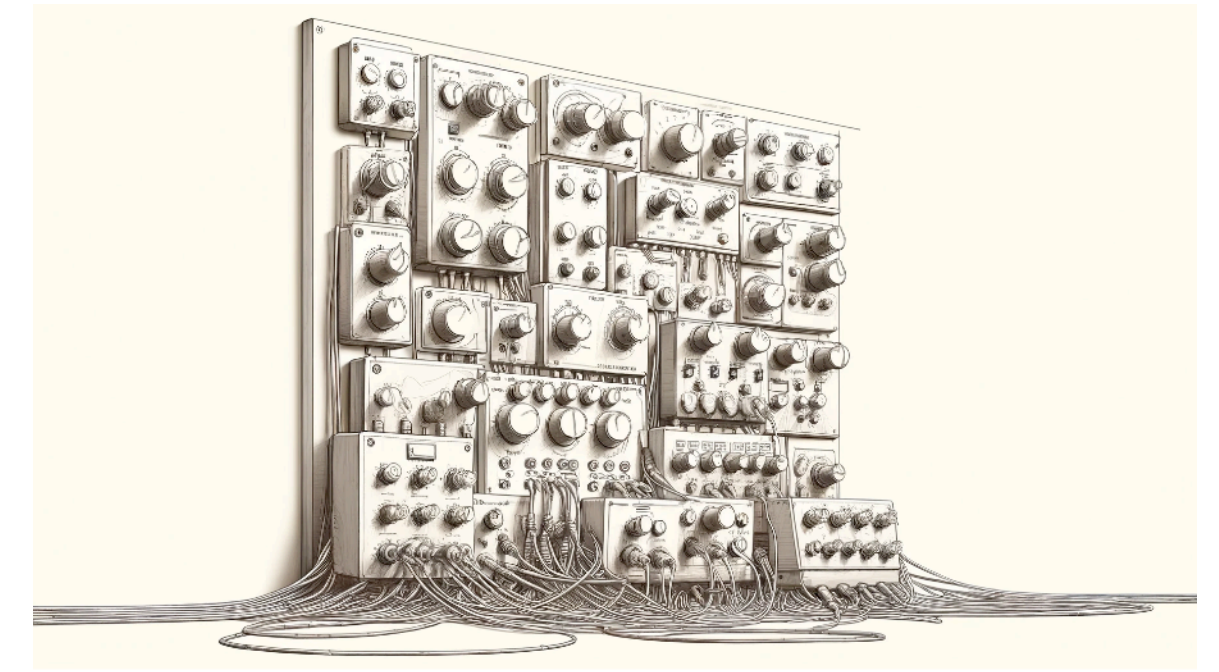


How do we get a machine to learn?

The neural network



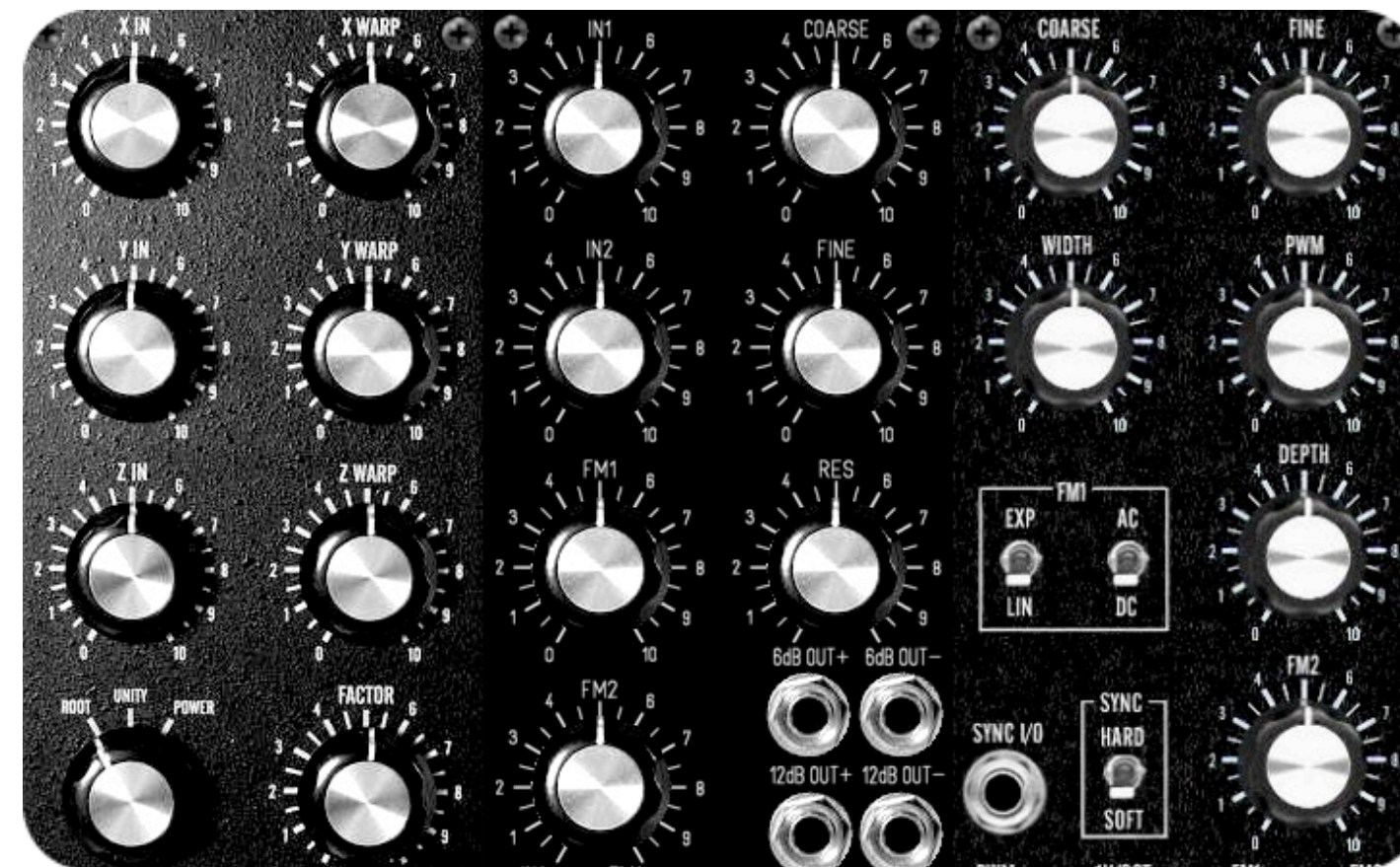
looks rather like:



Training:

Neural network at school

Data



‘Âne’

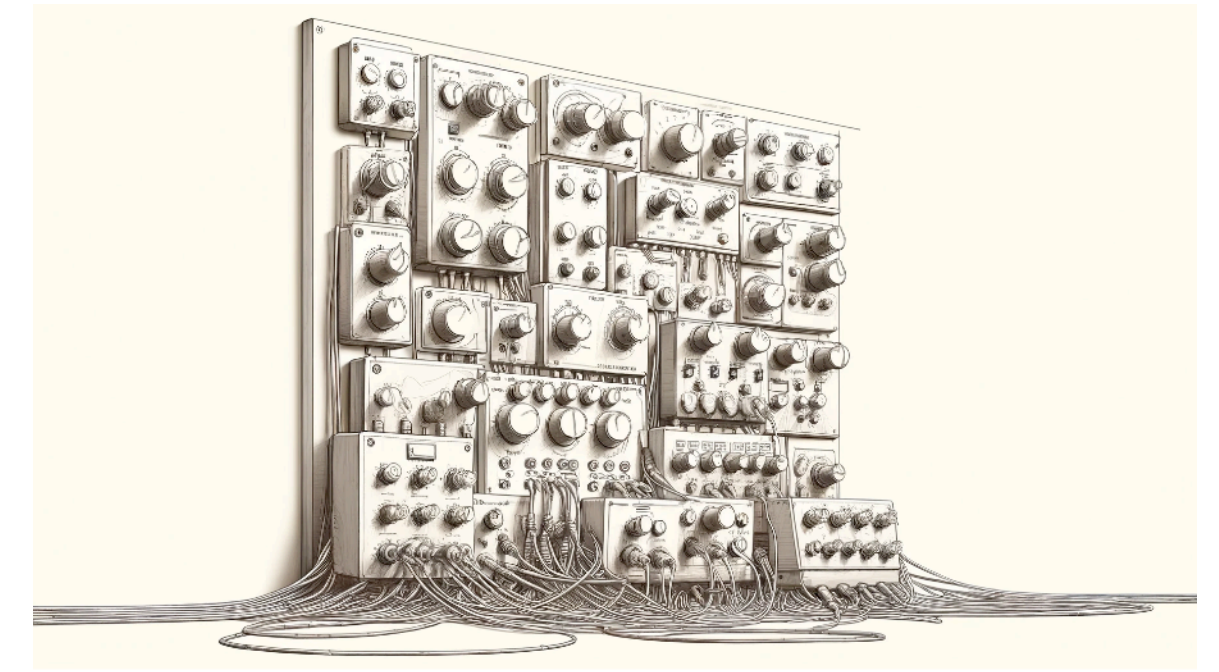


How do we get a machine to learn?

The neural network



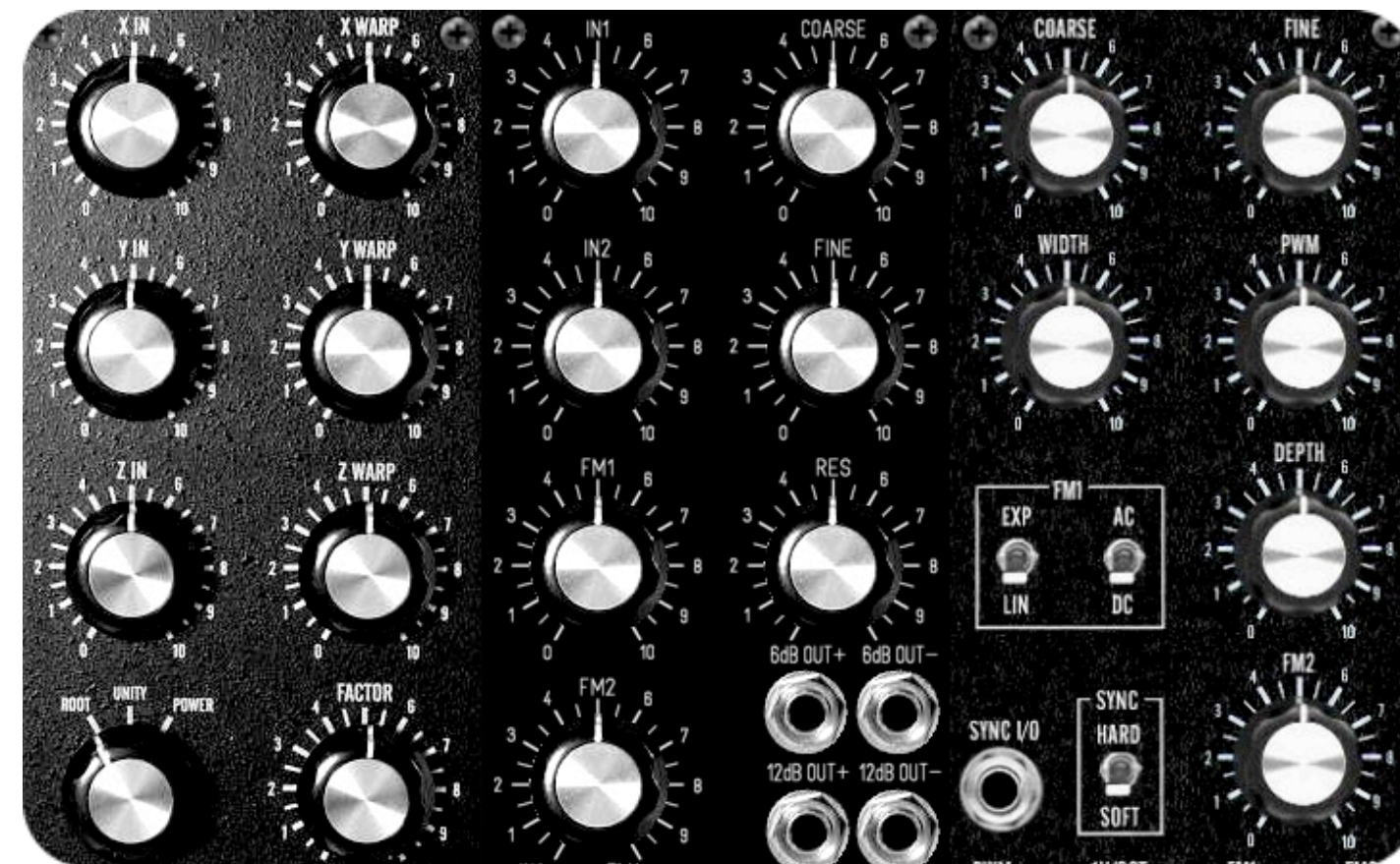
looks rather like:



Training:

Neural network at school

Data

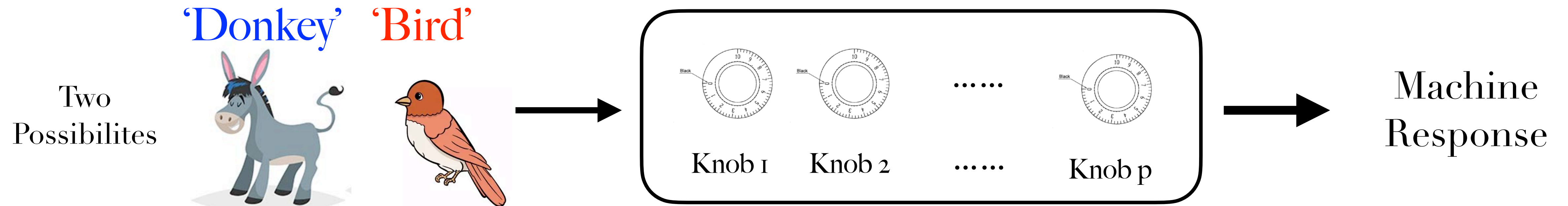


‘Âne’

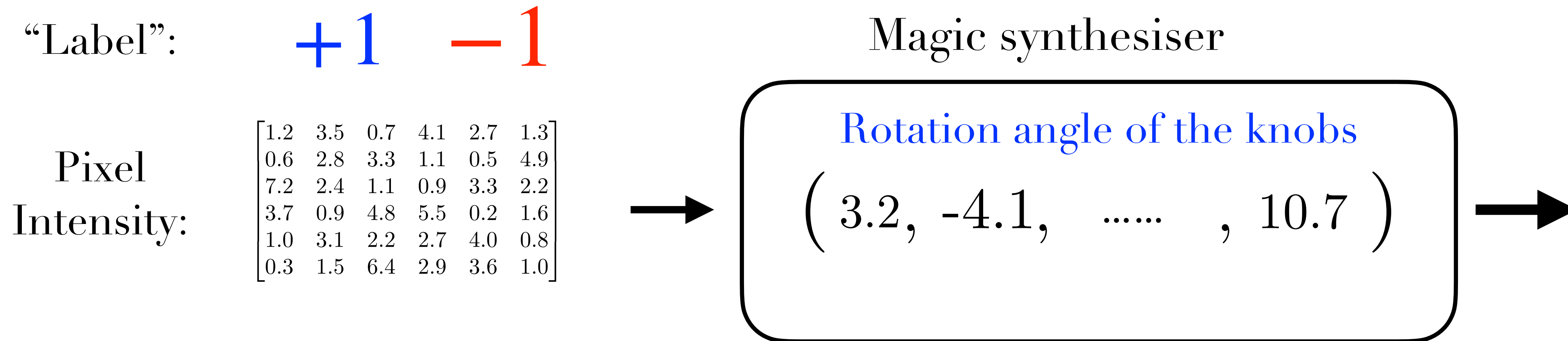
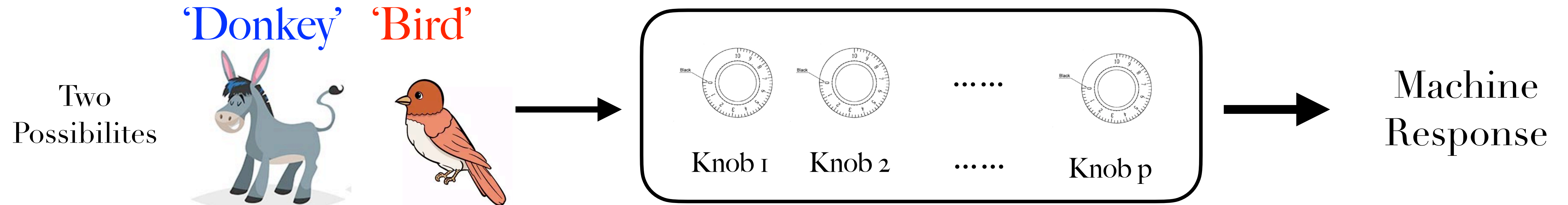


Training a neural network corresponds to turning knobs

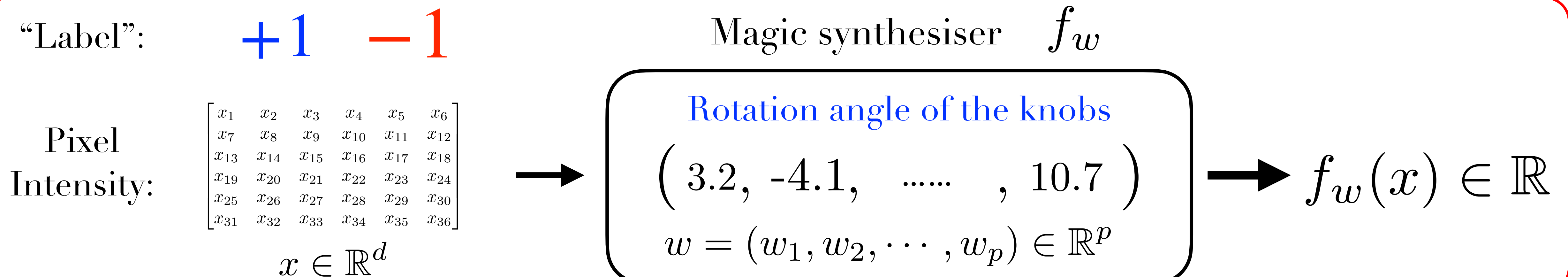
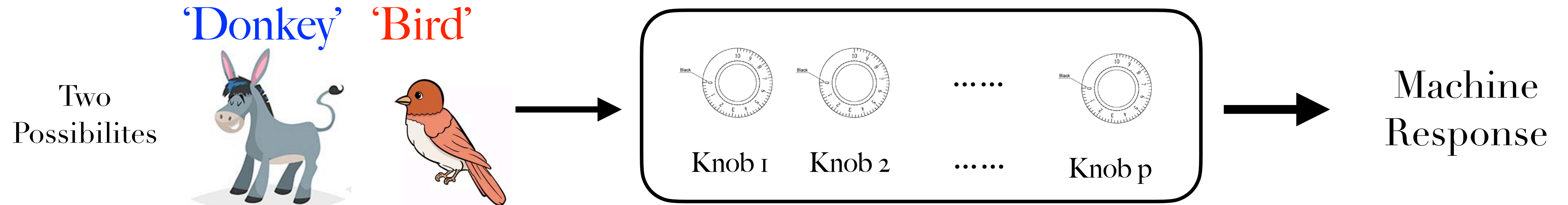
Same story, but with math



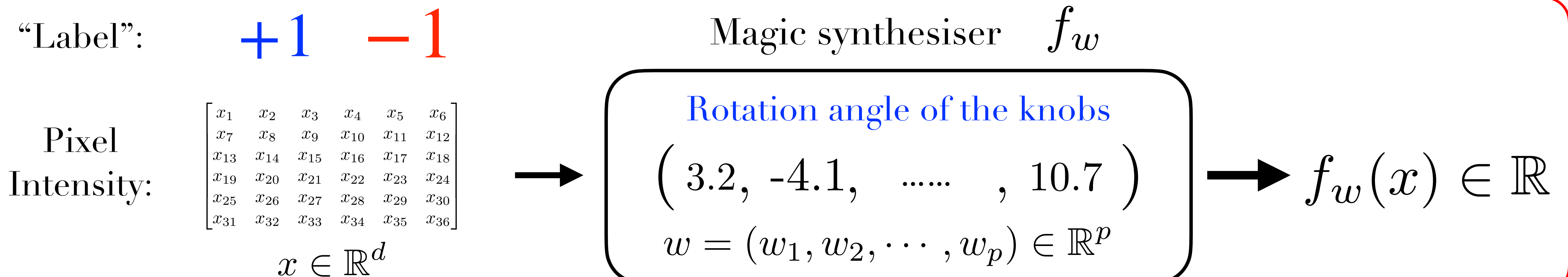
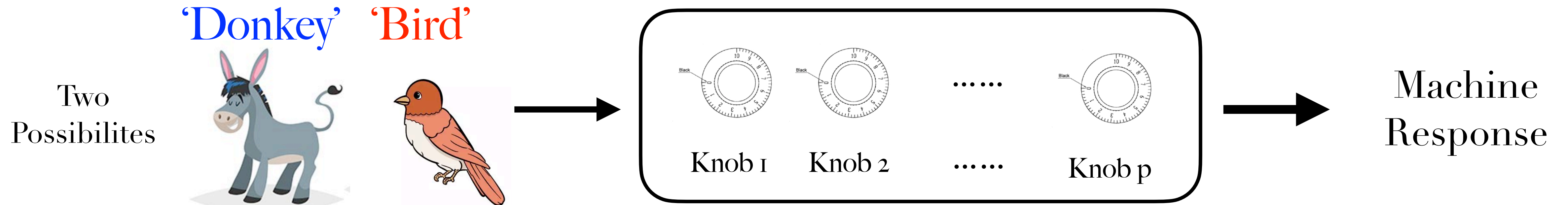
Same story, but with math



Same story, but with math

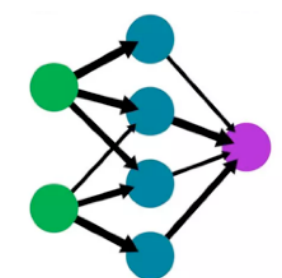


Same story, but with math

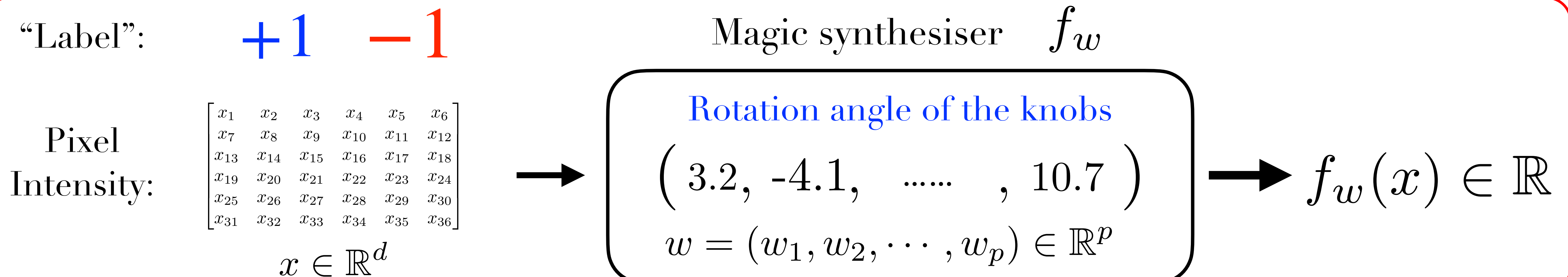
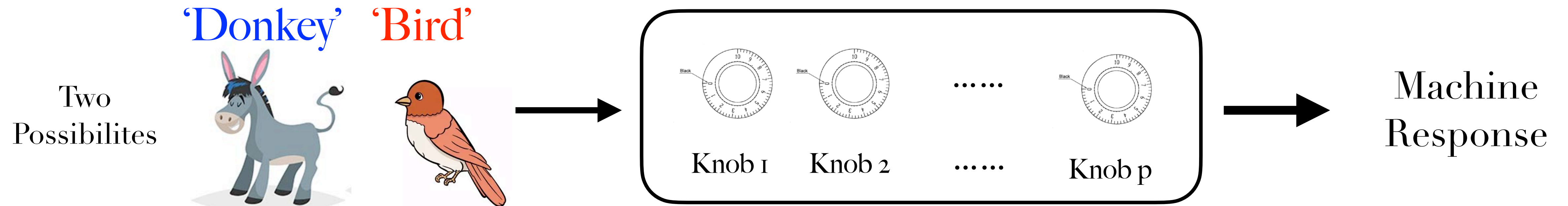


The magical synthesiser: neural networks

$$x \in \mathbb{R}^d \rightarrow f_w(x) = w_7 \max(0, w_1x_1 + w_2x_2 + w_3x_3) + w_8 \max(0, w_4x_4 + w_5x_5 + w_6x_6)$$



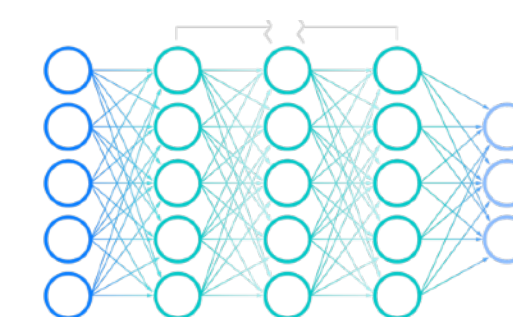
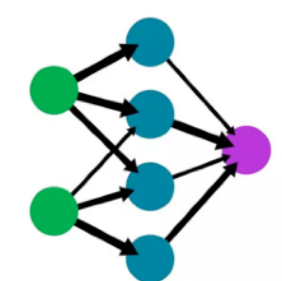
Same story, but with math



The magical synthesiser: neural networks

$$x \in \mathbb{R}^d \rightarrow f_w(x) = w_7 \max(0, w_1 x_1 + w_2 x_2 + w_3 x_3) + w_8 \max(0, w_4 x_4 + w_5 x_5 + w_6 x_6)$$

$$\rightarrow f_w(x) = W_L \phi(W_{L-1} \phi(\dots \phi(W_1 x + b_1) \dots) + b_{L-1}) + b_L.$$



The learning procedure, but with math

‘Donkey’



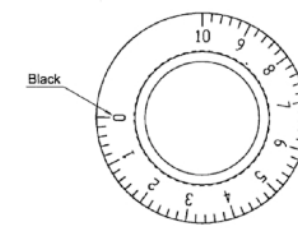
Label +1

$$\begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ x_7 & x_8 & x_9 & x_{10} & x_{11} & x_{12} \\ x_{13} & x_{14} & x_{15} & x_{16} & x_{17} & x_{18} \\ x_{19} & x_{20} & x_{21} & x_{22} & x_{23} & x_{24} \\ x_{25} & x_{26} & x_{27} & x_{28} & x_{29} & x_{30} \\ x_{31} & x_{32} & x_{33} & x_{34} & x_{35} & x_{36} \end{bmatrix}$$

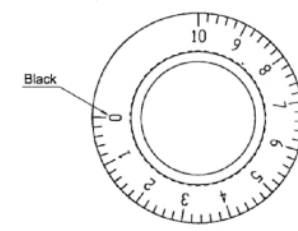
$$x \in \mathbb{R}^d$$



Magic synthesiser f_w



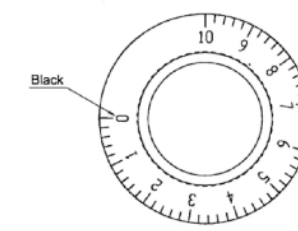
Knob 1



Knob 2

.....

.....



Knob p

$$w = (w_1, w_2, \dots, w_p) \in \mathbb{R}^p$$

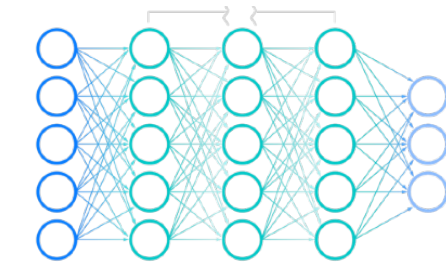


Machine
Response

$$f_w(x) \in \mathbb{R}$$

Magic synthesisers: neural networks

$$\mathcal{X} \rightarrow f_w(x) = w_7 \max(0, w_1 x_1 + w_2 x_2 + w_3 x_3) + w_8 \max(0, w_4 x_4 + w_5 x_5 + w_6 x_6)$$



The learning procedure, but with math

‘Donkey’



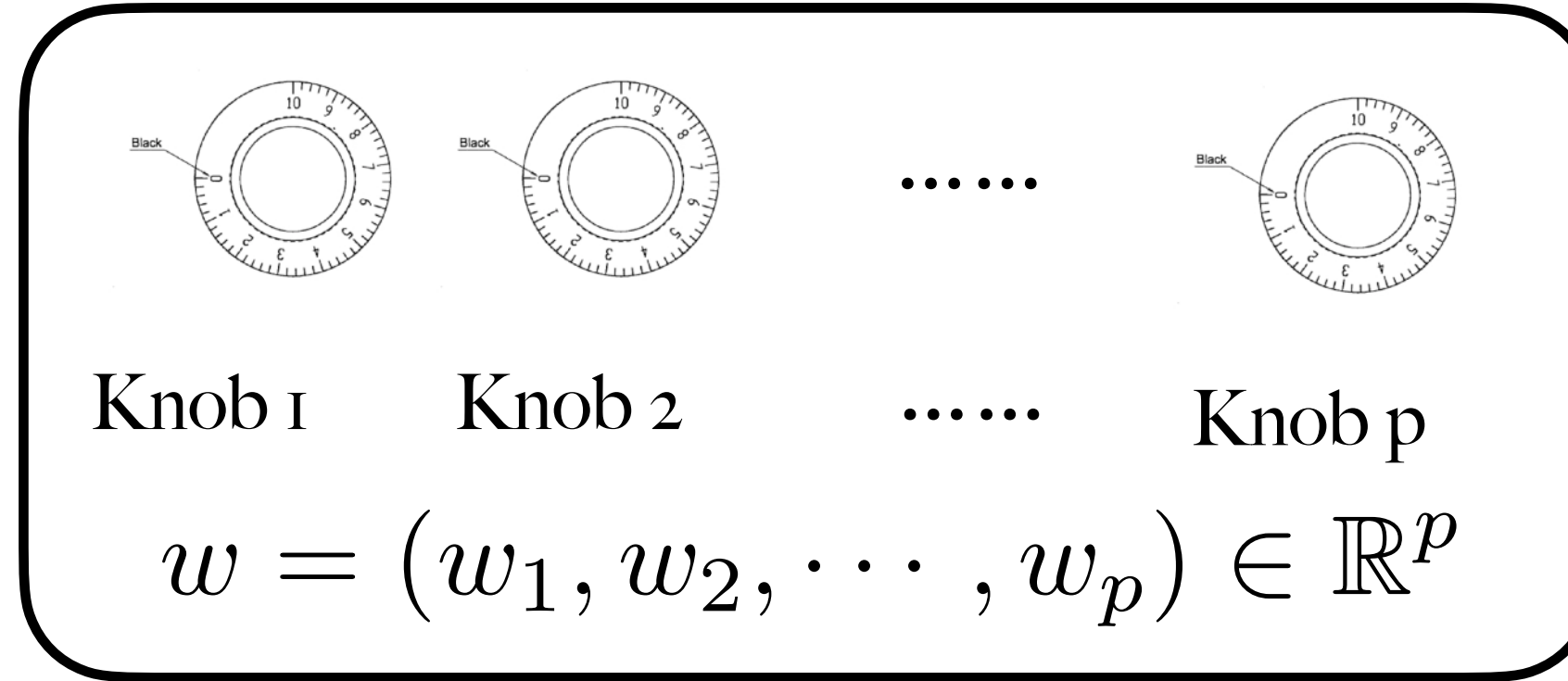
Label +1

$$\begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ x_7 & x_8 & x_9 & x_{10} & x_{11} & x_{12} \\ x_{13} & x_{14} & x_{15} & x_{16} & x_{17} & x_{18} \\ x_{19} & x_{20} & x_{21} & x_{22} & x_{23} & x_{24} \\ x_{25} & x_{26} & x_{27} & x_{28} & x_{29} & x_{30} \\ x_{31} & x_{32} & x_{33} & x_{34} & x_{35} & x_{36} \end{bmatrix}$$

$$x \in \mathbb{R}^d$$



Magic synthesiser f_w

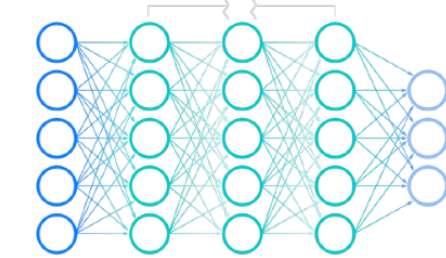


Machine
Response

$$f_w(x) \in \mathbb{R}$$

Magic synthesisers: neural networks

$$x \rightarrow f_w(x) = w_7 \max(0, w_1 x_1 + w_2 x_2 + w_3 x_3) + w_8 \max(0, w_4 x_4 + w_5 x_5 + w_6 x_6)$$



How can we get it to learn?

The learning procedure, but with math

‘Donkey’



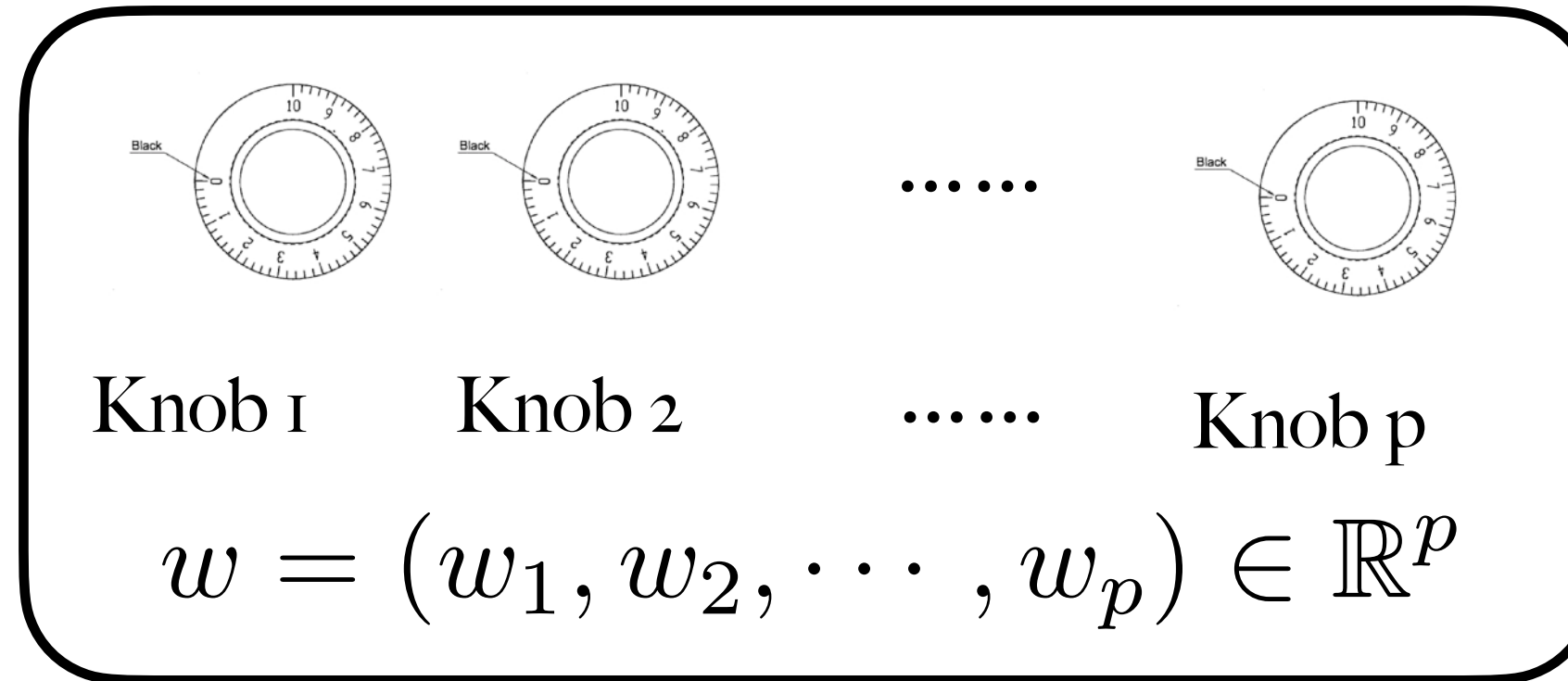
Label +1

$$\begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ x_7 & x_8 & x_9 & x_{10} & x_{11} & x_{12} \\ x_{13} & x_{14} & x_{15} & x_{16} & x_{17} & x_{18} \\ x_{19} & x_{20} & x_{21} & x_{22} & x_{23} & x_{24} \\ x_{25} & x_{26} & x_{27} & x_{28} & x_{29} & x_{30} \\ x_{31} & x_{32} & x_{33} & x_{34} & x_{35} & x_{36} \end{bmatrix}$$

$$x \in \mathbb{R}^d$$



Magic synthesiser f_w

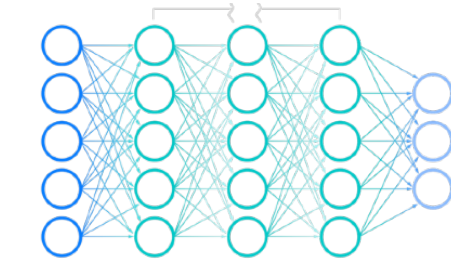


Machine Response

$$f_w(x) \in \mathbb{R}$$

Magic synthesisers: neural networks

$$\mathcal{X} \rightarrow f_w(x) = w_7 \max(0, w_1 x_1 + w_2 x_2 + w_3 x_3) + w_8 \max(0, w_4 x_4 + w_5 x_5 + w_6 x_6)$$



How can we get it to learn?

Before training, it predicts $f_w(x)$ (-234.5), and we want it to predict “Donkey” ($+1$)

We can model it's mistake by

The learning procedure, but with math

‘Donkey’

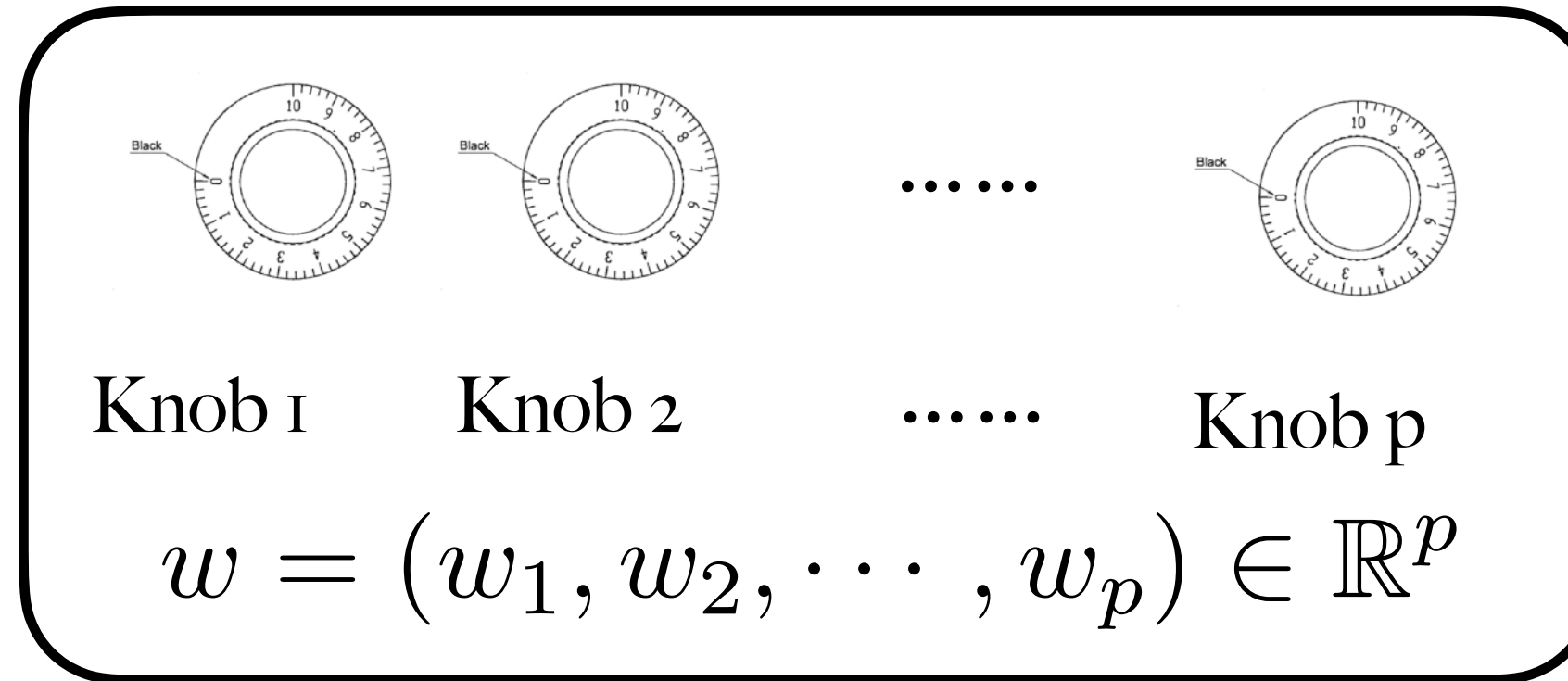


Label +1

$$\begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ x_7 & x_8 & x_9 & x_{10} & x_{11} & x_{12} \\ x_{13} & x_{14} & x_{15} & x_{16} & x_{17} & x_{18} \\ x_{19} & x_{20} & x_{21} & x_{22} & x_{23} & x_{24} \\ x_{25} & x_{26} & x_{27} & x_{28} & x_{29} & x_{30} \\ x_{31} & x_{32} & x_{33} & x_{34} & x_{35} & x_{36} \end{bmatrix}$$

$$x \in \mathbb{R}^d$$

Magic synthesiser f_w

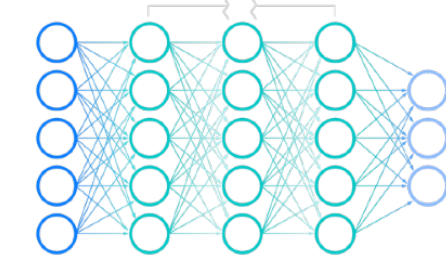


Machine
Response

$$f_w(x) \in \mathbb{R}$$

Magic synthesisers: neural networks

$$\mathcal{X} \rightarrow f_w(x) = w_7 \max(0, w_1 x_1 + w_2 x_2 + w_3 x_3) + w_8 \max(0, w_4 x_4 + w_5 x_5 + w_6 x_6)$$



How can we get it to learn?

Before training, it predicts $f_w(x)$ (-234.5), and we want it to predict “Donkey” (+1)

We can model it's mistake by $L(w) = (f_w(x) - 1)^2$

The learning procedure, but with math

‘Donkey’

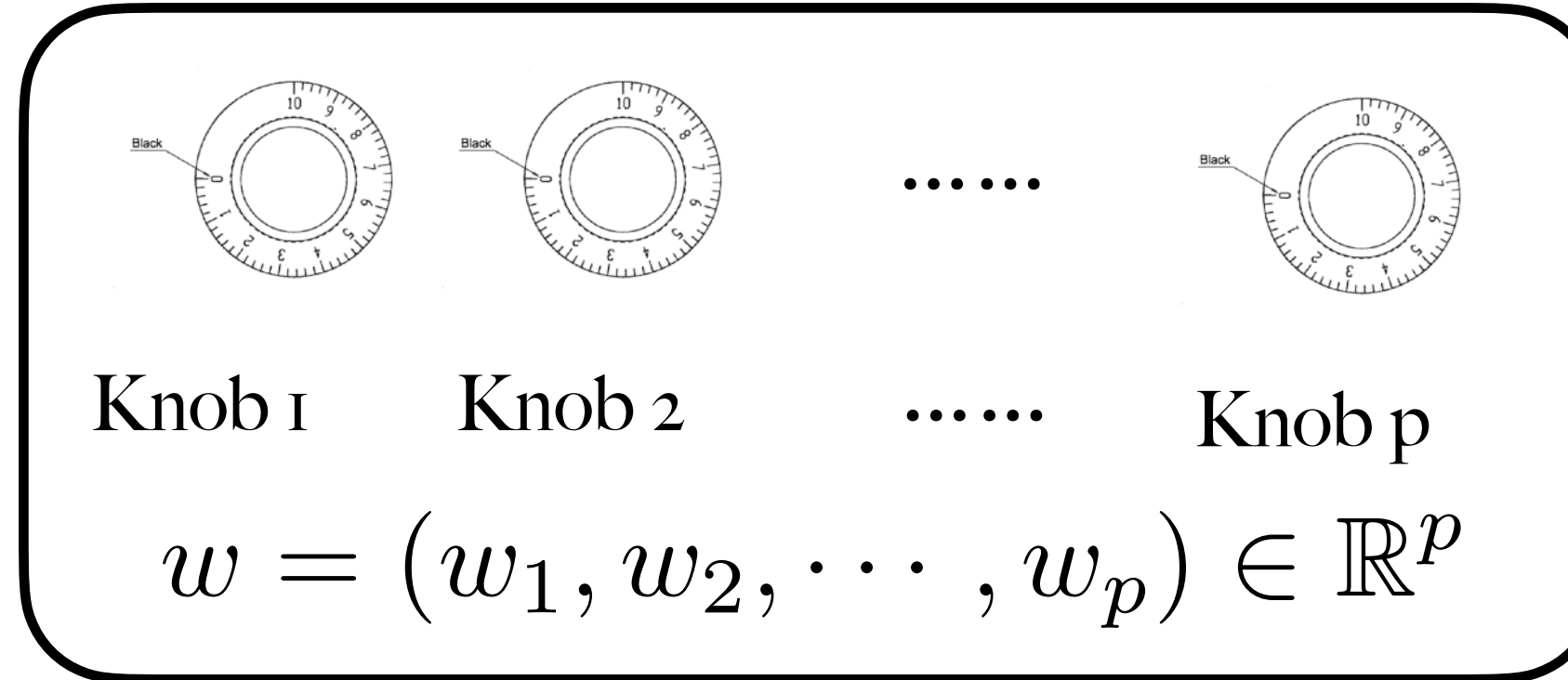


Label +1

$$\begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ x_7 & x_8 & x_9 & x_{10} & x_{11} & x_{12} \\ x_{13} & x_{14} & x_{15} & x_{16} & x_{17} & x_{18} \\ x_{19} & x_{20} & x_{21} & x_{22} & x_{23} & x_{24} \\ x_{25} & x_{26} & x_{27} & x_{28} & x_{29} & x_{30} \\ x_{31} & x_{32} & x_{33} & x_{34} & x_{35} & x_{36} \end{bmatrix}$$

$$x \in \mathbb{R}^d$$

Magic synthesiser f_w

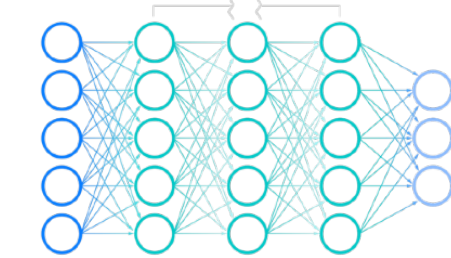


Machine Response

$$f_w(x) \in \mathbb{R}$$

Magic synthesisers: neural networks

$$\mathcal{X} \rightarrow f_w(x) = w_7 \max(0, w_1x_1 + w_2x_2 + w_3x_3) + w_8 \max(0, w_4x_4 + w_5x_5 + w_6x_6)$$



How can we get it to learn?

Before training, it predicts $f_w(x)$ (-234.5), and we want it to predict “Donkey” ($+1$)

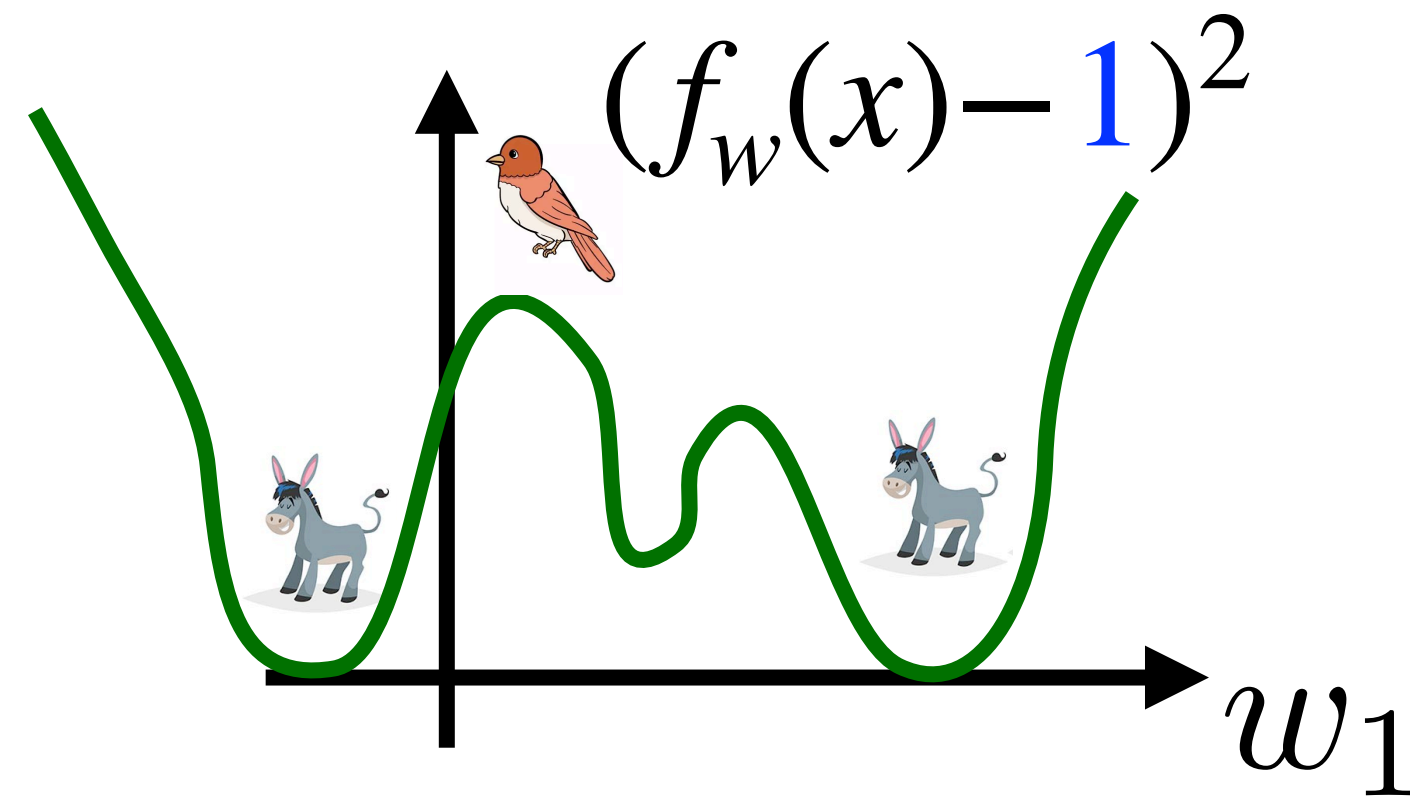
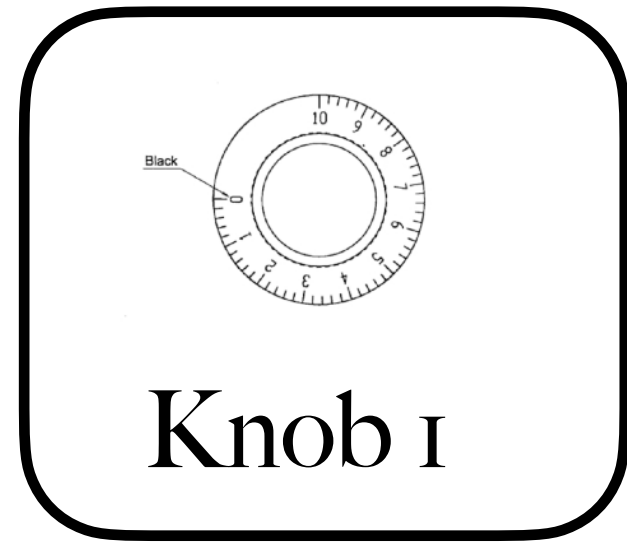
We can model it's mistake by $L(w) = (f_w(x) - 1)^2$

We want to minimise the mistake! $\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$

It boils down to minimising a function!

$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

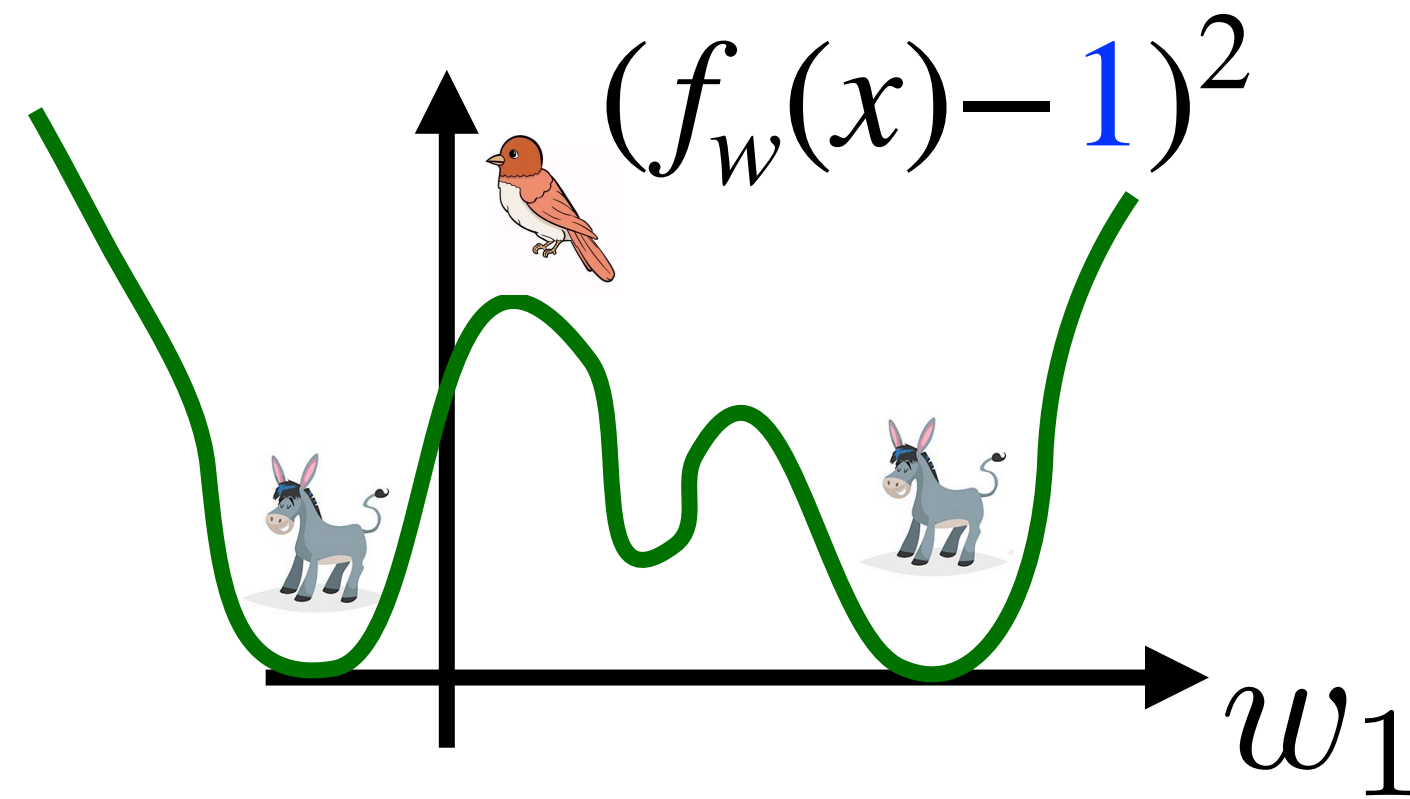
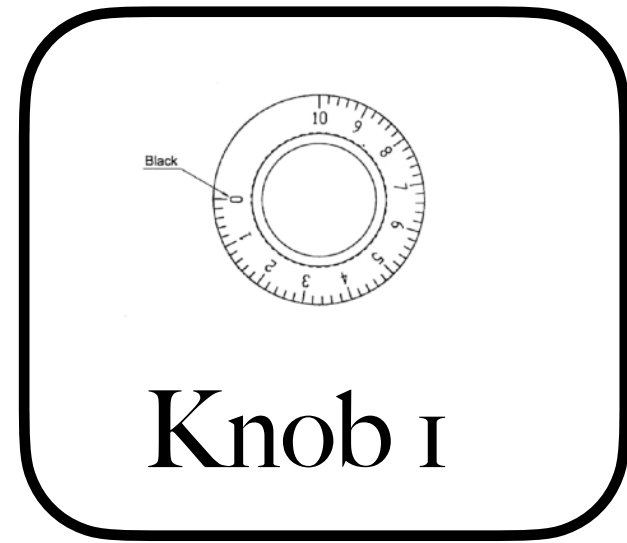
$$w = (w_1)$$



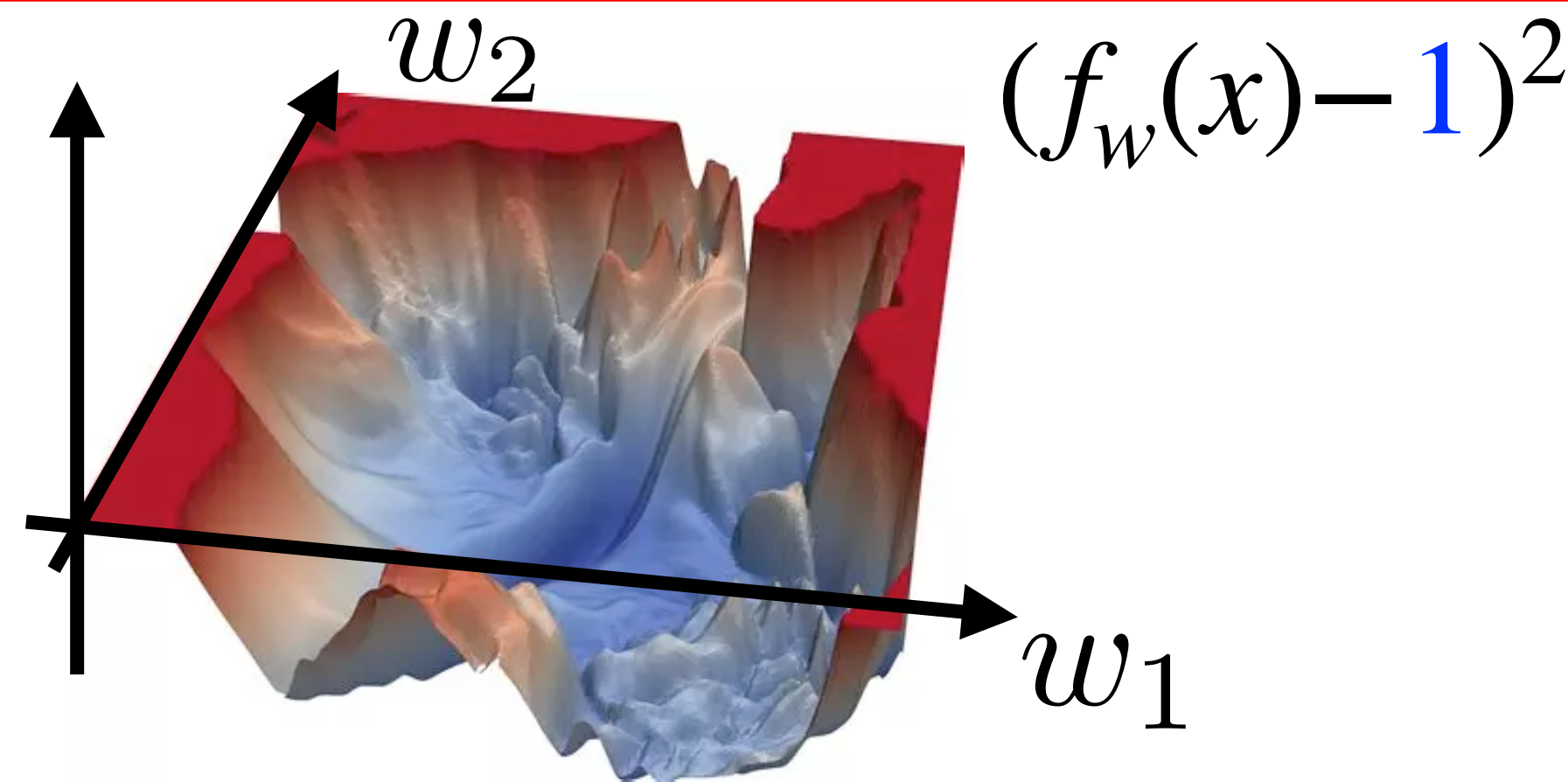
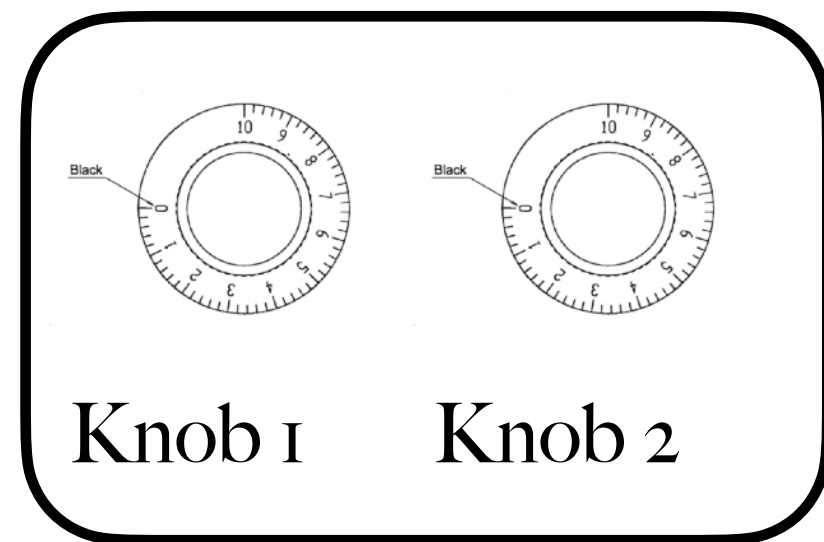
It boils down to minimising a function!

$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

$$w = (w_1)$$



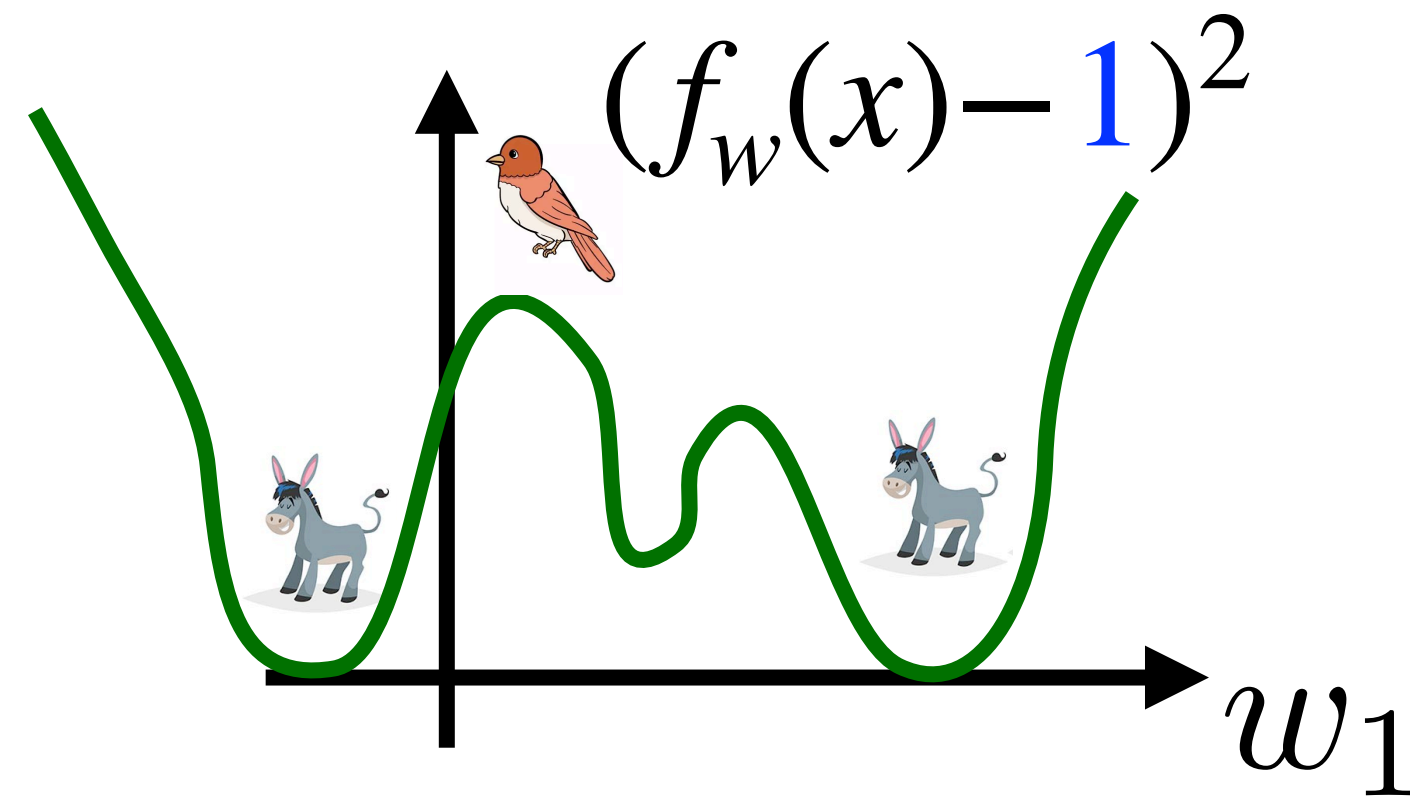
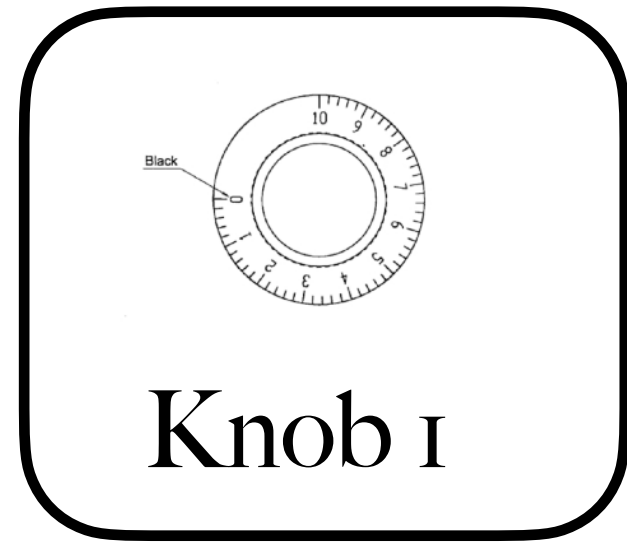
$$w = (w_1, w_2)$$



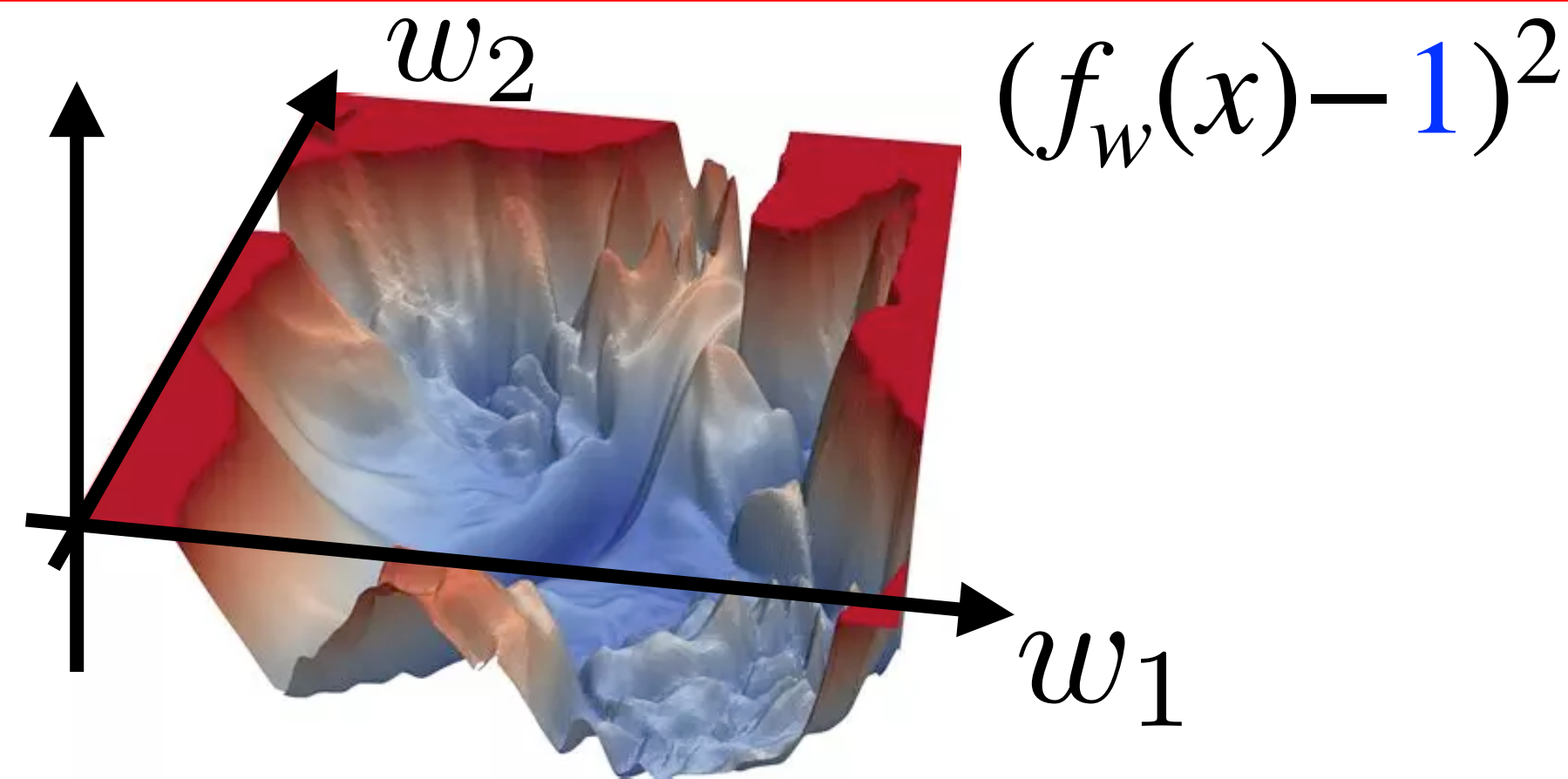
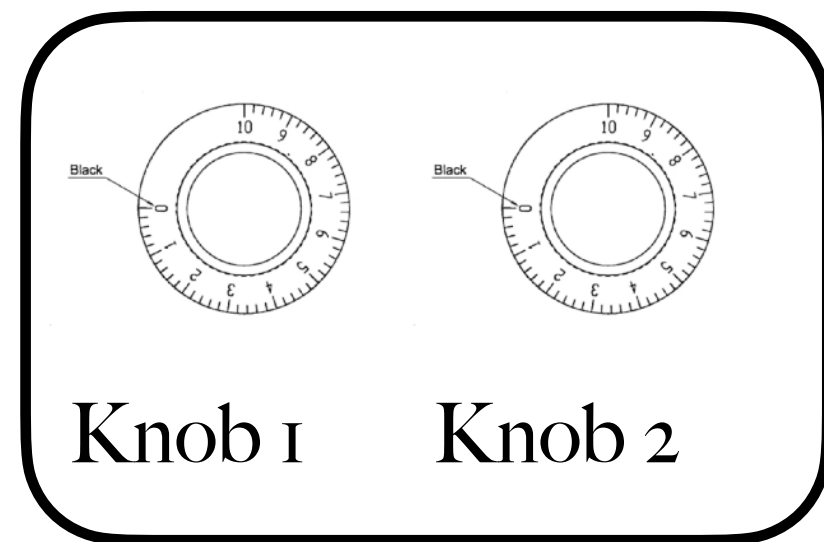
It boils down to minimising a function!

$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

$$w = (w_1)$$



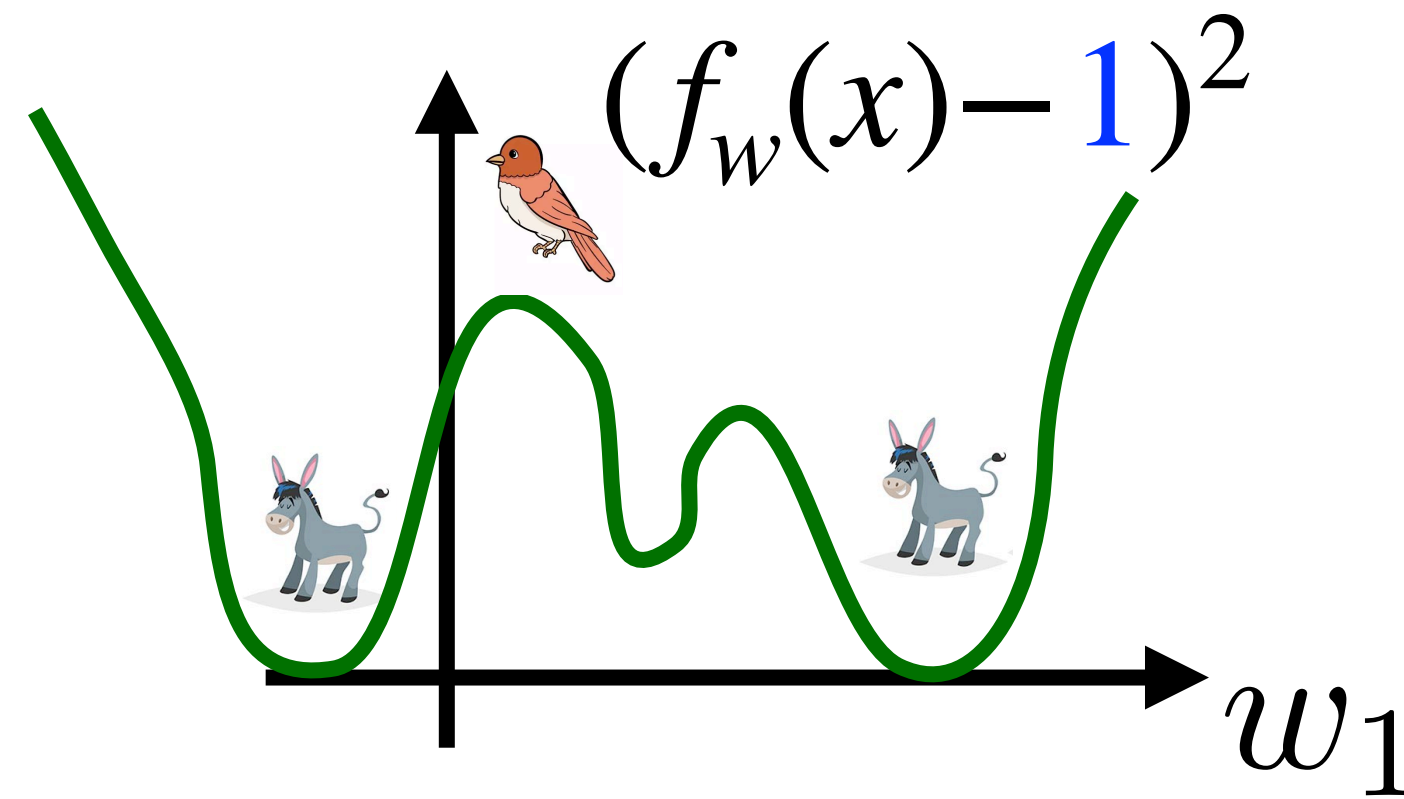
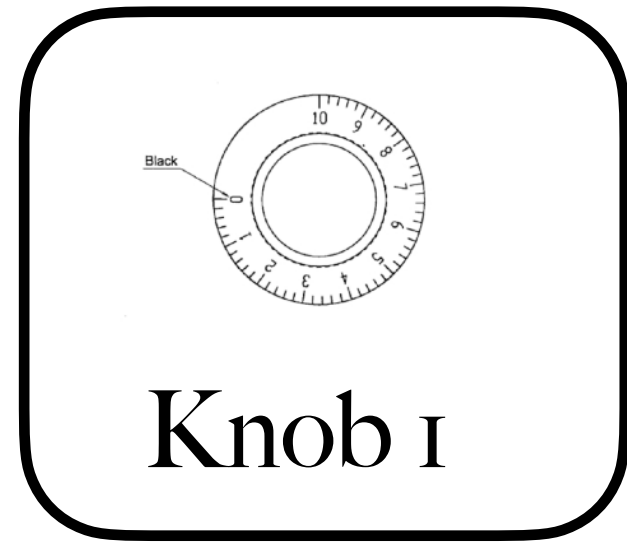
$$w = (w_1, w_2)$$



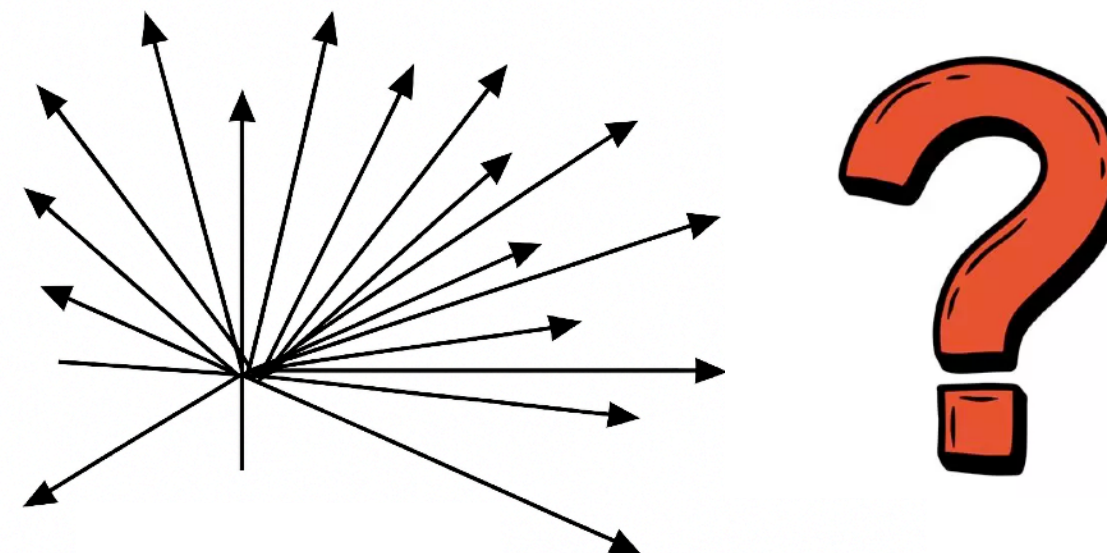
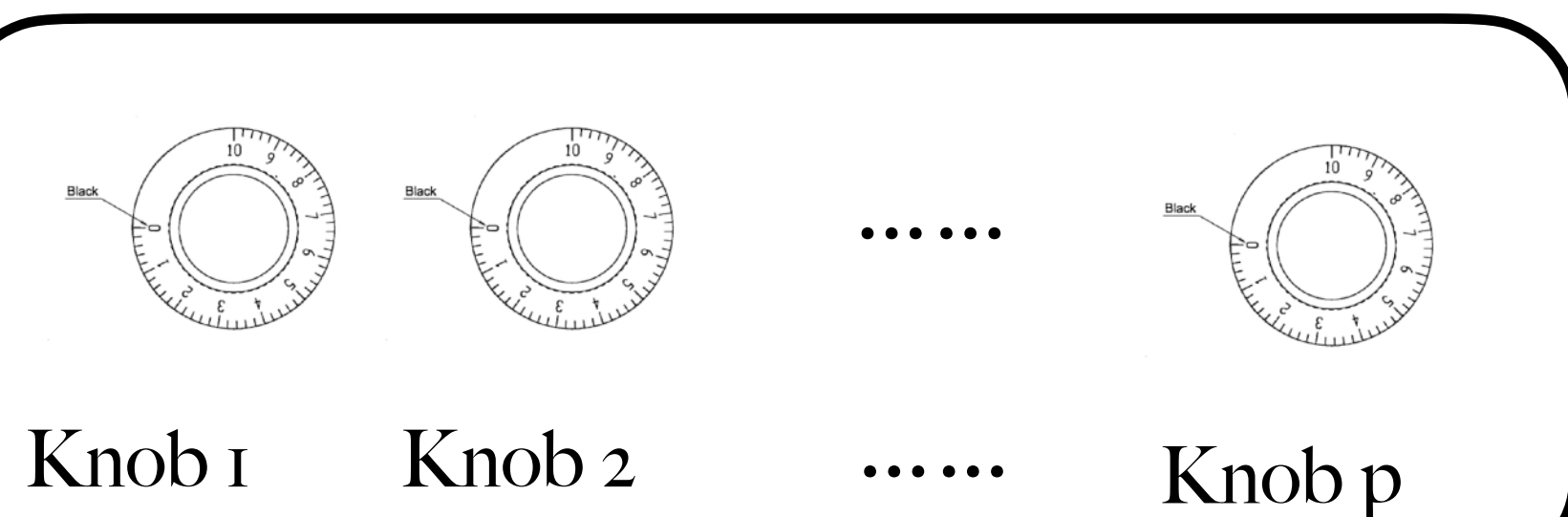
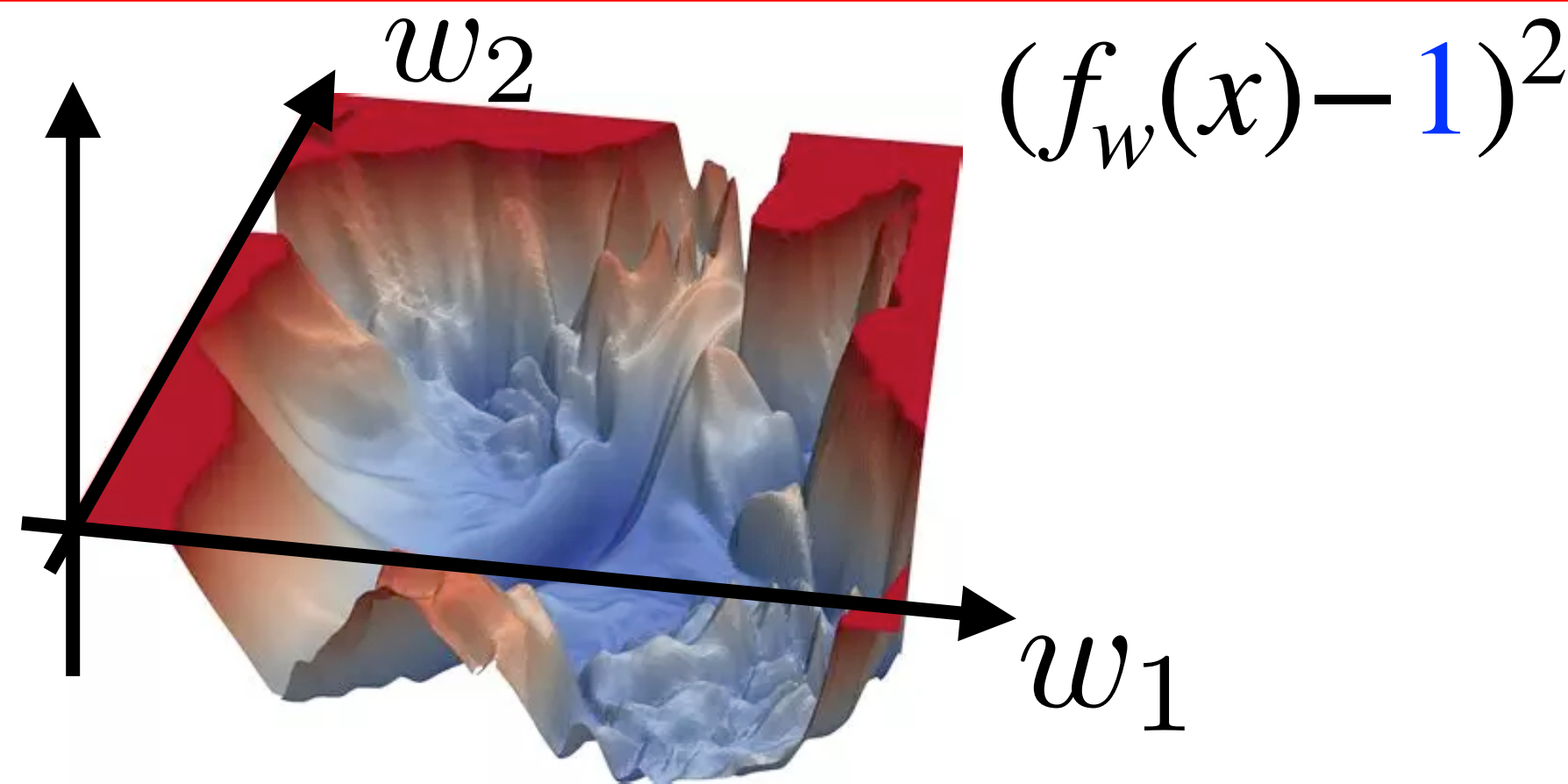
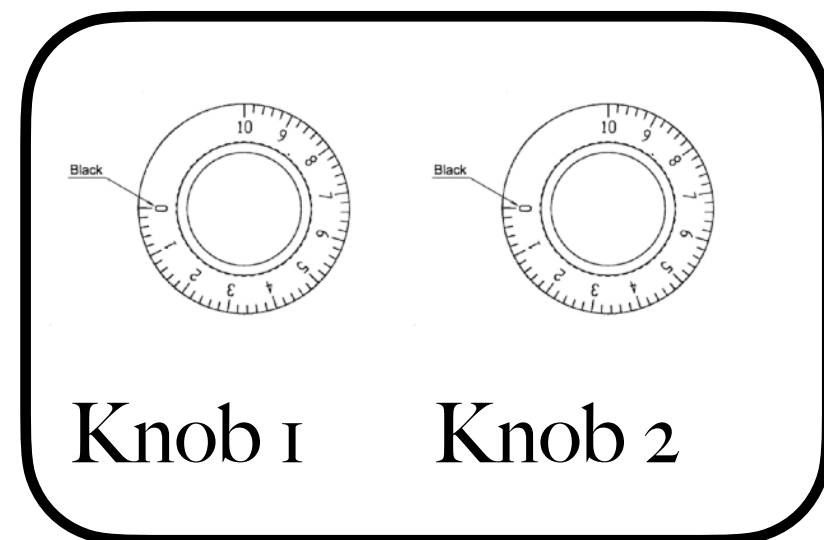
It boils down to minimising a function!

$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

$$w = (w_1)$$



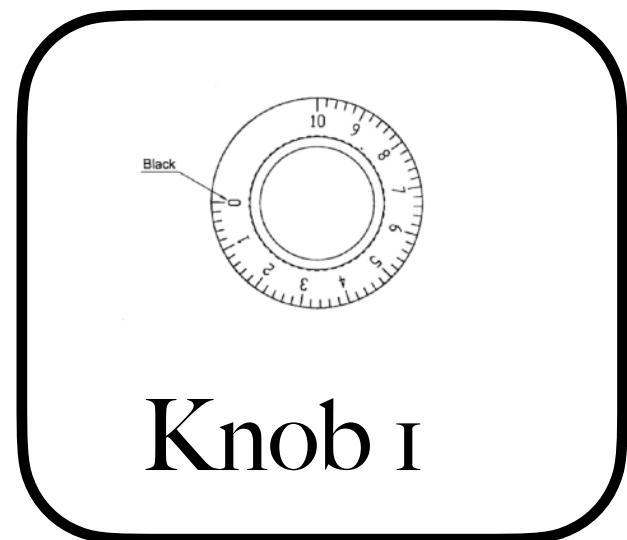
$$w = (w_1, w_2)$$



It boils down to minimising a function!

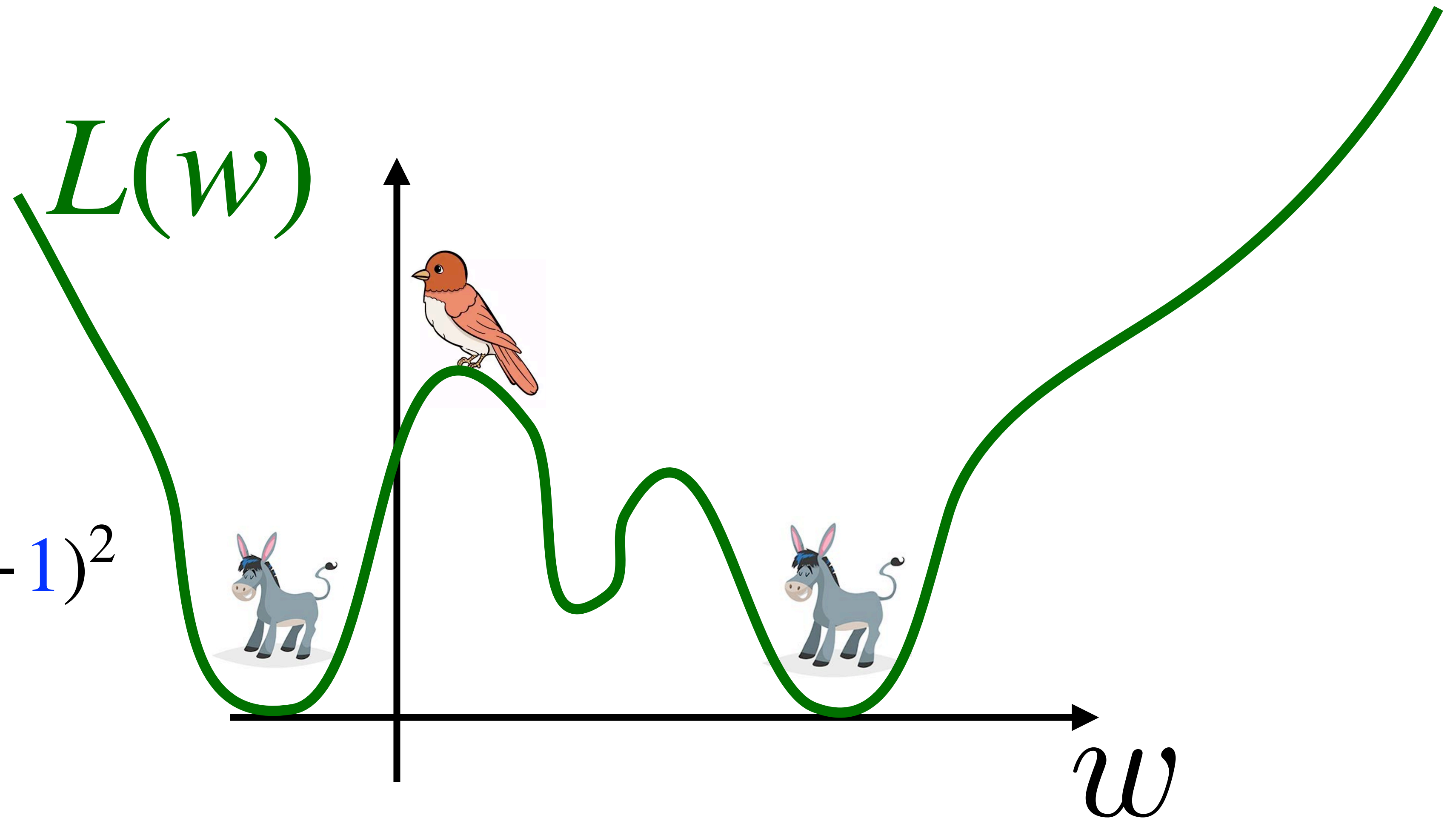
$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

Back to 1D:



$$\min_{w \in \mathbb{R}^p} L(w)$$

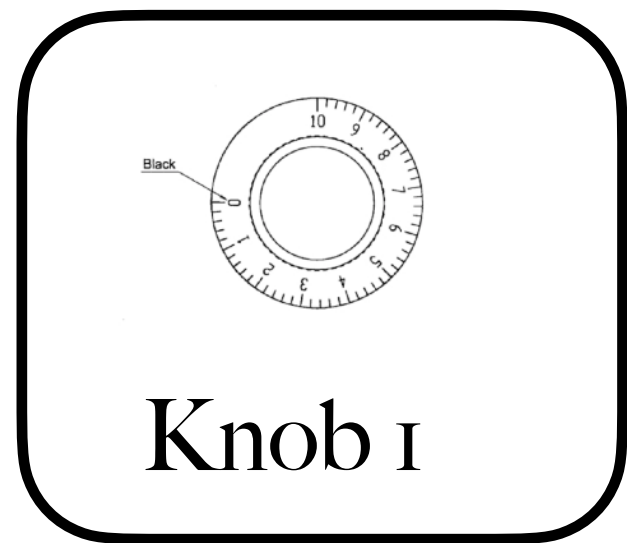
$$L(w) = (f_w(x) - 1)^2$$



It boils down to minimising a function!

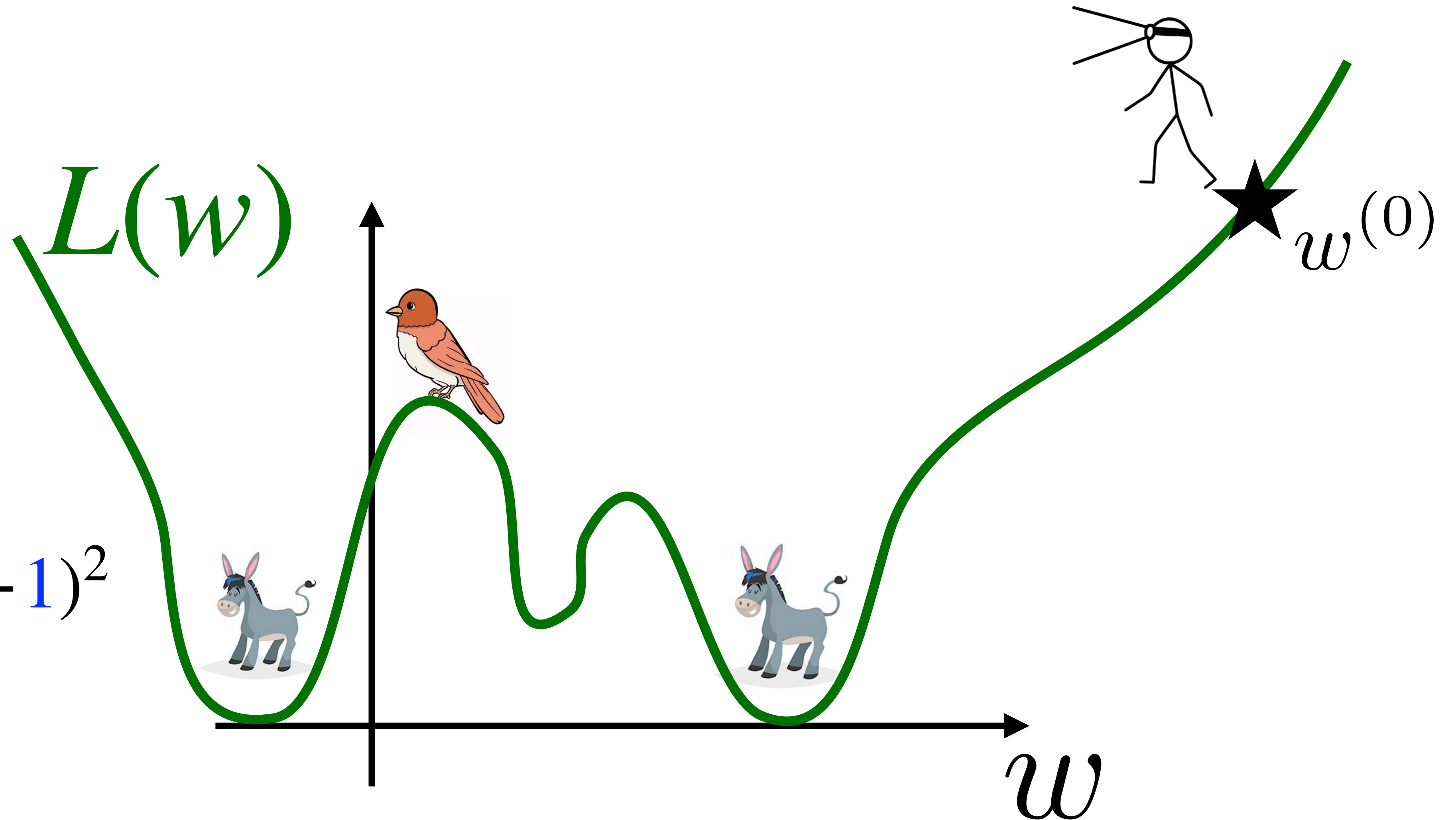
$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

Back to 1D:



$$\min_{w \in \mathbb{R}^p} L(w)$$

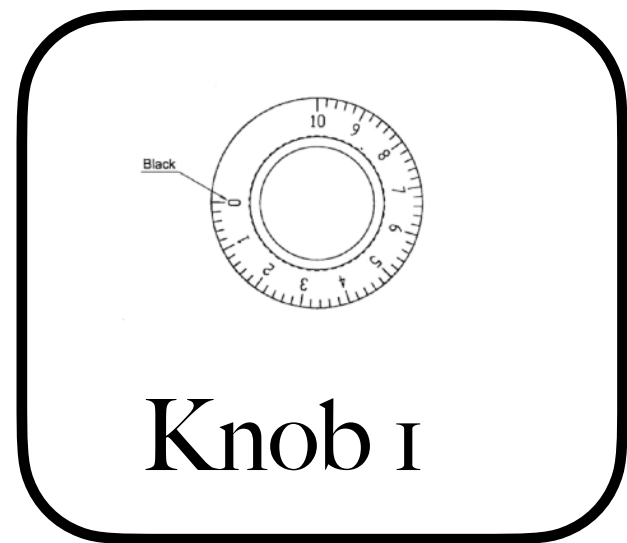
$$L(w) = (f_w(x) - 1)^2$$



It boils down to minimising a function!

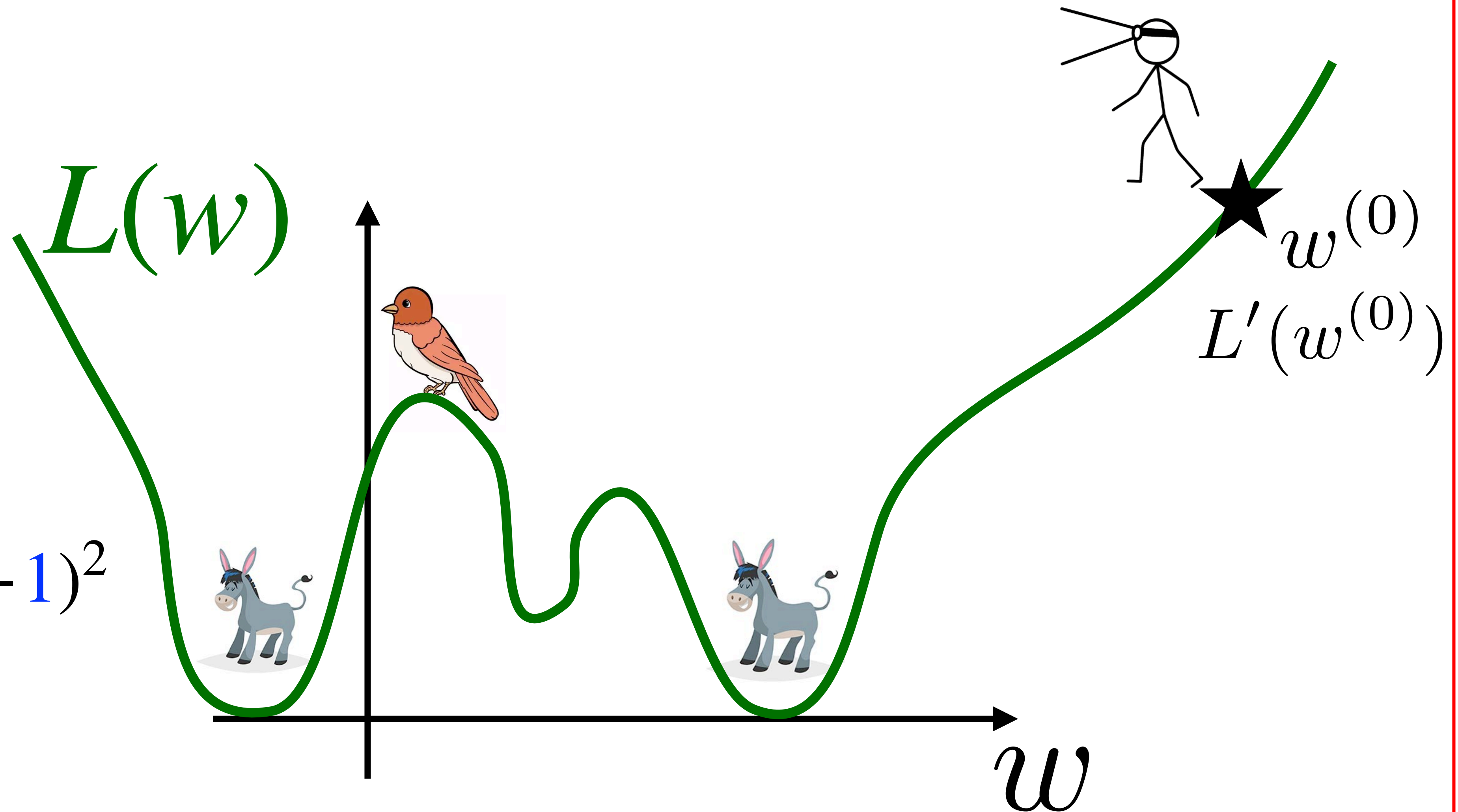
$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

Back to 1D:



$$\min_{w \in \mathbb{R}^p} L(w)$$

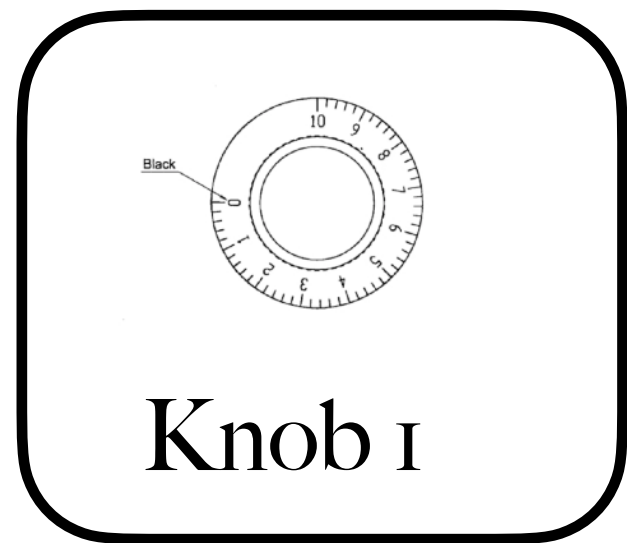
$$L(w) = (f_w(x) - 1)^2$$



It boils down to minimising a function!

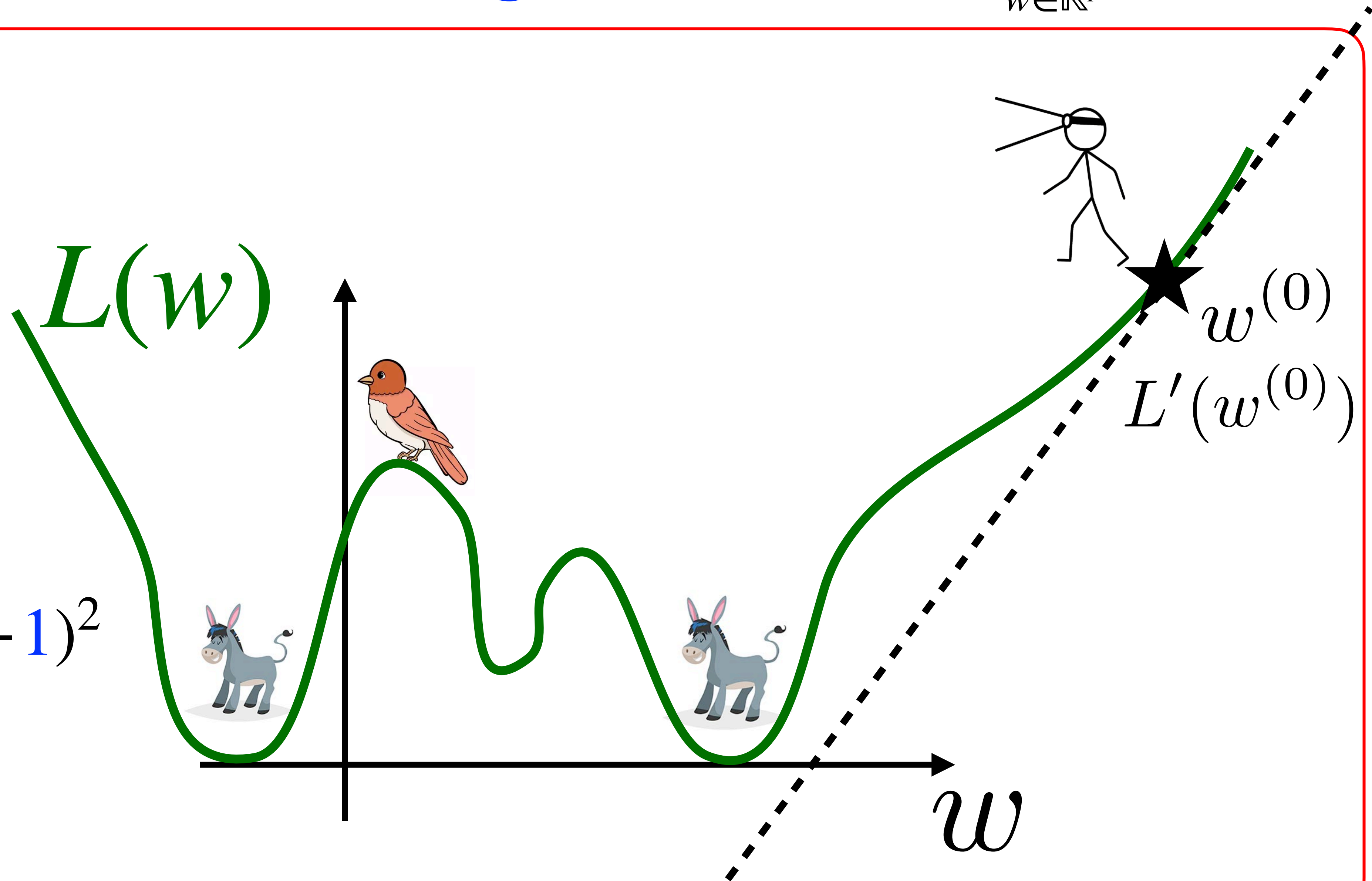
$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

Back to 1D:



$$\min_{w \in \mathbb{R}^p} L(w)$$

$$L(w) = (f_w(x) - 1)^2$$

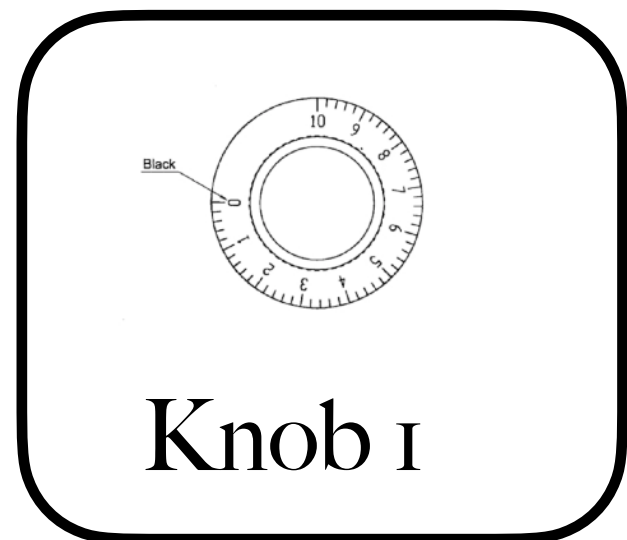


It boils down to minimising a function!

$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

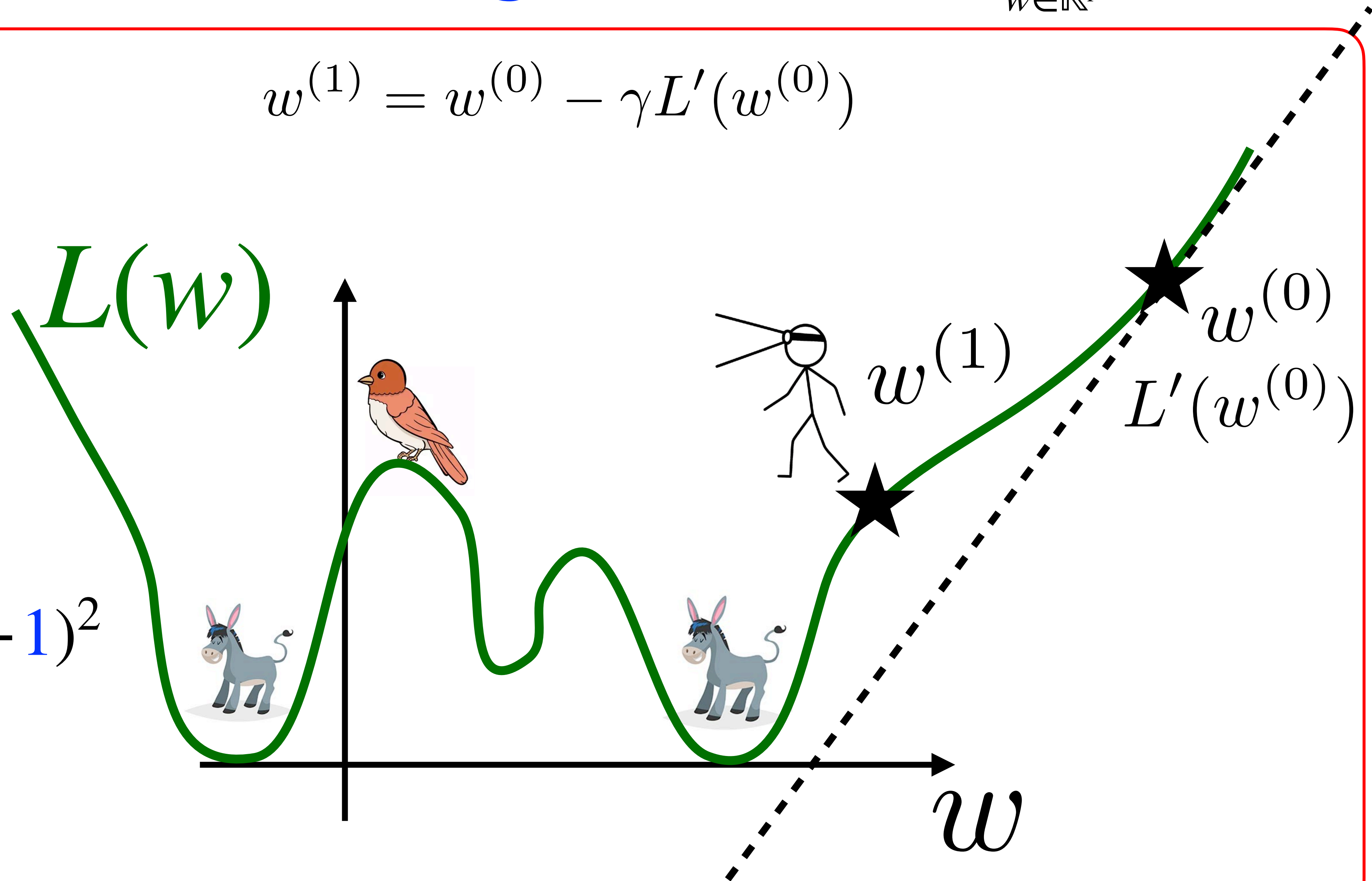
Back to 1D:

$$w^{(1)} = w^{(0)} - \gamma L'(w^{(0)})$$



$$\min_{w \in \mathbb{R}^p} L(w)$$

$$L(w) = (f_w(x) - 1)^2$$

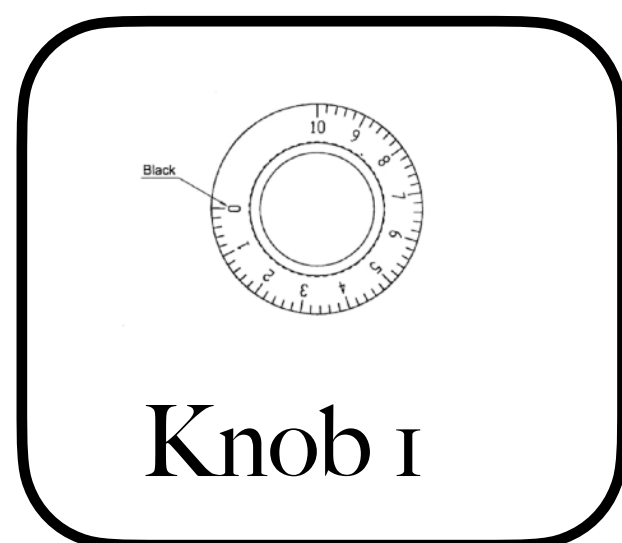


It boils down to minimising a function!

$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

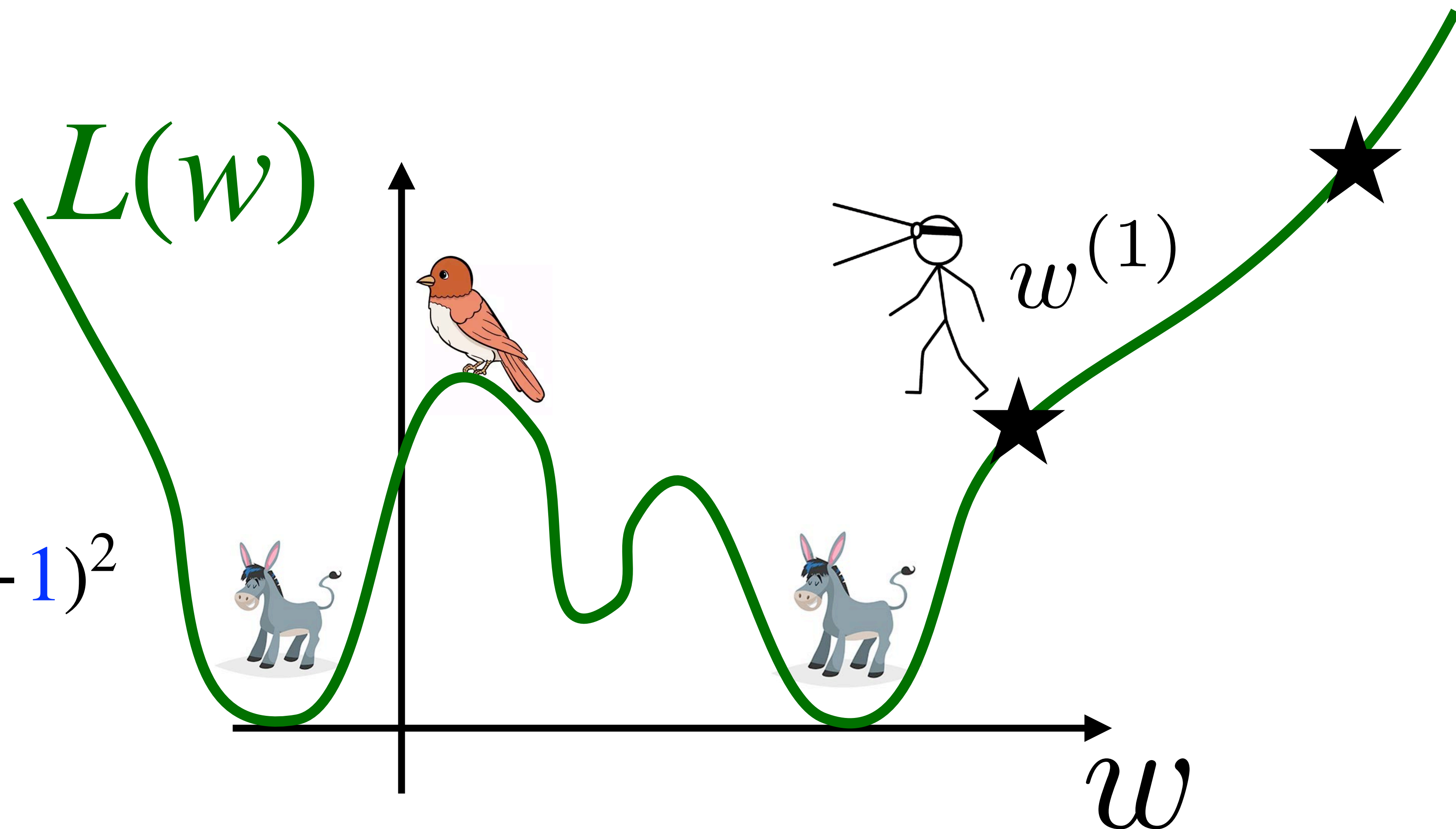
Back to 1D:

$$w^{(1)} = w^{(0)} - \gamma L'(w^{(0)})$$



$$\min_{w \in \mathbb{R}^p} L(w)$$

$$L(w) = (f_w(x) - 1)^2$$

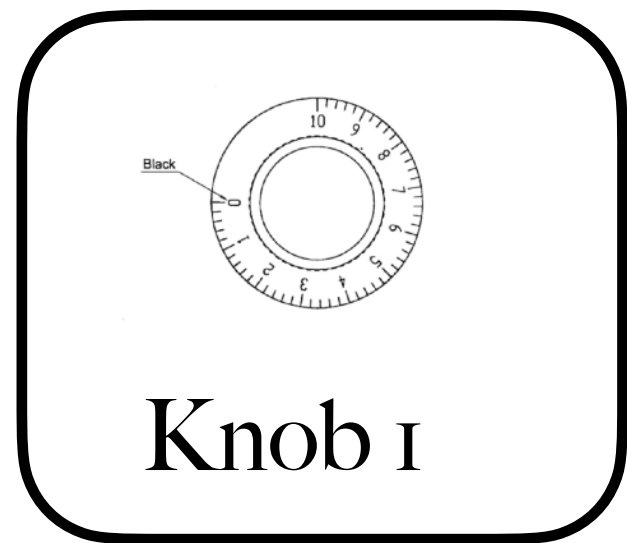


It boils down to minimising a function!

$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

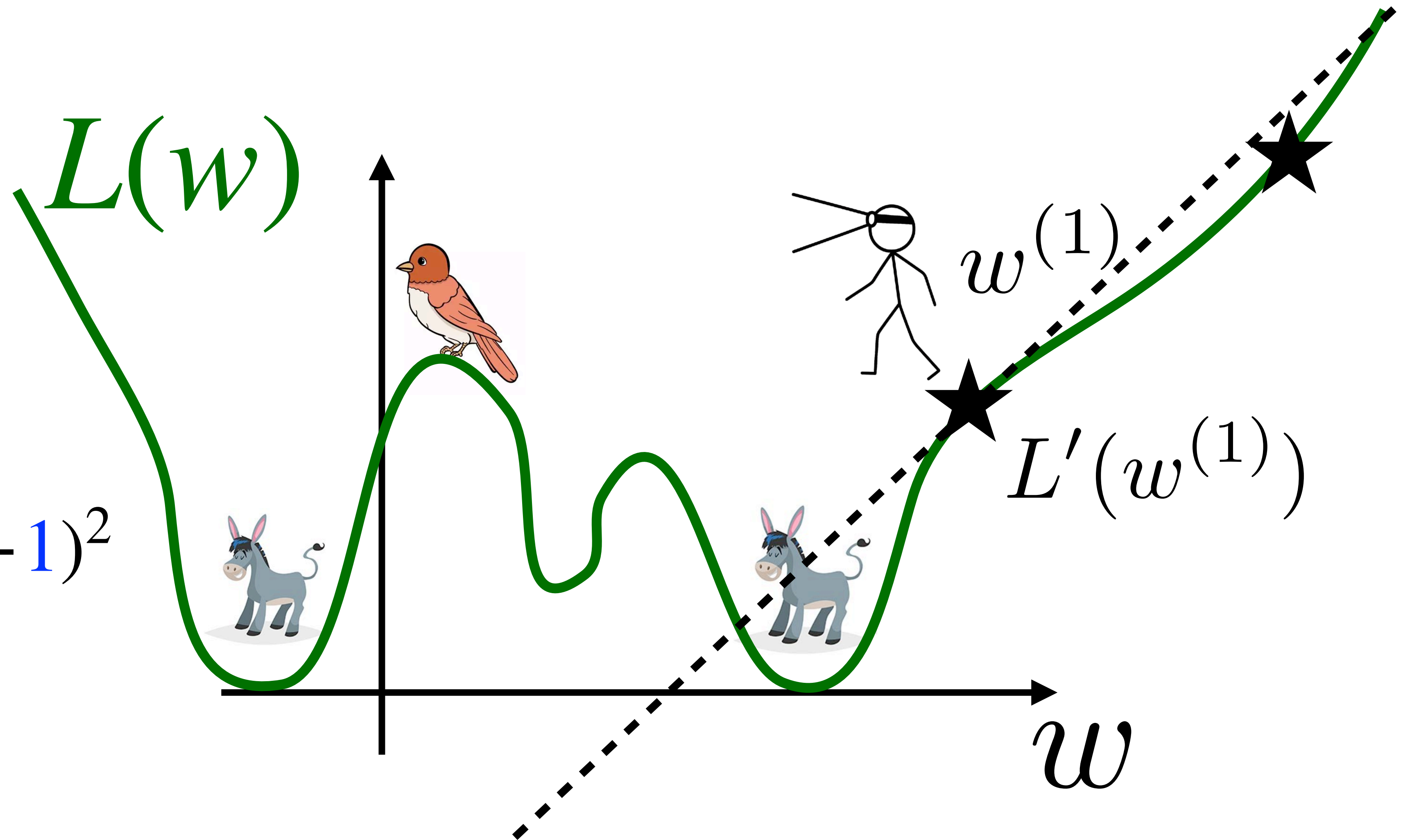
Back to 1D:

$$w^{(1)} = w^{(0)} - \gamma L'(w^{(0)})$$



$$\min_{w \in \mathbb{R}^p} L(w)$$

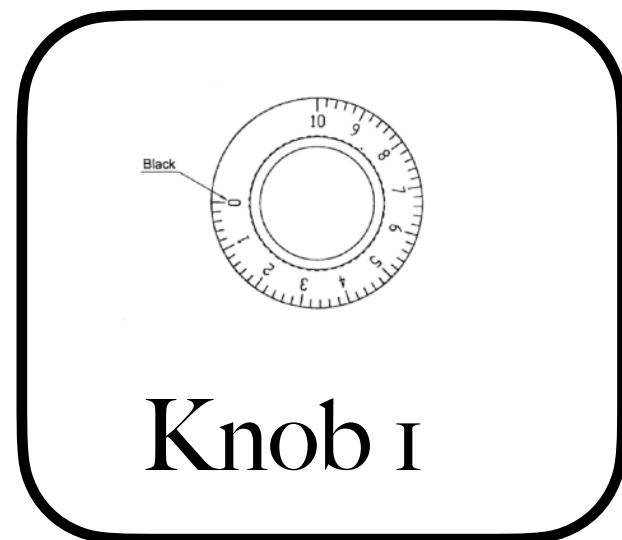
$$L(w) = (f_w(x) - 1)^2$$



It boils down to minimising a function!

$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

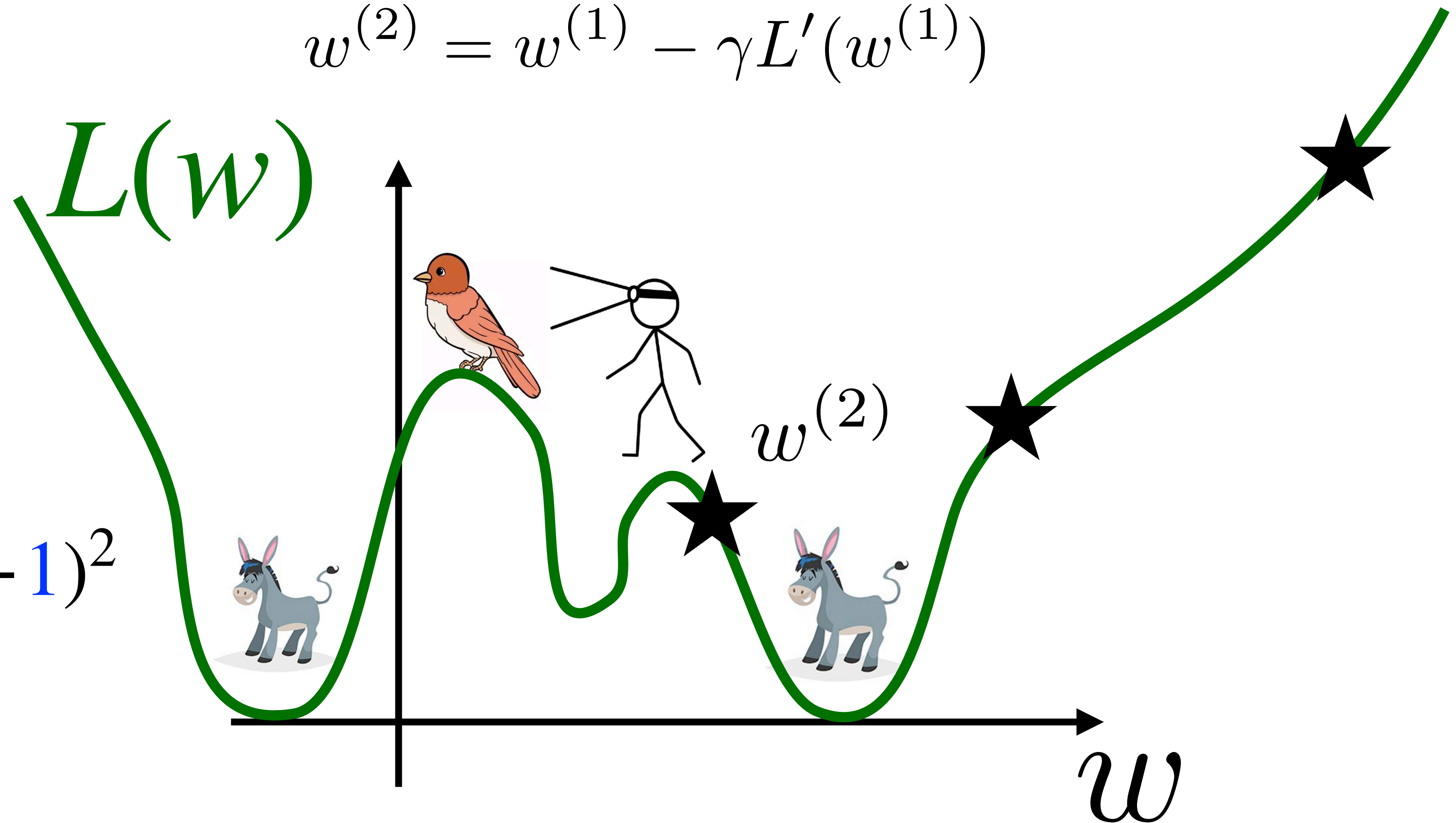
Back to 1D:



$$\min_{w \in \mathbb{R}^p} L(w)$$

$$L(w) = (f_w(x) - 1)^2$$

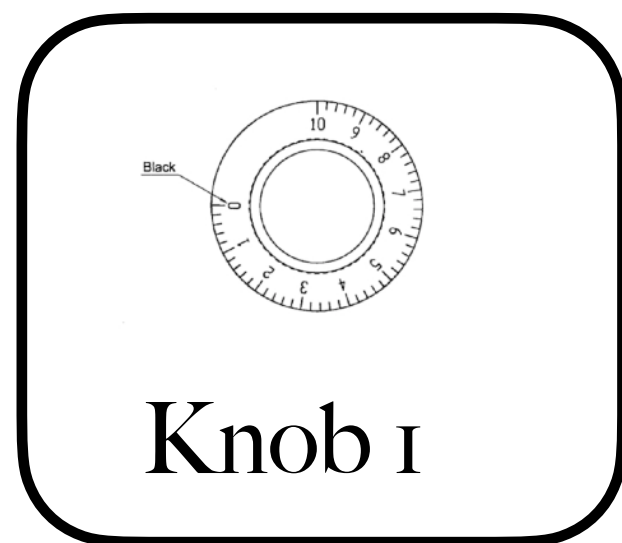
$$w^{(1)} = w^{(0)} - \gamma L'(w^{(0)})$$
$$w^{(2)} = w^{(1)} - \gamma L'(w^{(1)})$$



It boils down to minimising a function!

$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

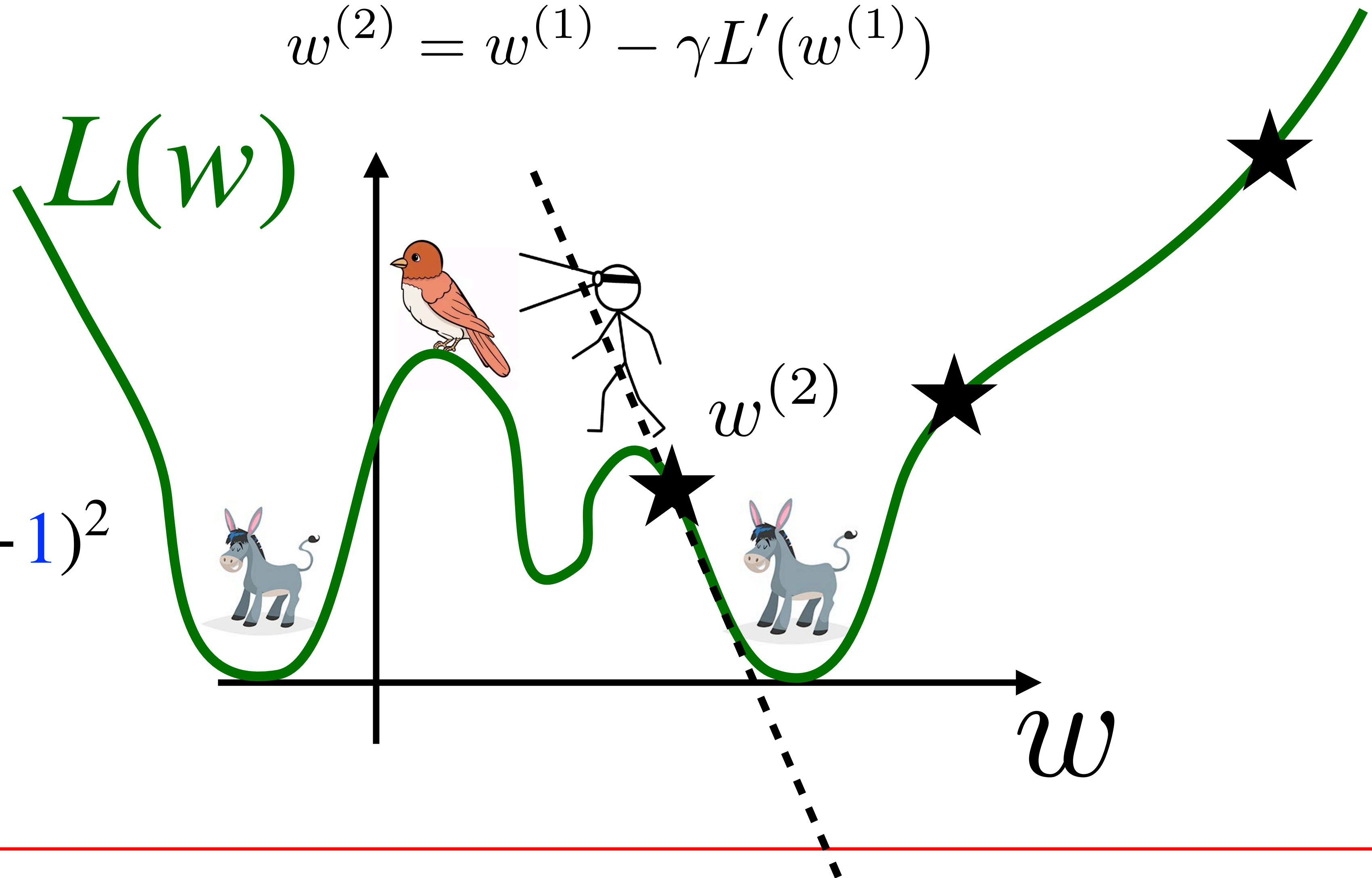
Back to 1D:



$$\min_{w \in \mathbb{R}^p} L(w)$$

$$L(w) = (f_w(x) - 1)^2$$

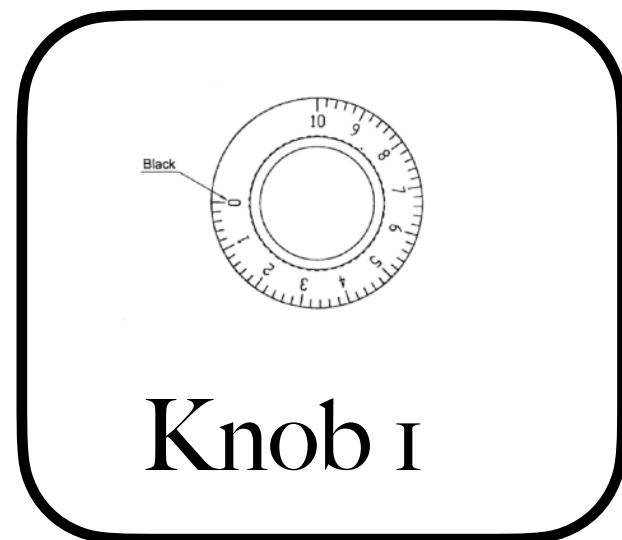
$$w^{(1)} = w^{(0)} - \gamma L'(w^{(0)})$$
$$w^{(2)} = w^{(1)} - \gamma L'(w^{(1)})$$



It boils down to minimising a function!

$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

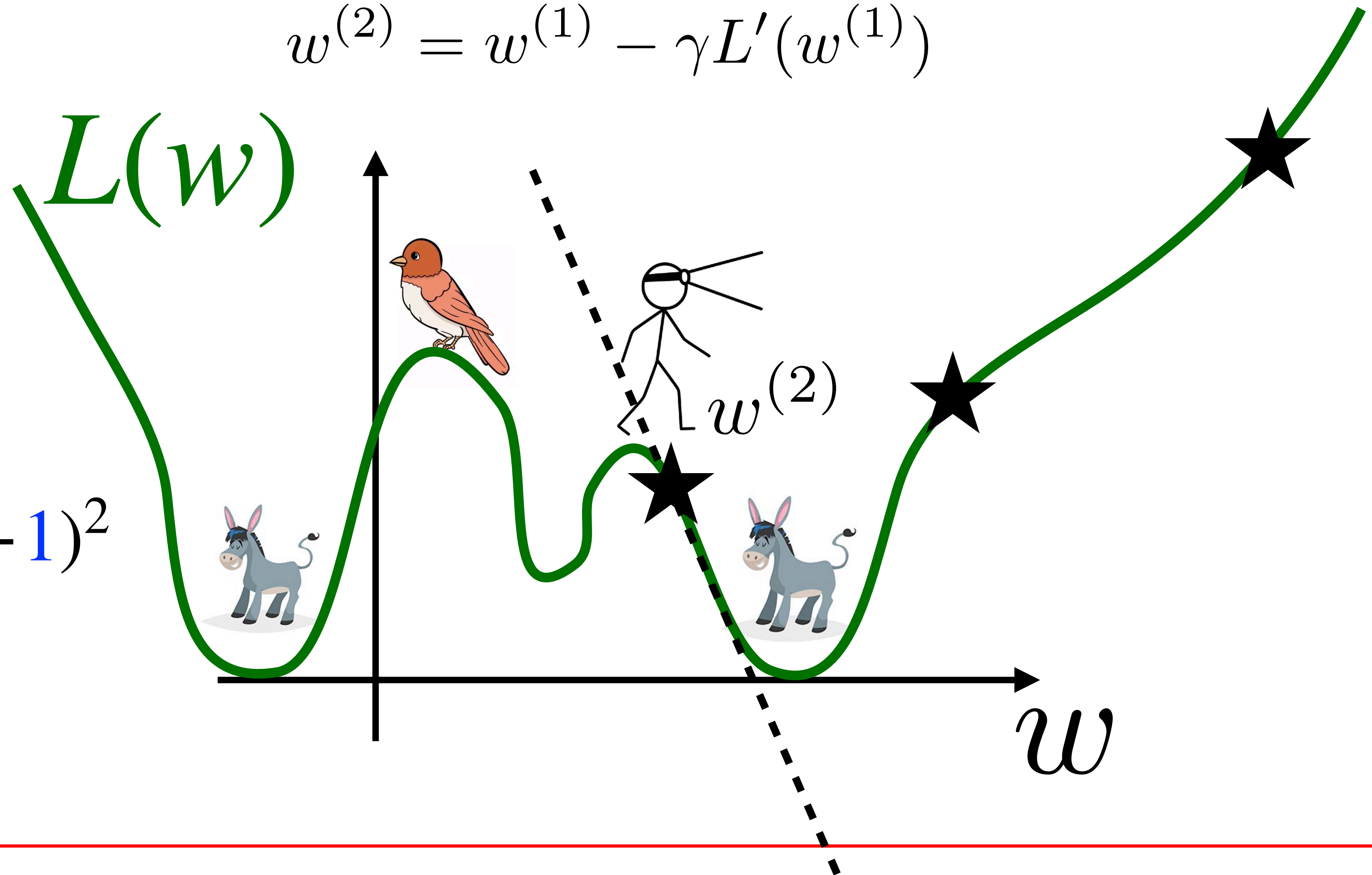
Back to 1D:



$$\min_{w \in \mathbb{R}^p} L(w)$$

$$L(w) = (f_w(x) - 1)^2$$

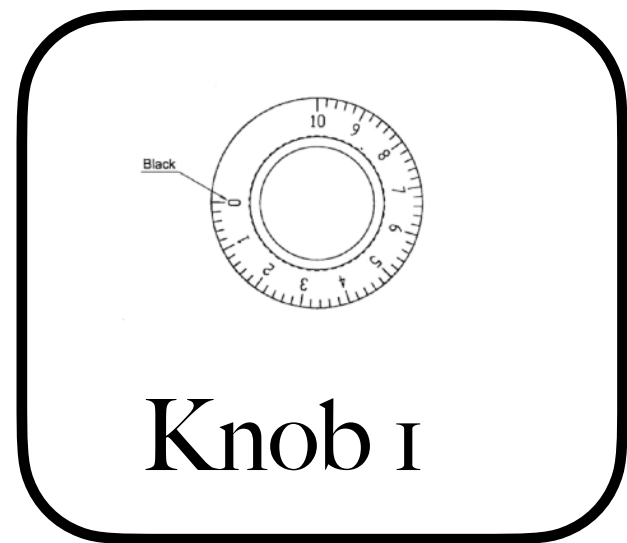
$$w^{(1)} = w^{(0)} - \gamma L'(w^{(0)})$$
$$w^{(2)} = w^{(1)} - \gamma L'(w^{(1)})$$



It boils down to minimising a function!

$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

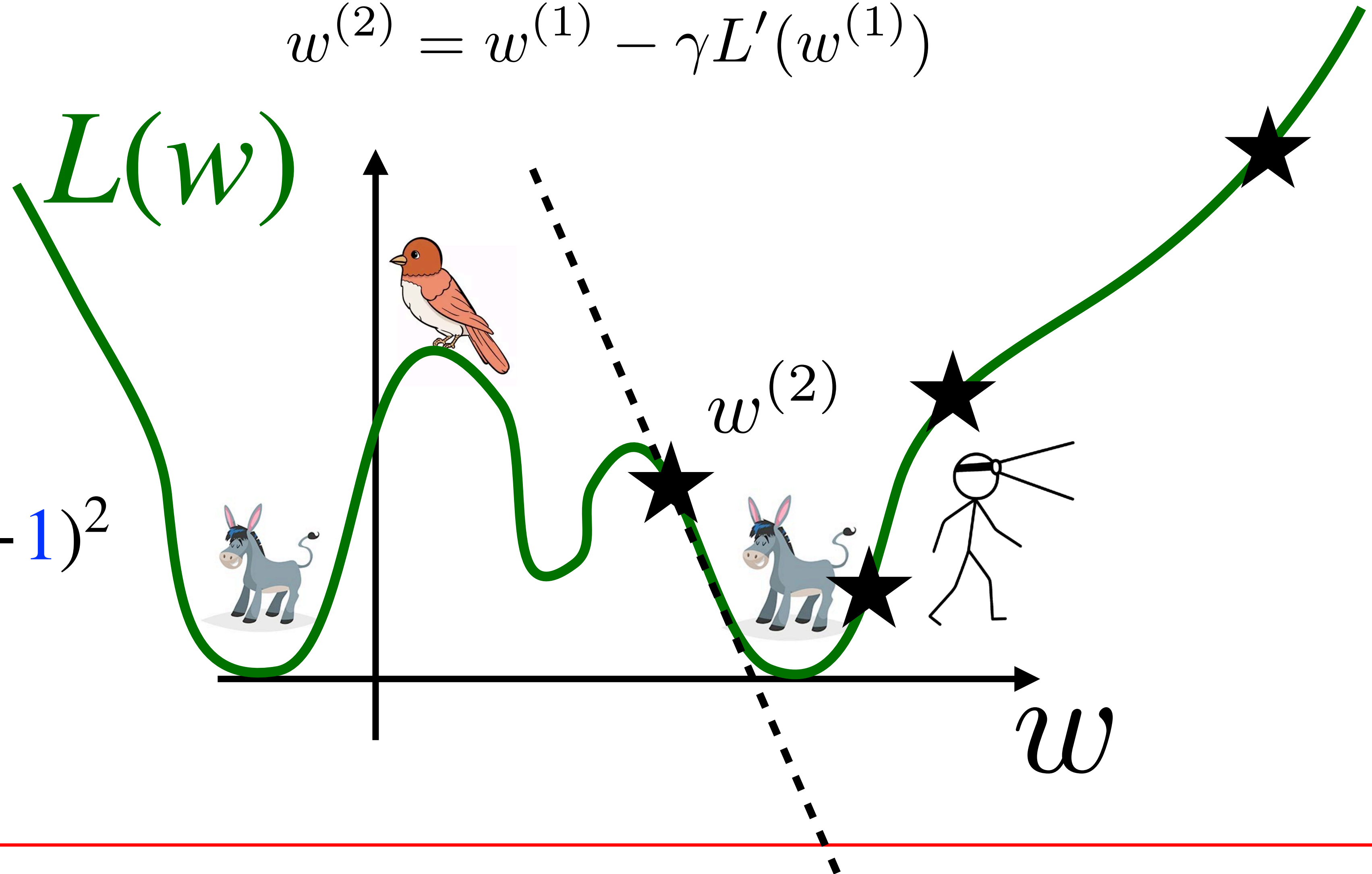
Back to 1D:



$$\min_{w \in \mathbb{R}^p} L(w)$$

$$L(w) = (f_w(x) - 1)^2$$

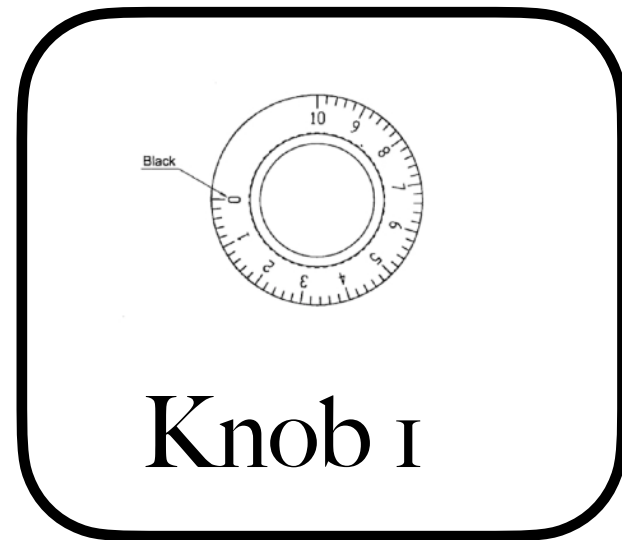
$$w^{(1)} = w^{(0)} - \gamma L'(w^{(0)})$$
$$w^{(2)} = w^{(1)} - \gamma L'(w^{(1)})$$



It boils down to minimising a function!

$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

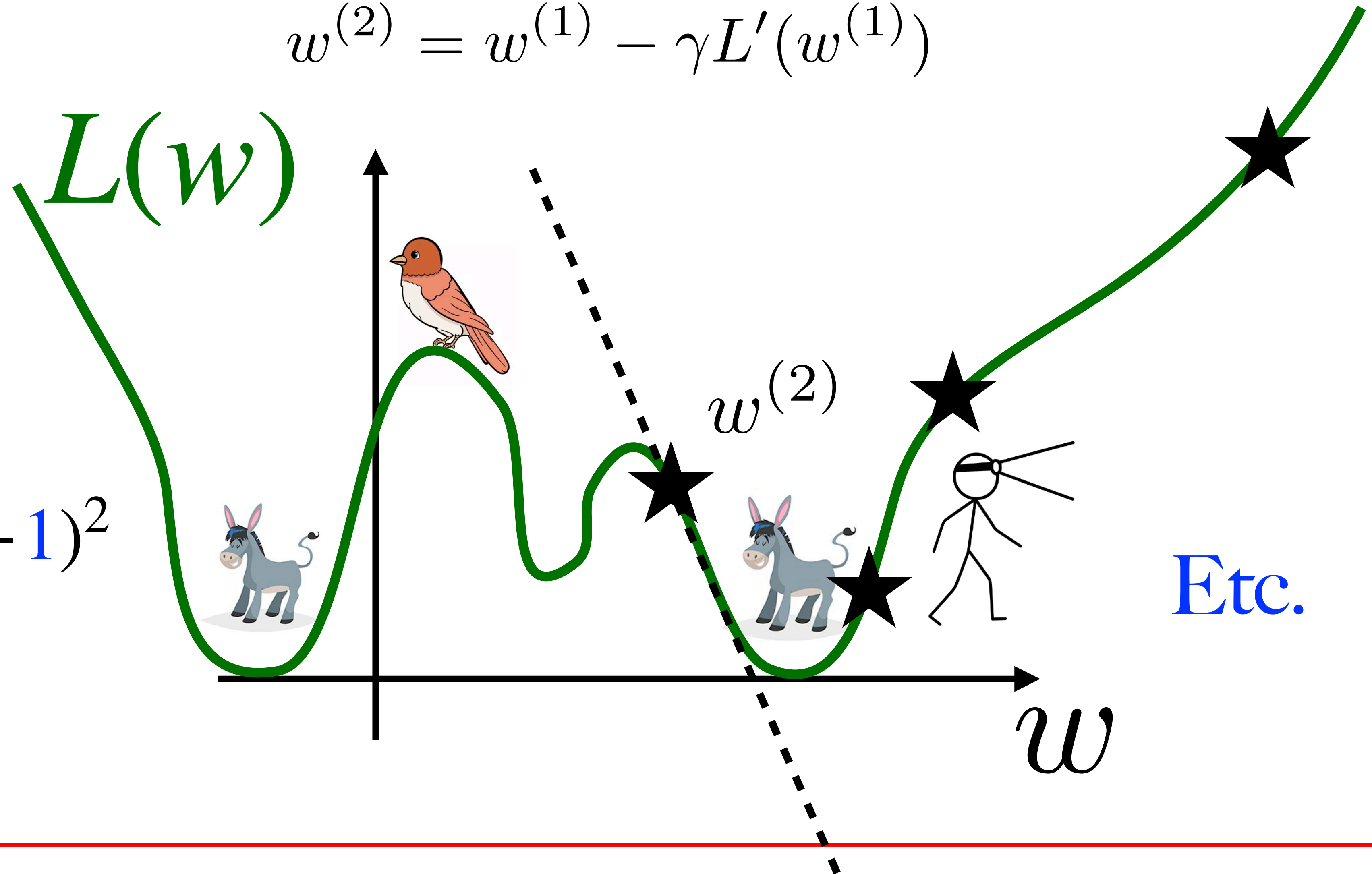
Back to 1D:



$$\min_{w \in \mathbb{R}^p} L(w)$$

$$L(w) = (f_w(x) - 1)^2$$

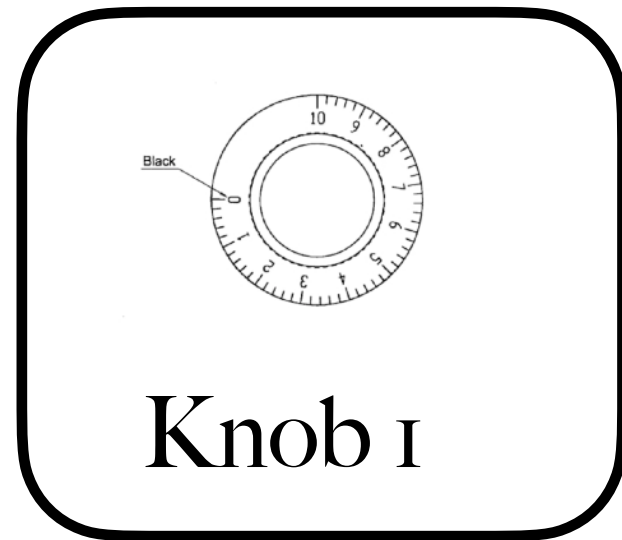
$$w^{(1)} = w^{(0)} - \gamma L'(w^{(0)})$$
$$w^{(2)} = w^{(1)} - \gamma L'(w^{(1)})$$



It boils down to minimising a function!

$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

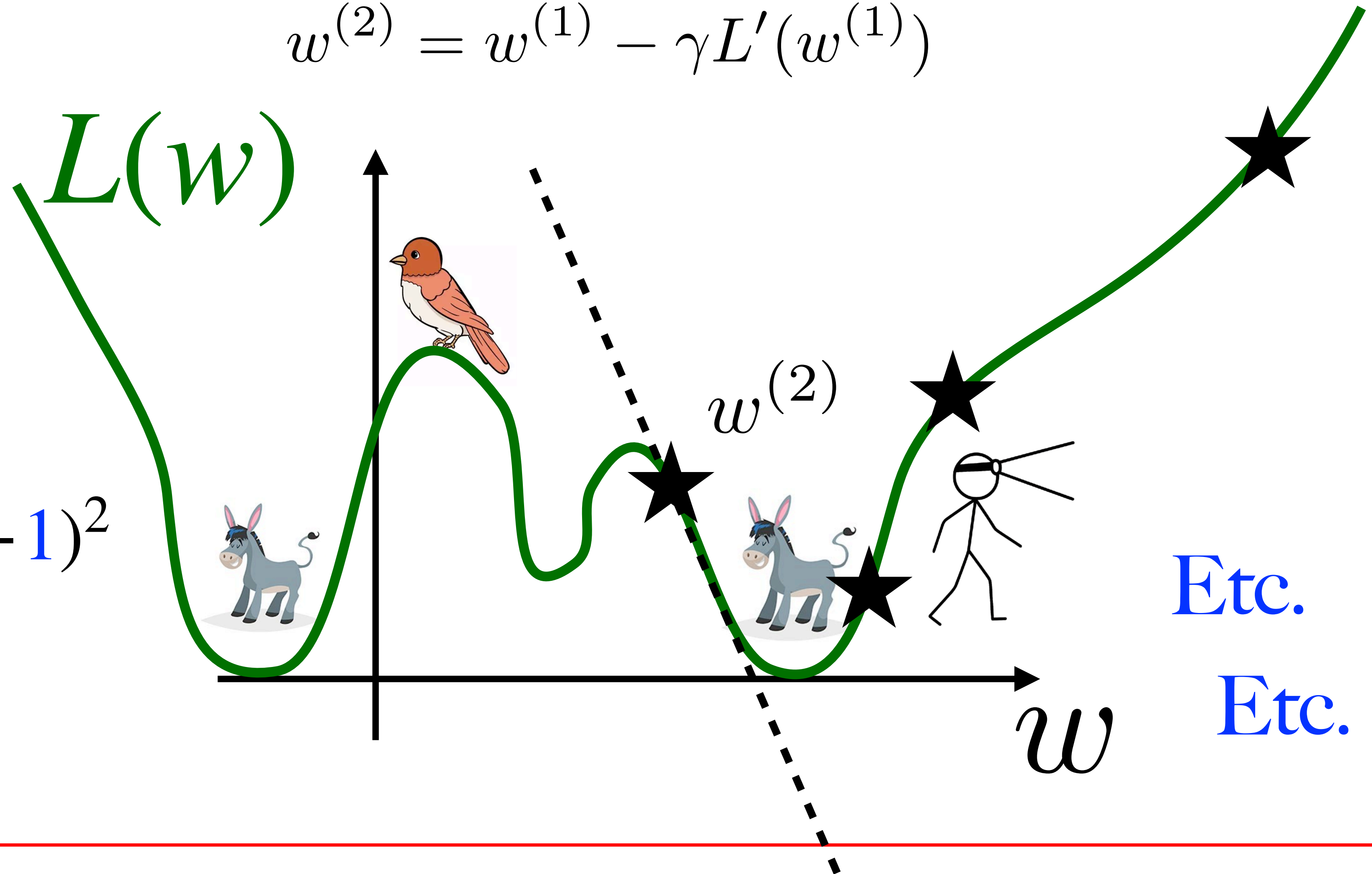
Back to 1D:



$$\min_{w \in \mathbb{R}^p} L(w)$$

$$L(w) = (f_w(x) - 1)^2$$

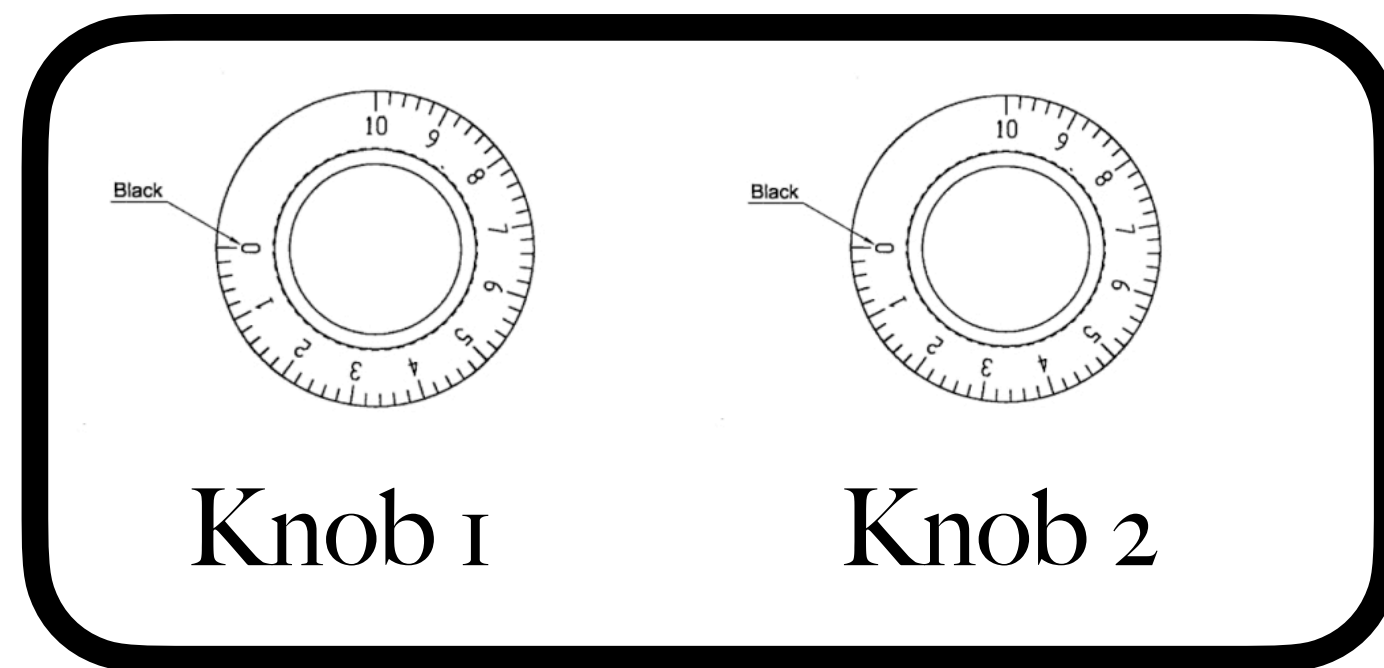
$$w^{(1)} = w^{(0)} - \gamma L'(w^{(0)})$$
$$w^{(2)} = w^{(1)} - \gamma L'(w^{(1)})$$



It boils down to minimising a function!

$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

En 2D:



$$\min_{w \in \mathbb{R}^p} L(w)$$

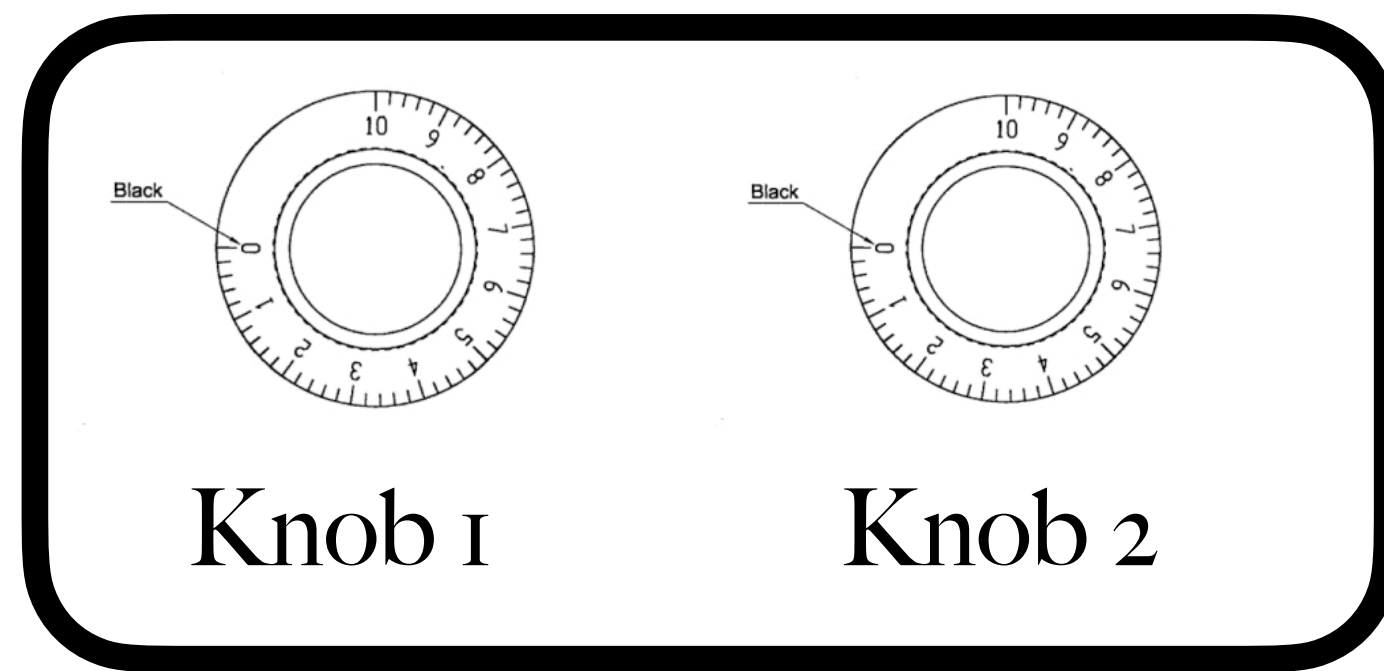
$$L(w) = (f_w(x) - 1)^2$$



It boils down to minimising a function!

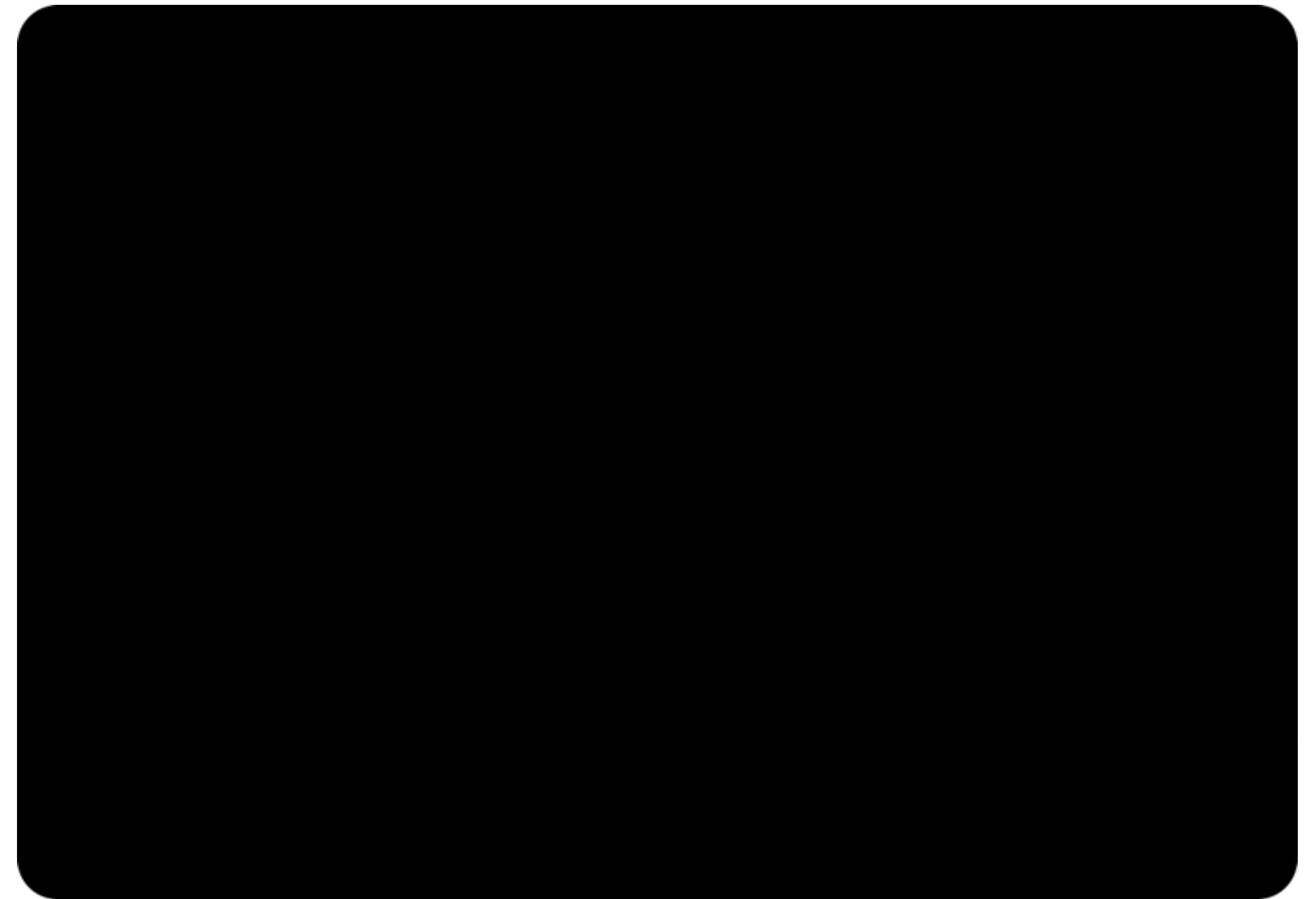
$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

En 2D:



$$\min_{w \in \mathbb{R}^p} L(w)$$

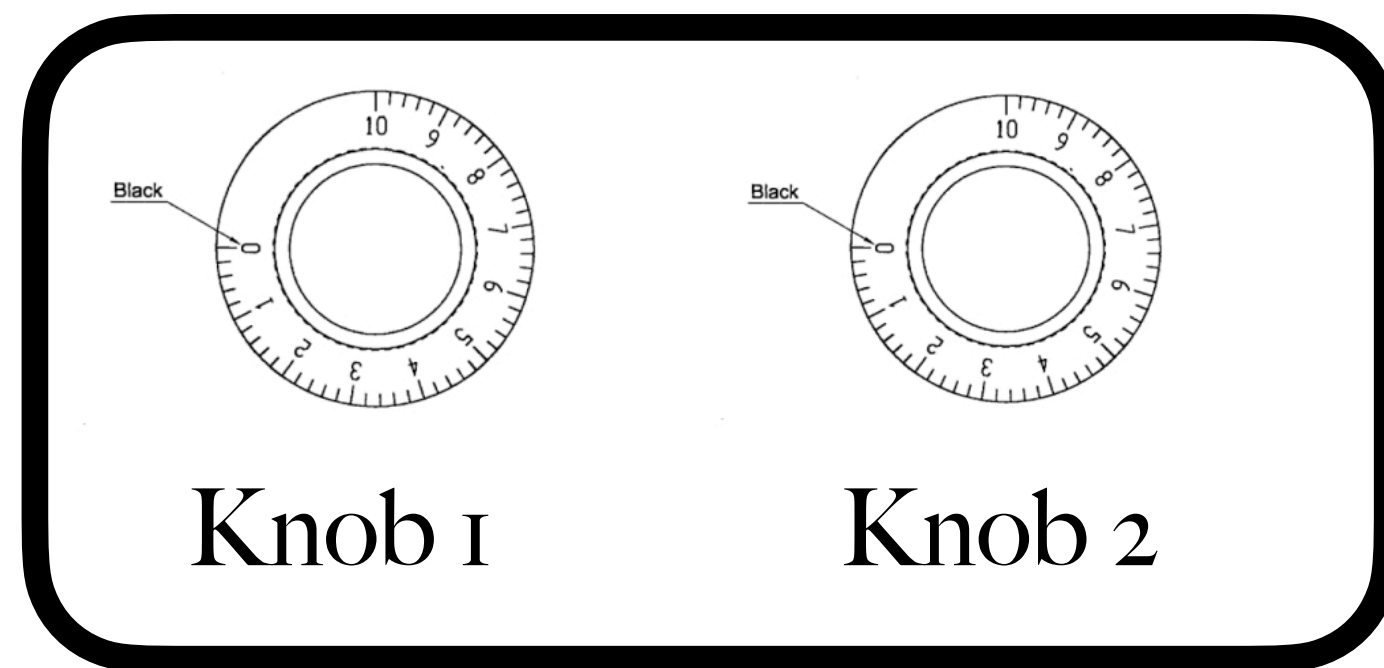
$$L(w) = (f_w(x) - 1)^2$$



It boils down to minimising a function!

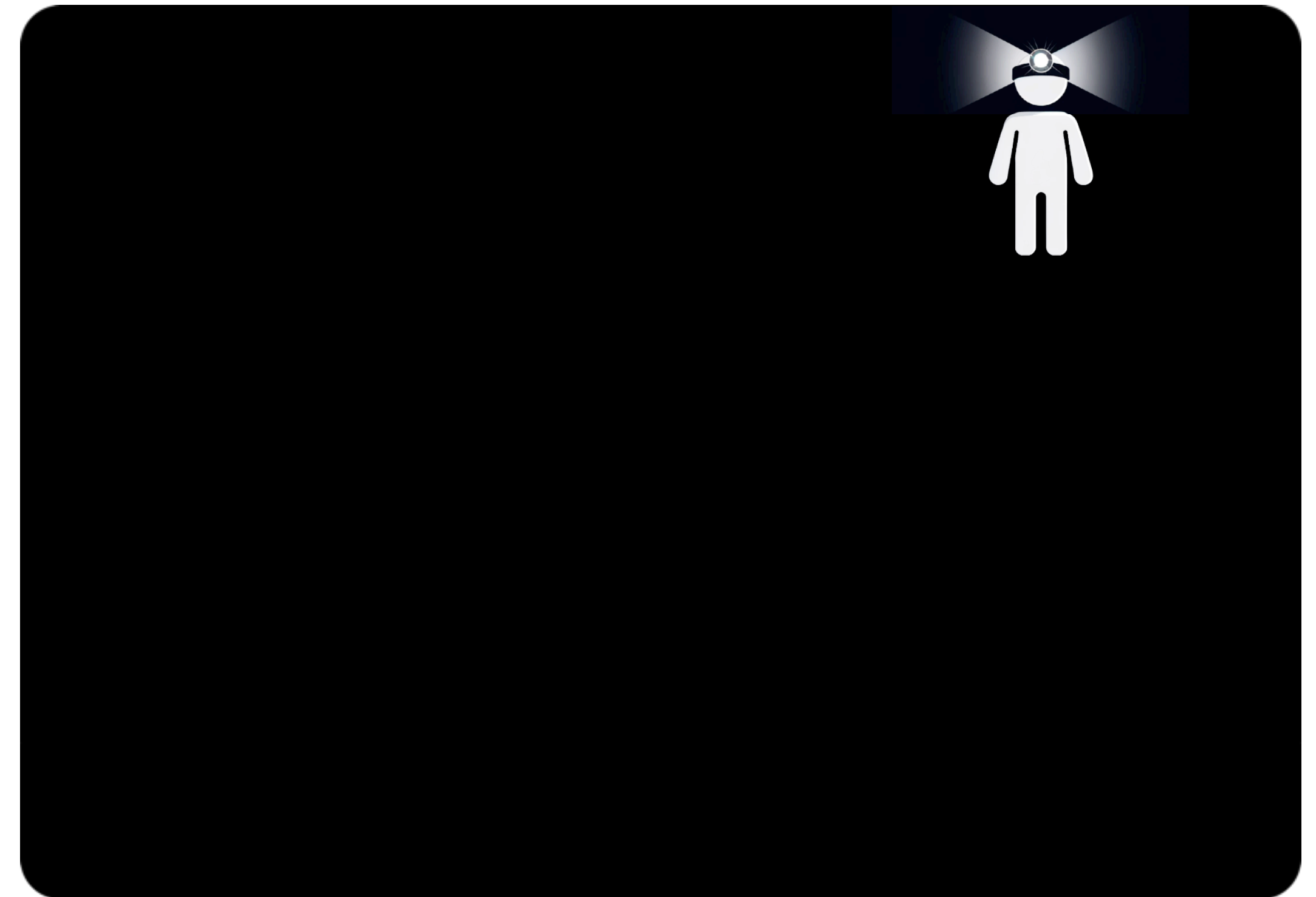
$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

En 2D:



$$\min_{w \in \mathbb{R}^p} L(w)$$

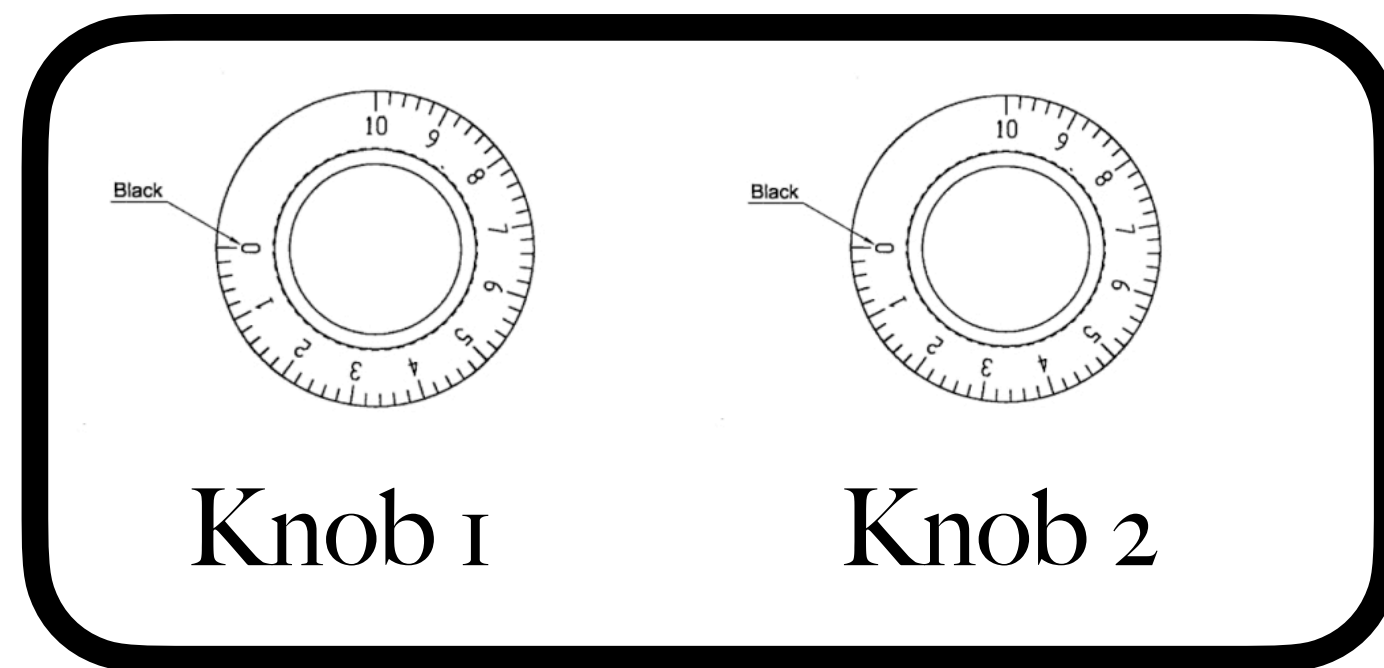
$$L(w) = (f_w(x) - 1)^2$$



It boils down to minimising a function!

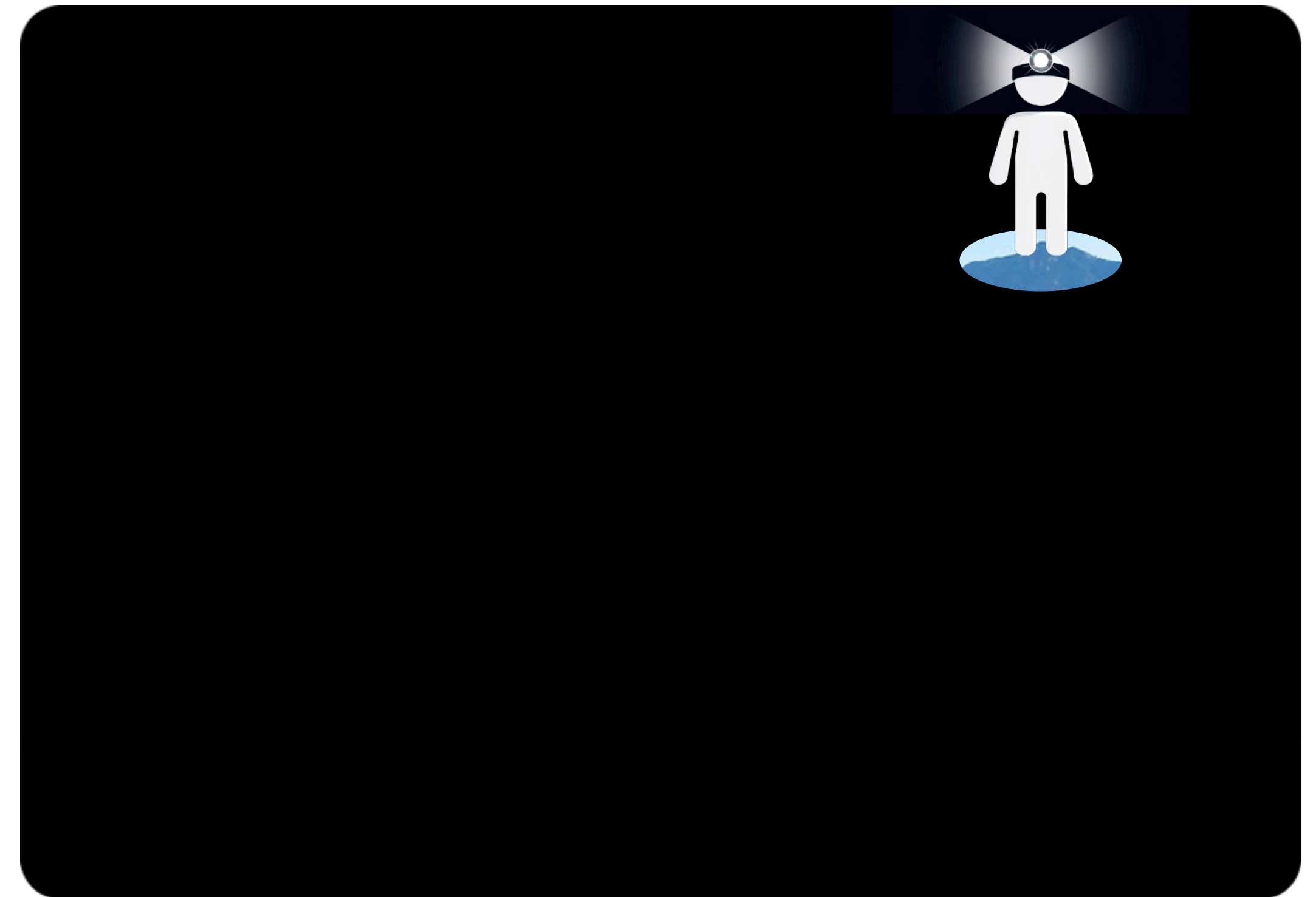
$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

En 2D:



$$\min_{w \in \mathbb{R}^p} L(w)$$

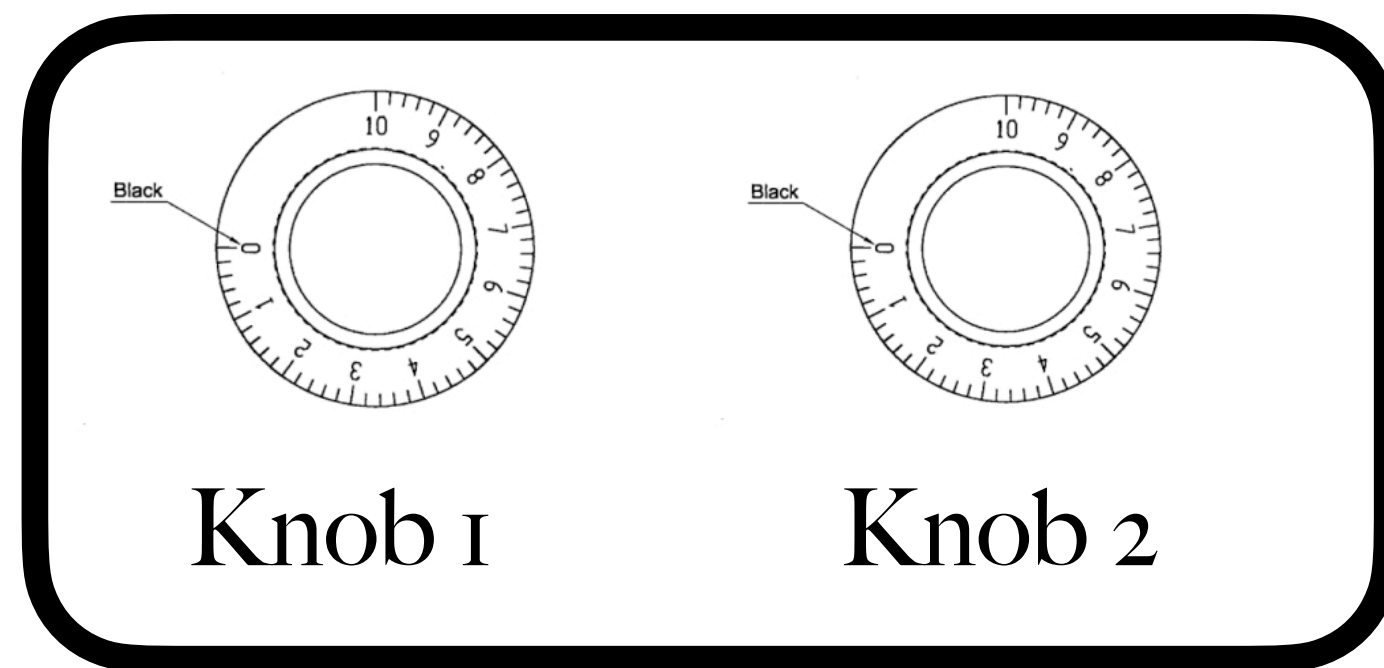
$$L(w) = (f_w(x) - 1)^2$$



It boils down to minimising a function!

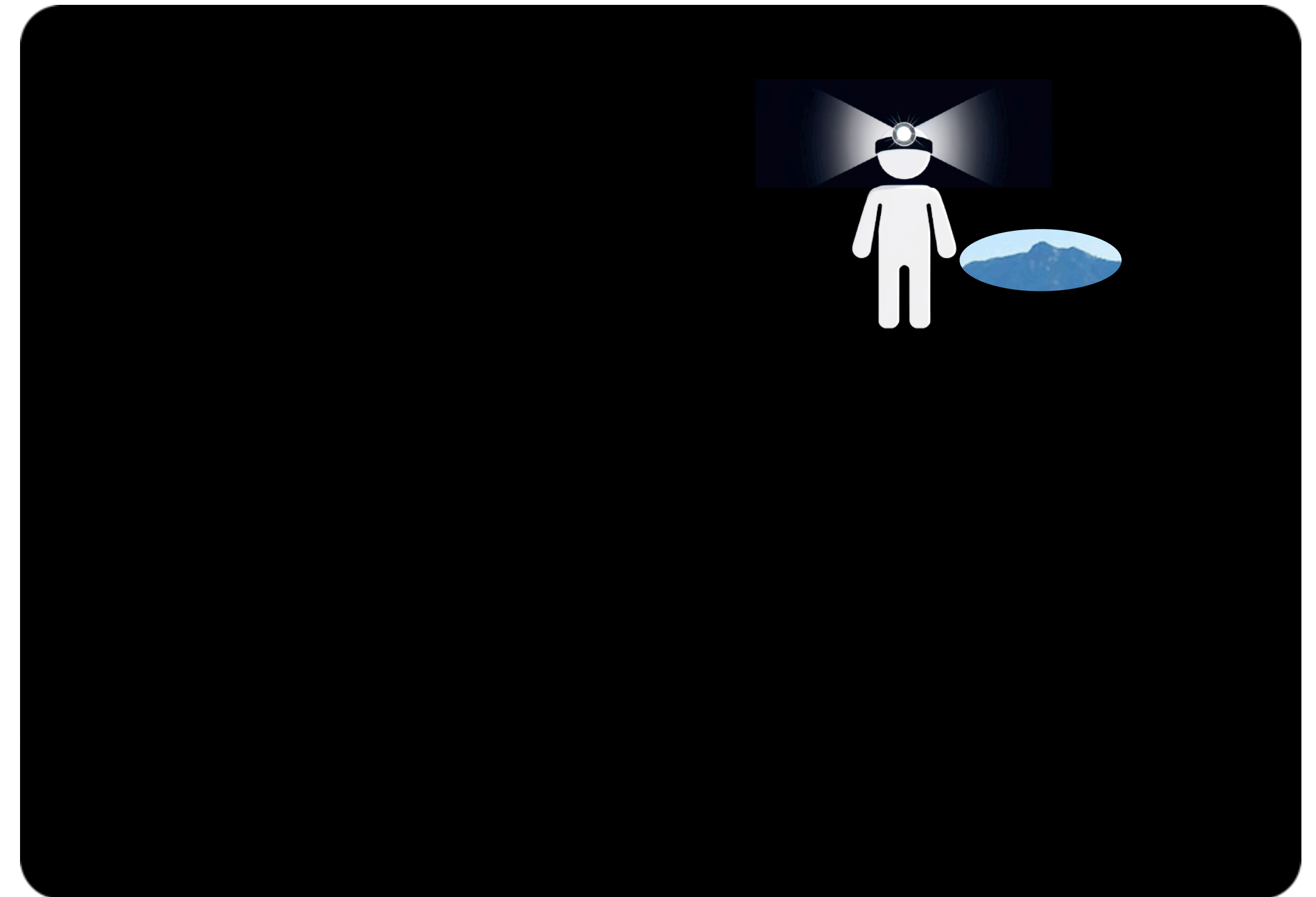
$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

En 2D:



$$\min_{w \in \mathbb{R}^p} L(w)$$

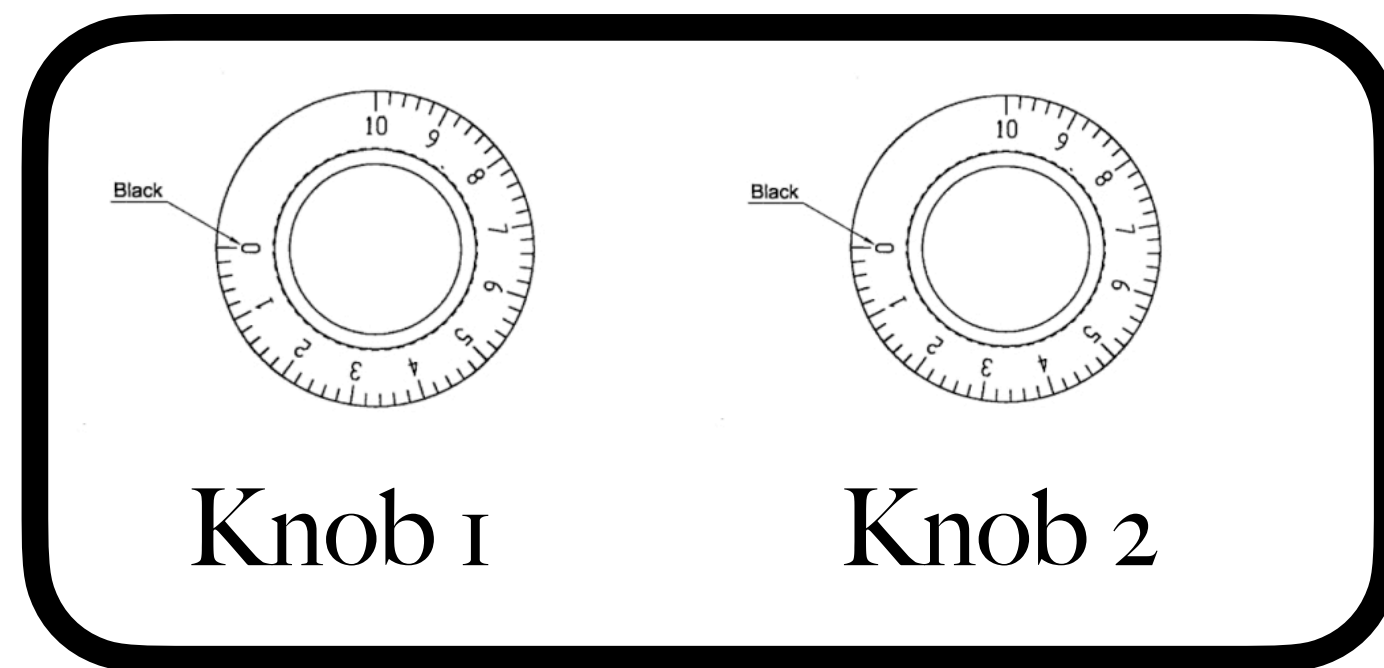
$$L(w) = (f_w(x) - 1)^2$$



It boils down to minimising a function!

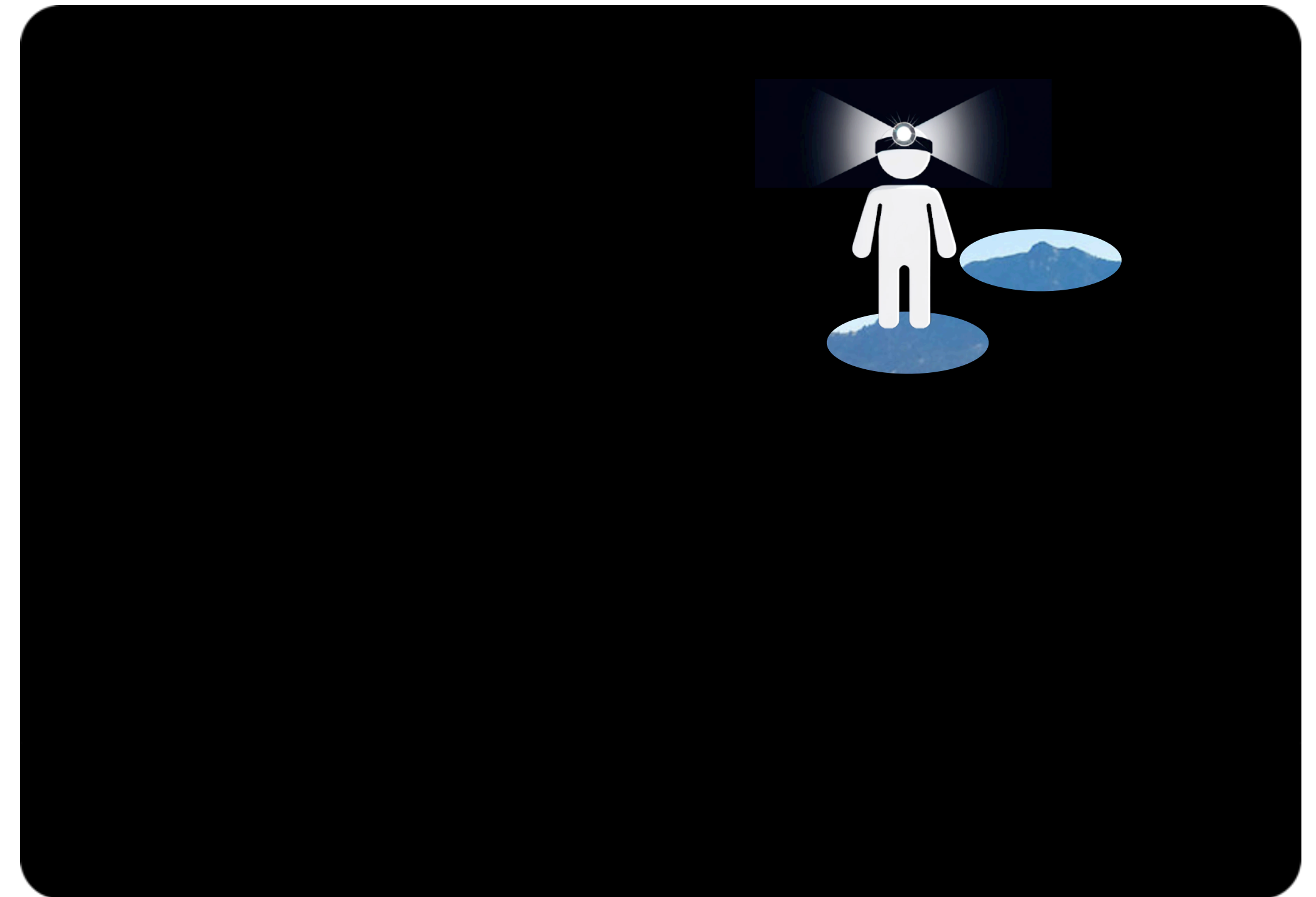
$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

En 2D:



$$\min_{w \in \mathbb{R}^p} L(w)$$

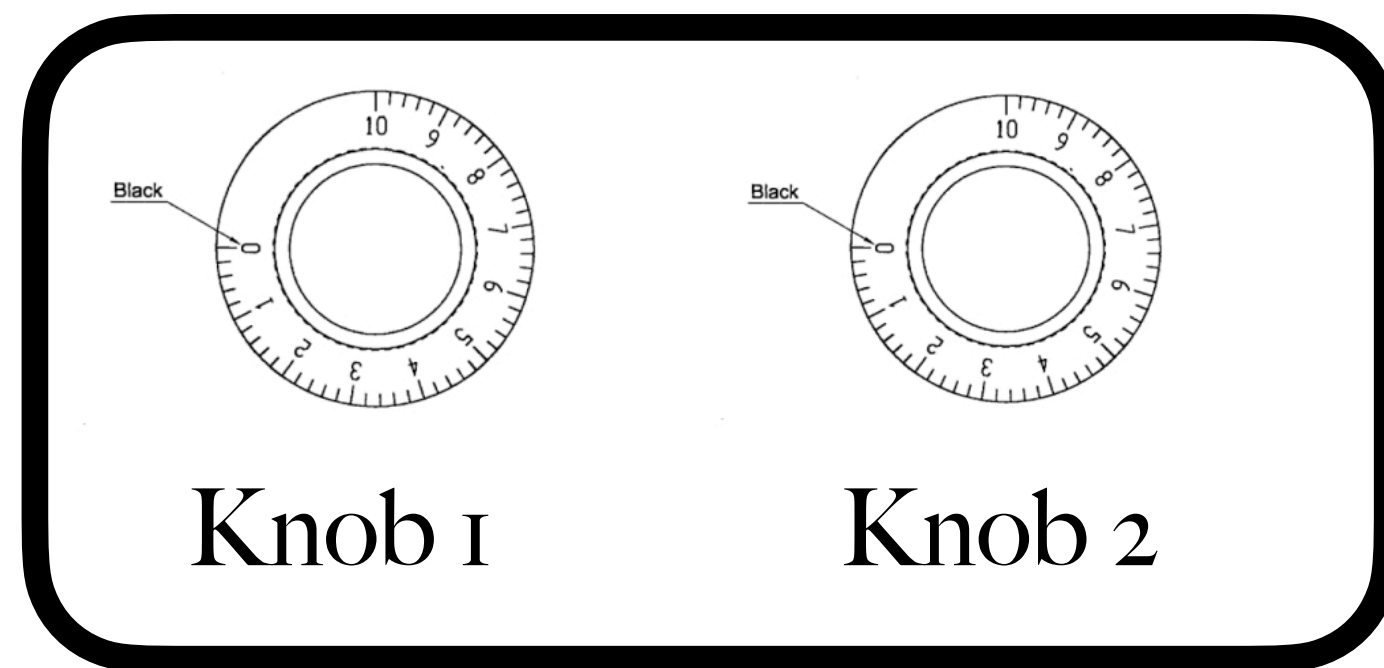
$$L(w) = (f_w(x) - 1)^2$$



It boils down to minimising a function!

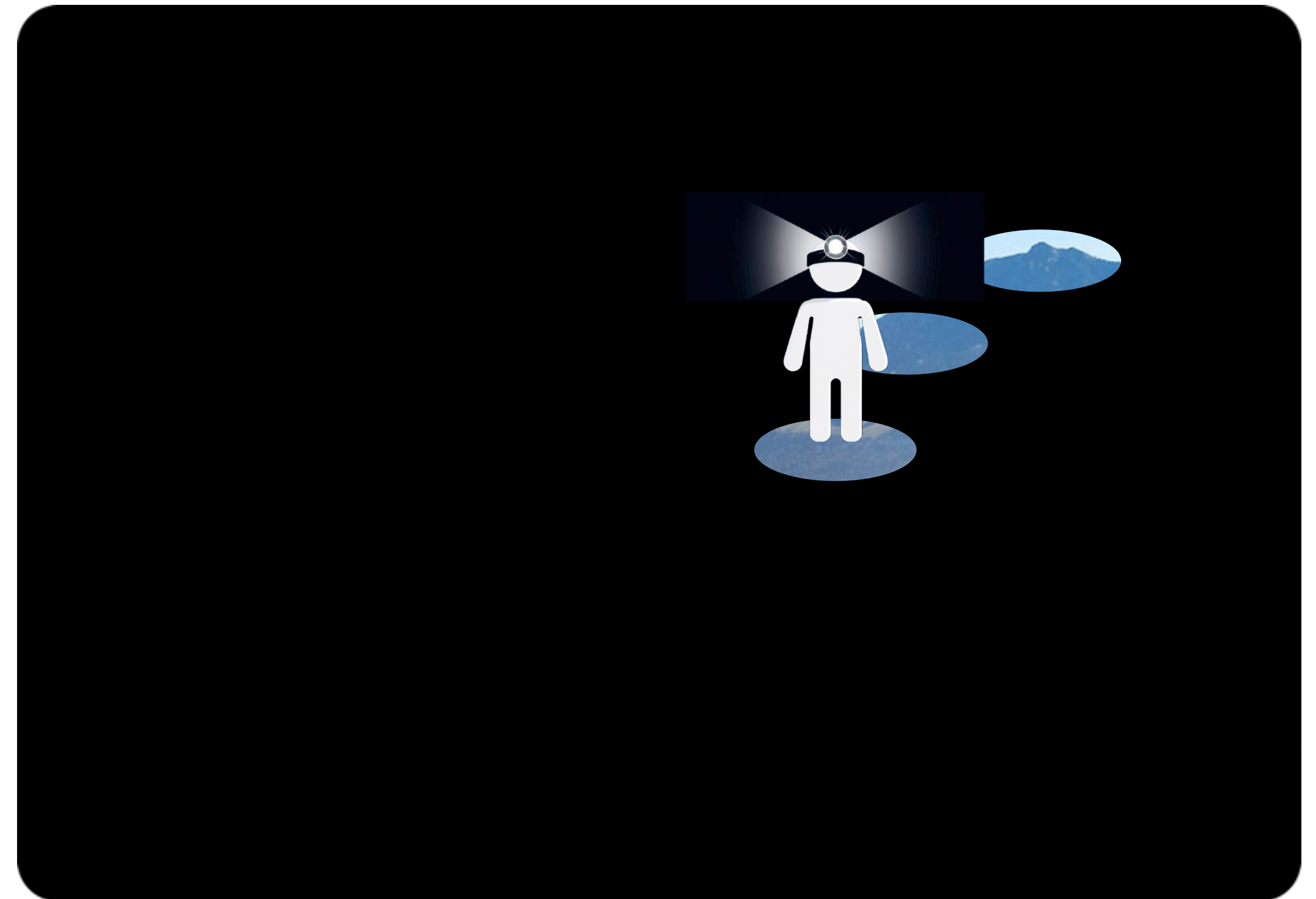
$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

En 2D:



$$\min_{w \in \mathbb{R}^p} L(w)$$

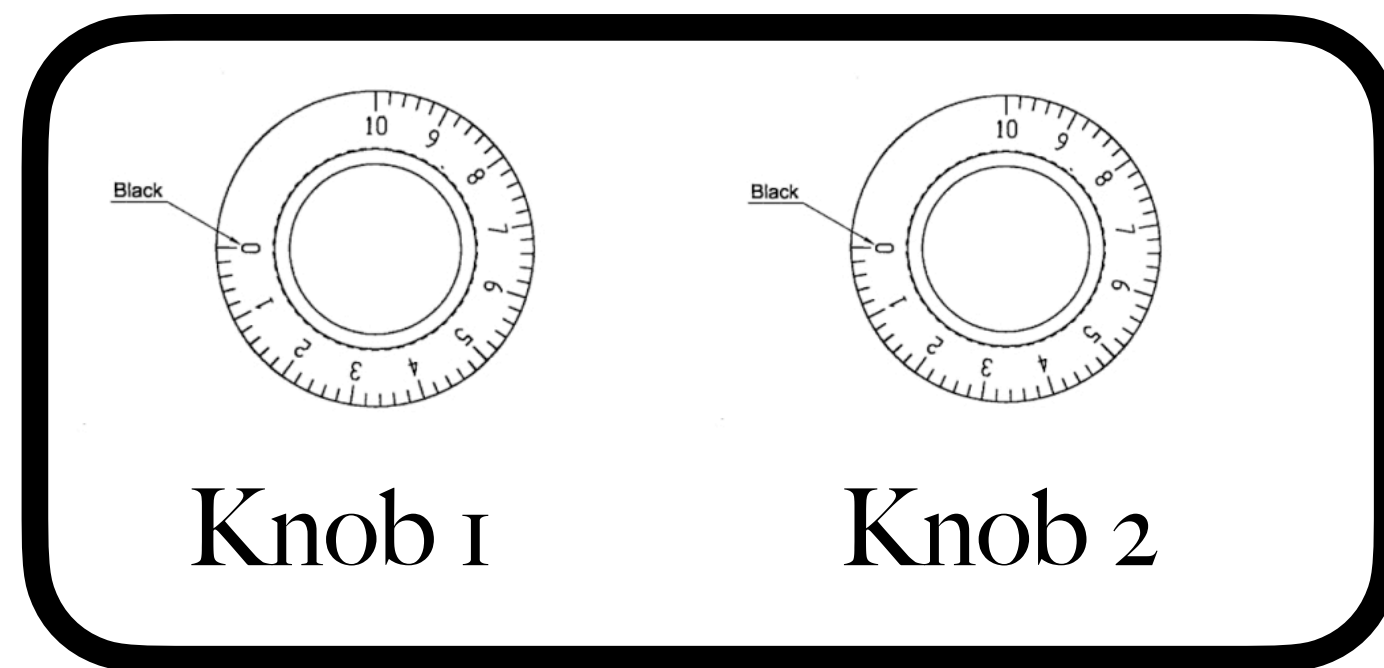
$$L(w) = (f_w(x) - 1)^2$$



It boils down to minimising a function!

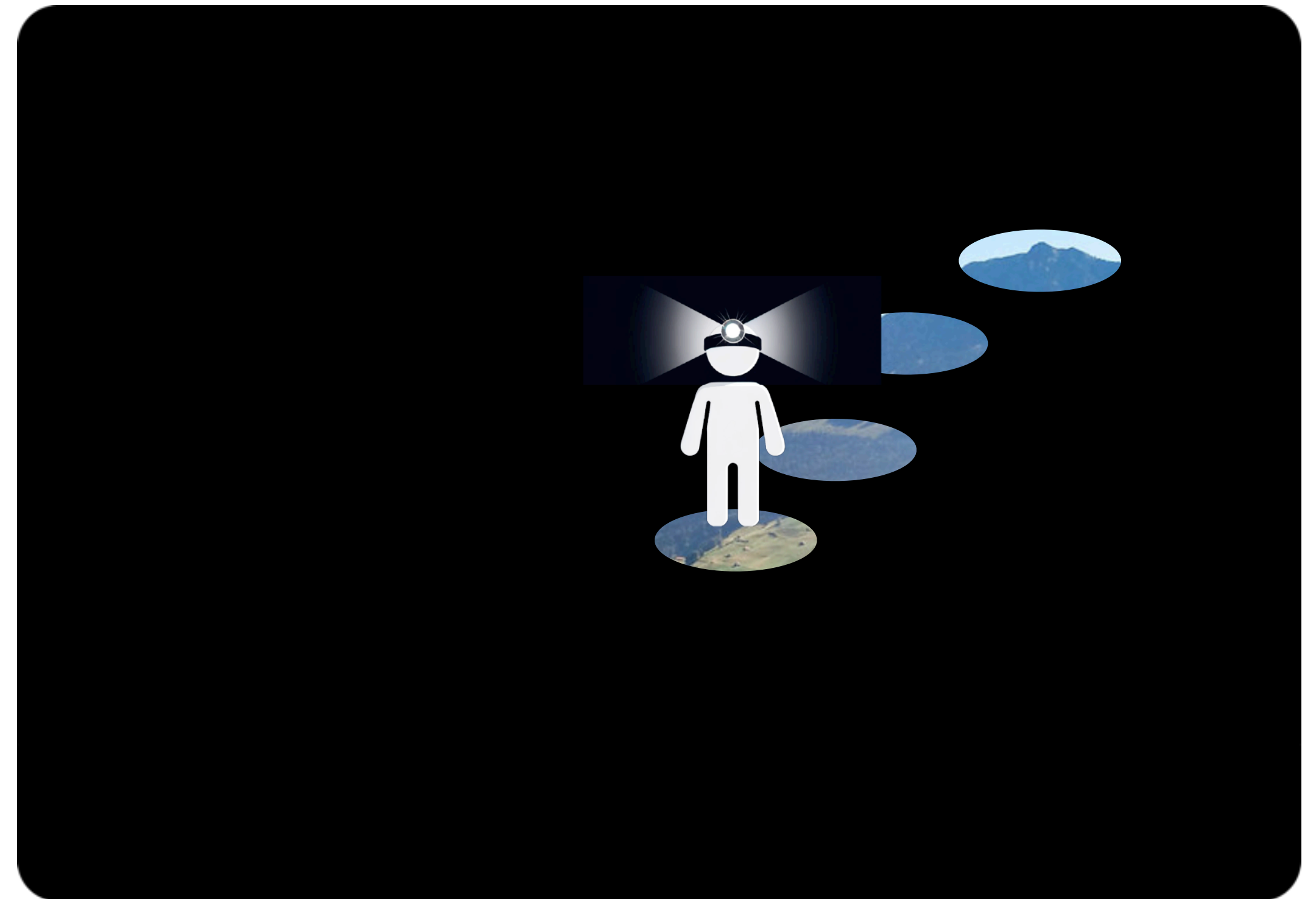
$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

En 2D:



$$\min_{w \in \mathbb{R}^p} L(w)$$

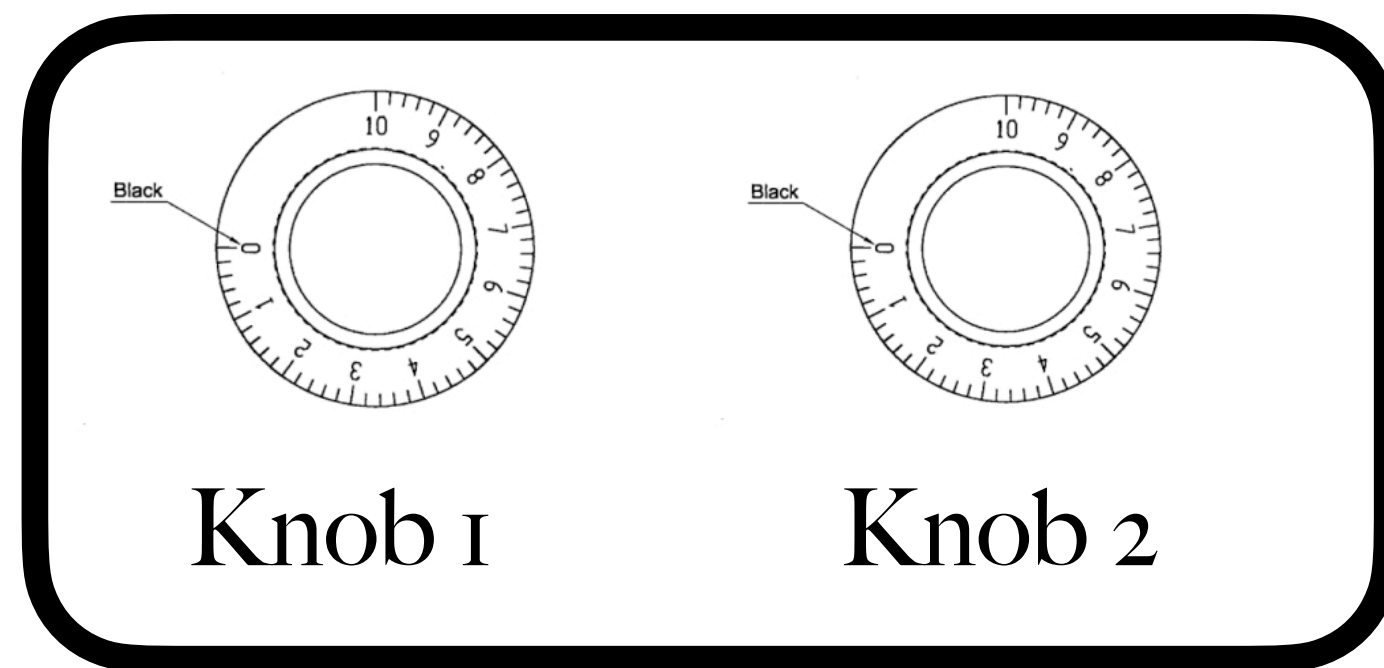
$$L(w) = (f_w(x) - 1)^2$$



It boils down to minimising a function!

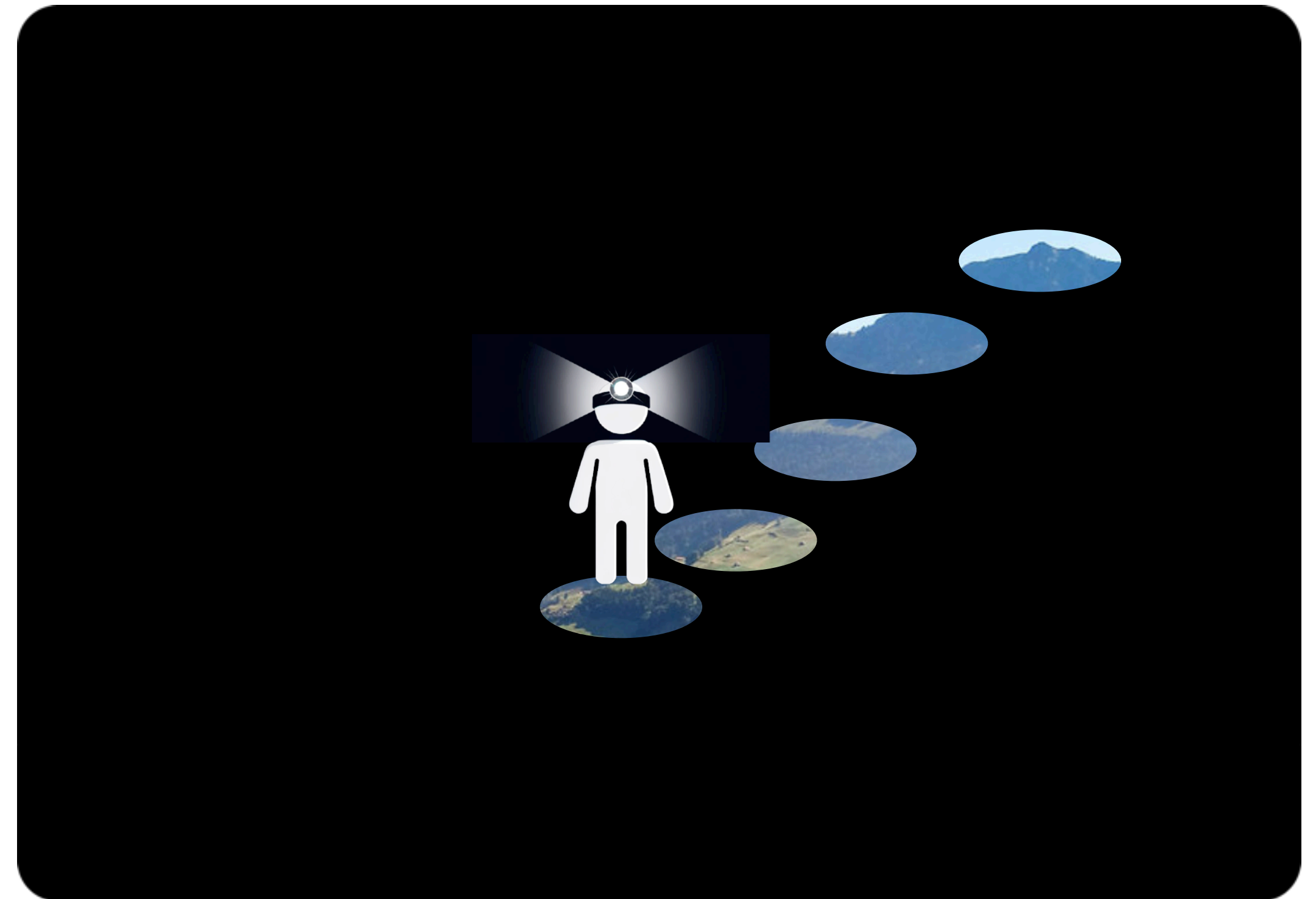
$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

En 2D:



$$\min_{w \in \mathbb{R}^p} L(w)$$

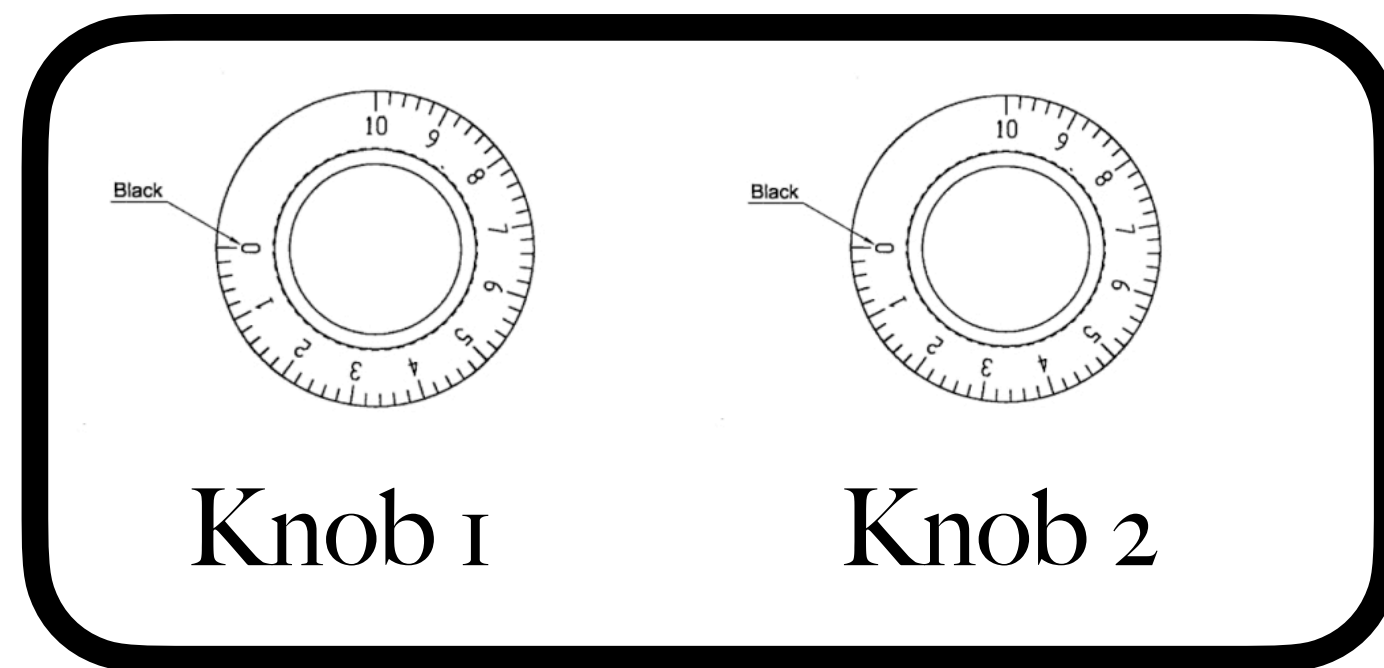
$$L(w) = (f_w(x) - 1)^2$$



It boils down to minimising a function!

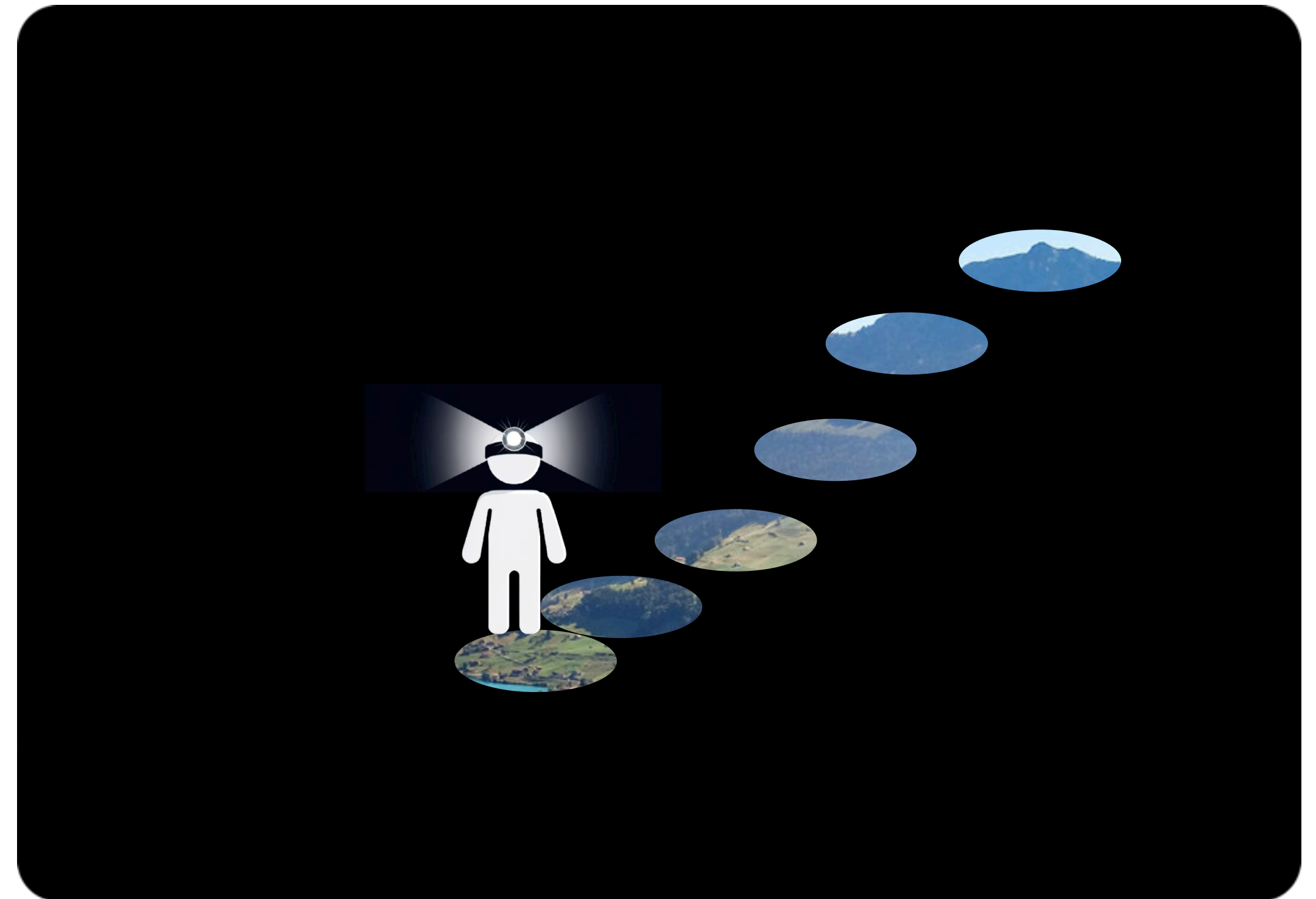
$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

En 2D:



$$\min_{w \in \mathbb{R}^p} L(w)$$

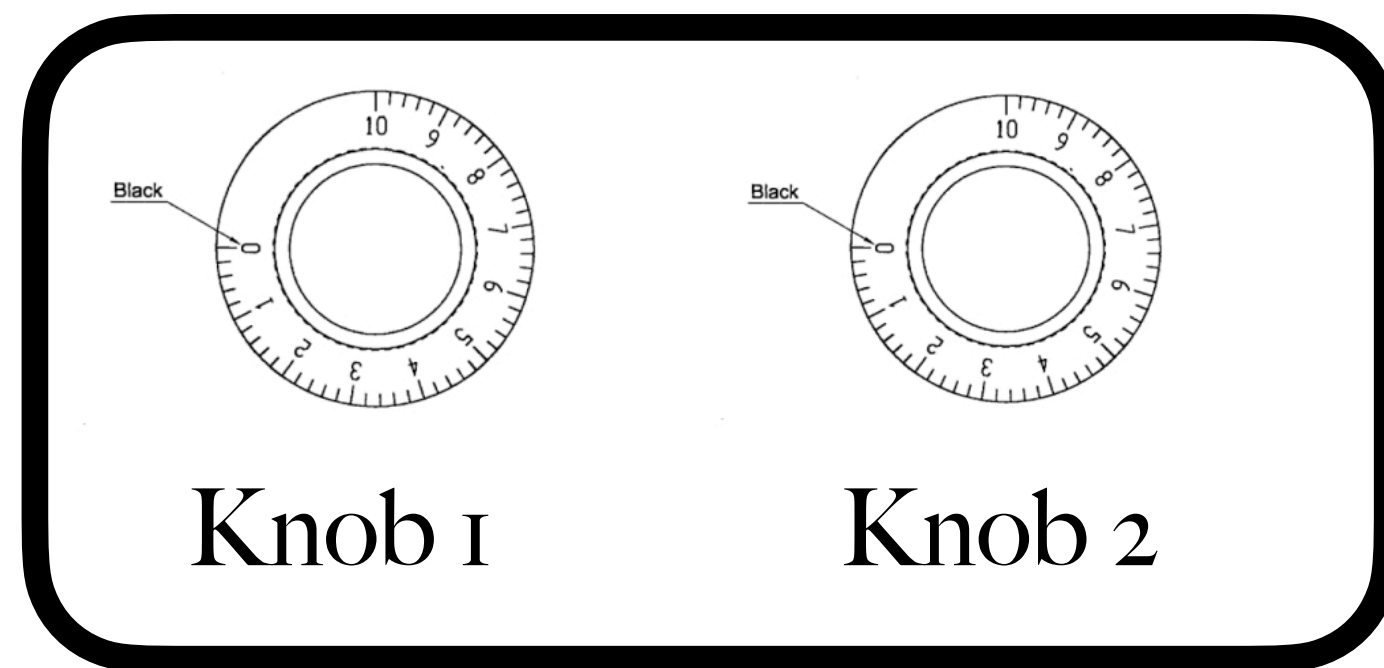
$$L(w) = (f_w(x) - 1)^2$$



It boils down to minimising a function!

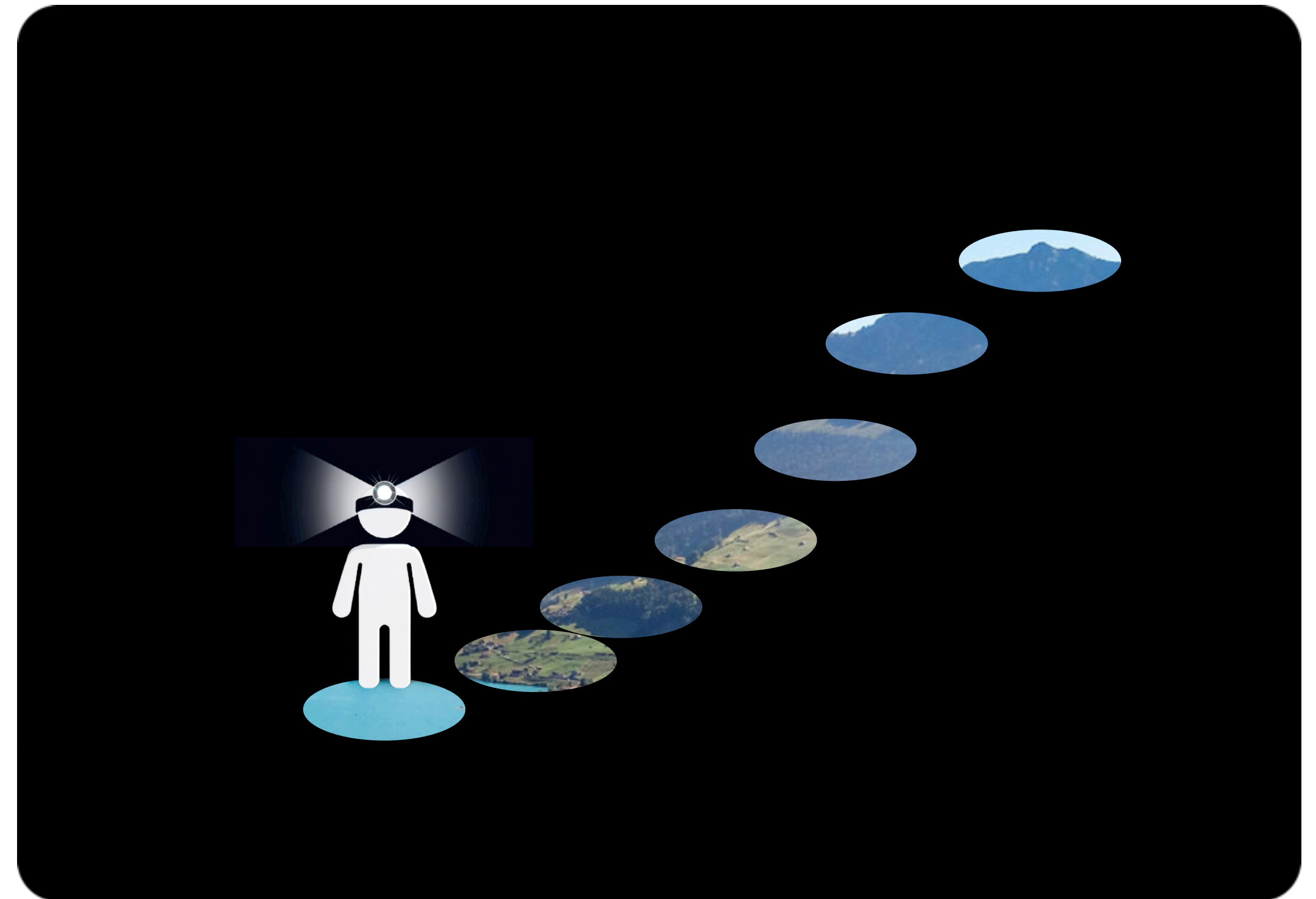
$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$

En 2D:



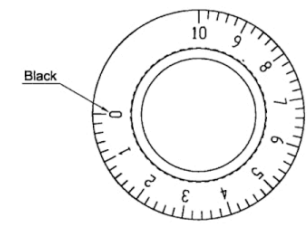
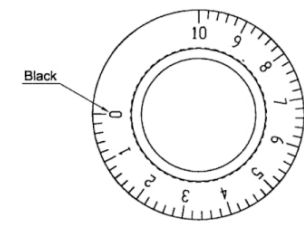
$$\min_{w \in \mathbb{R}^p} L(w)$$

$$L(w) = (f_w(x) - 1)^2$$

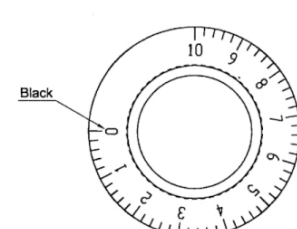


It boils down to minimising a function!

$$\min_{w \in \mathbb{R}^p} (f_w(x) - 1)^2$$



.....

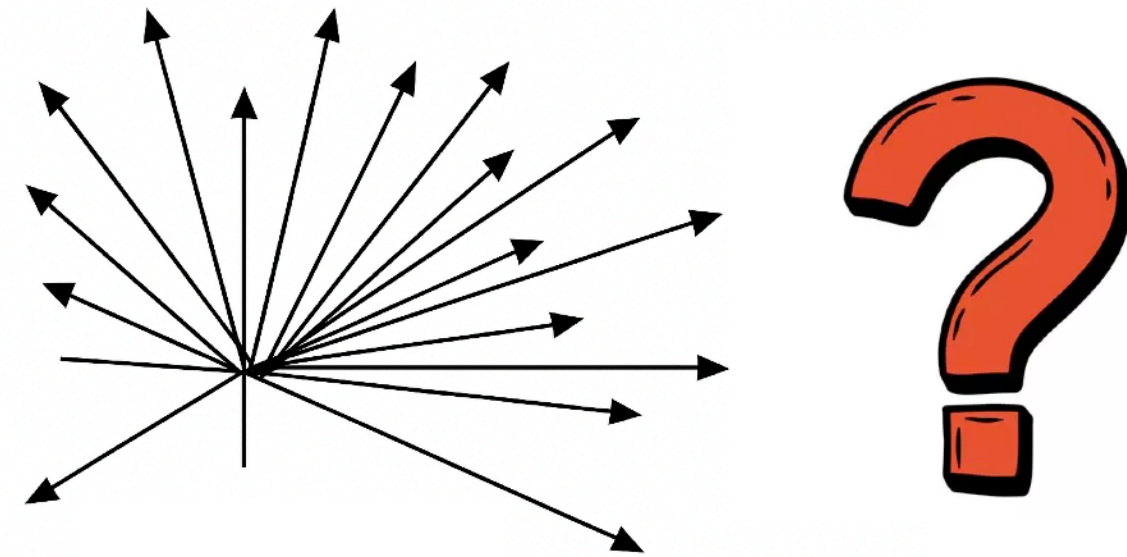


Bouton 1

Bouton 2

.....

Bouton p



$$\min_{w \in \mathbb{R}^p} L(w)$$

$$L(w) = (f_w(x) - 1)^2$$

Gradient descent method:

$$w^{(k+1)} = w^{(k)} - \gamma \nabla L(w^{(k)})$$

New knob
Configuration

=

Old
configuration

+

How much
to turn them

×

In which
direction

If “everything goes well”: $L(w^{(k)}) \xrightarrow{k \rightarrow \infty} 0$ $f_{w^{(k)}}(x) \xrightarrow{k \rightarrow \infty} 1$

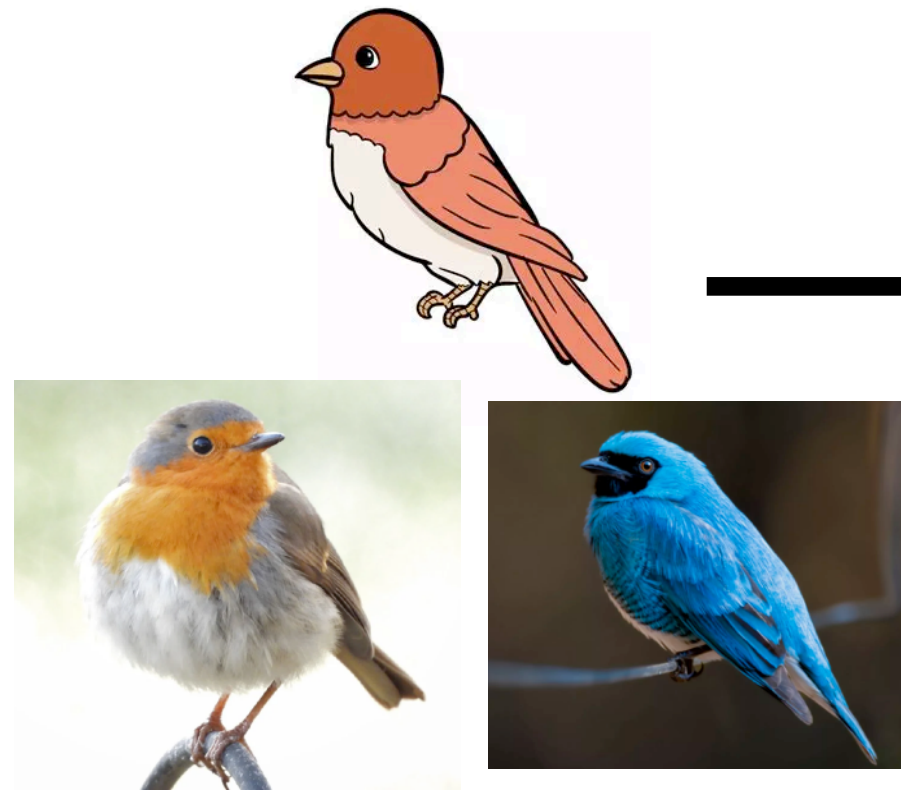


But there is more than one image!

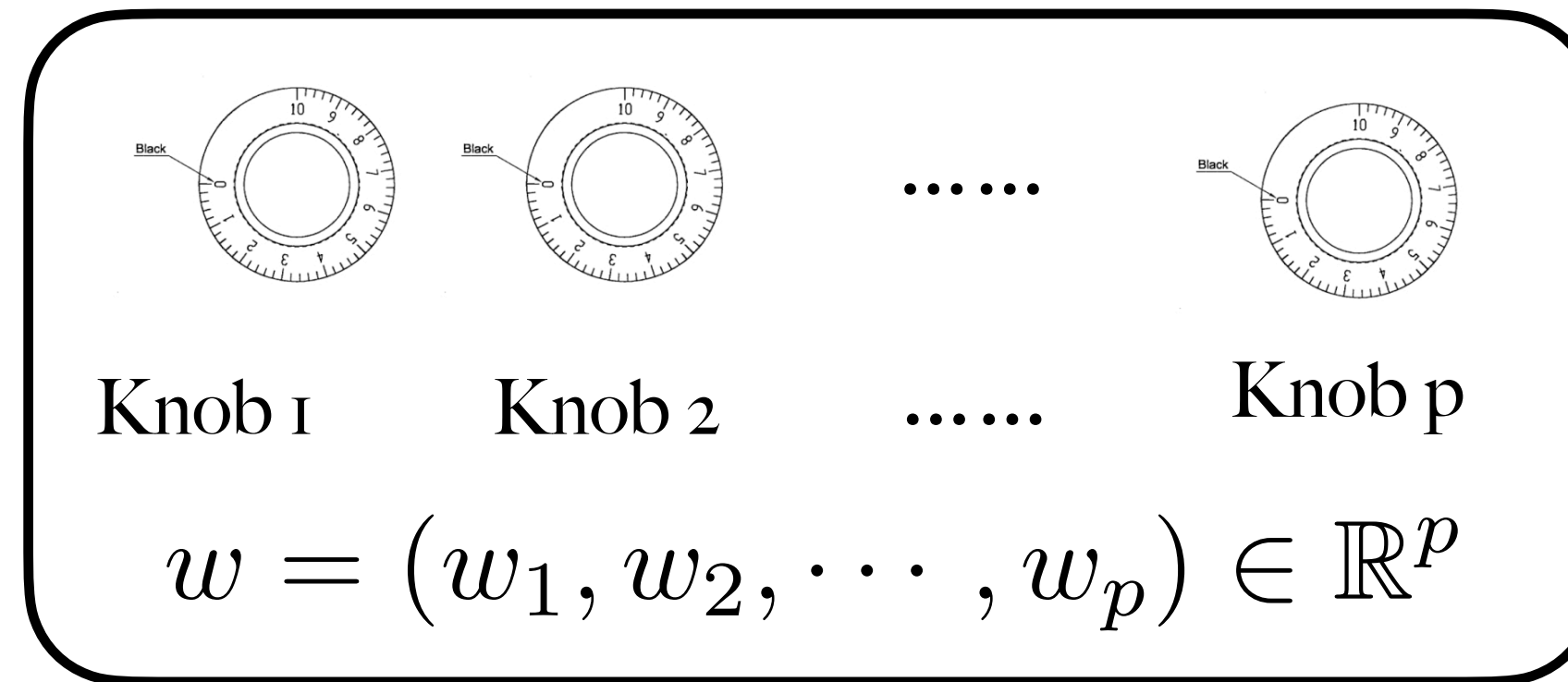
'Donkey'



'Bird'



Magic synthesiser f_w



Machine's
Response

$$f_w(x) \in \mathbb{R}$$

Data:

$$(x_1, y_1), \dots, (x_n, y_n)$$

1.2	3.5	0.7	4.1	2.7	1.3
0.6	2.8	3.3	1.1	0.5	4.9
7.2	2.4	1.1	0.9	3.3	2.2
3.7	0.9	4.8	5.5	0.2	1.6
1.0	3.1	2.2	2.7	4.0	0.8
0.3	1.5	6.4	2.9	3.6	1.0

+1

3.1	0.8	2.4	5.2	1.9	3.6
4.7	1.2	0.9	3.4	2.8	5.0
0.5	3.3	4.1	1.8	2.0	0.7
2.6	5.5	1.1	0.4	3.9	2.3
3.0	2.7	5.8	1.5	0.6	4.2
1.4	3.6	0.2	4.5	2.1	5.3

-1

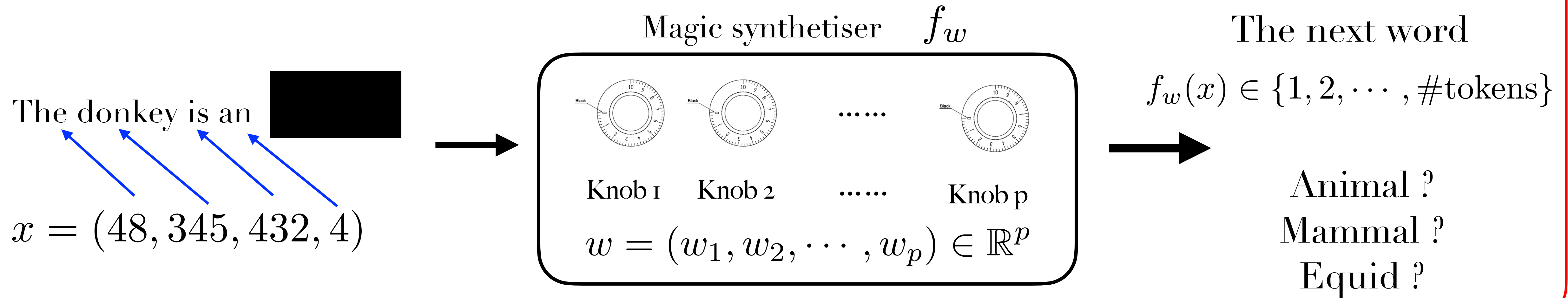
Loss to minimise:

$$\min_w \sum_{i=1}^n (y_i - f_w(x_i))^2$$

“Supervised learning”

What about ChatGPT?

The donkey is an equid. As a herbivore, it frequently consumes fibrous plants.



Data (sentences):

$$x_1, x_2, \dots, x_n$$

Function to minimise

$$\min_w \sum_{i=1}^n (\text{mot d'après} - f_w(x_i))^2$$

“Self-supervised” learning

Taking a step back

Three ingredients behind the success of modern machine learning:

Three ingredients behind the success of modern machine learning:

Data: internet (text et images)

Three ingredients behind the success of modern machine learning:

Data: internet (text et images)

Compute power: graphics processing units (GPU)

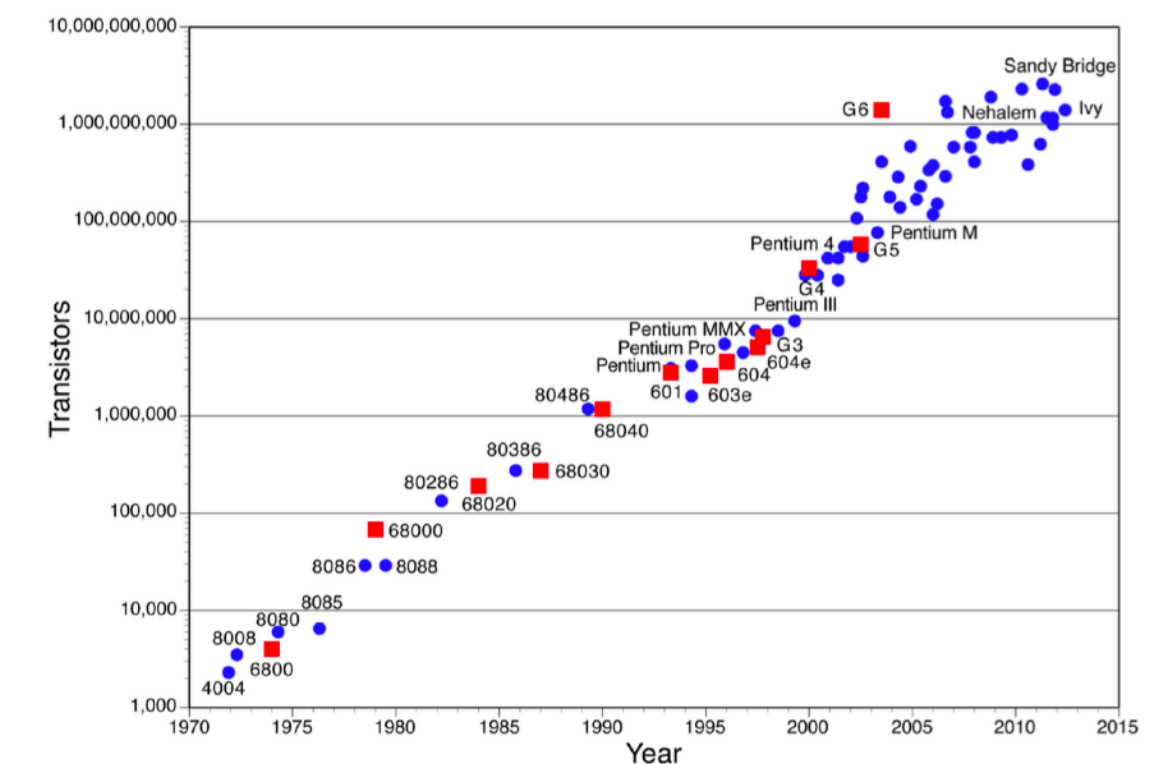
Three ingredients behind the success of modern machine learning:

Data: internet (text et images)

DeepBlue
1 tonne



Moore's Law



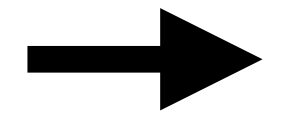
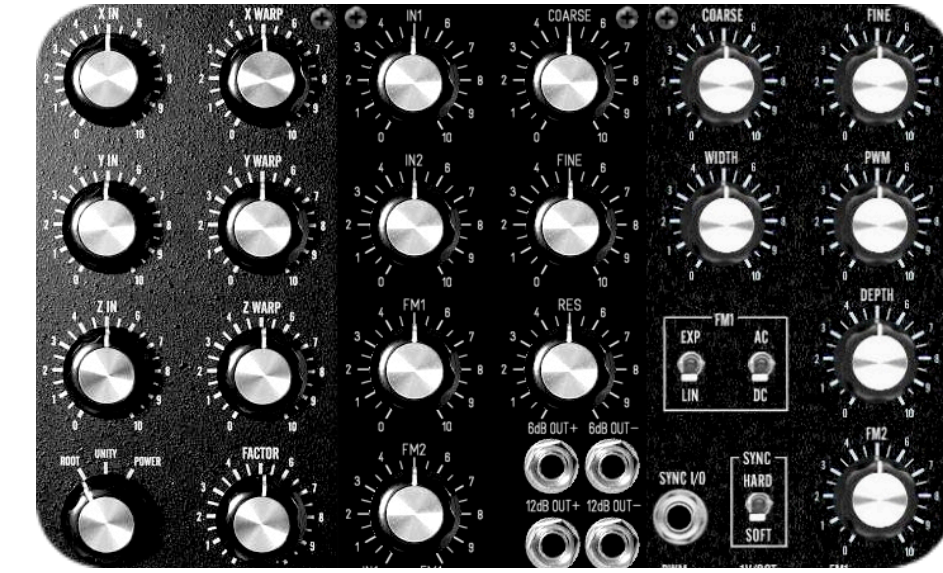
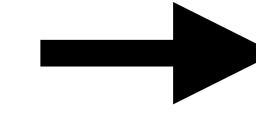
Compute power: graphics processing units (GPU)

Architectures et training techniques:

Neural networks, training algorithms, ...

Training Chat-GPT, some numbers

What's the capital
of Switzerland?



Bern

Size of the synthesiser:

A trillion (10^{12}) trainable knobs $\approx$ Paris covered with knobs

Energy to turn the knobs: 10GWh energy consumption (€2M euros electricity bill)

To compare: *Bringing up a child to its twenties:*
2000Kcal / day $\rightarrow$ 100Watts
20 MWh from 0 to 20 years old
therefore: Training ChatGPT $\approx$ 500 x 

0 $\rightarrow$ 20

But not much compared to the $\approx$ 1GWh daily consumption due to queries!

Text seen by ChatGPT:

Trained on a 10 trillion (10^{13}) words $\approx$ 10 million copies of War and Peace $\approx$ 100 000 years to read (no sleeping)

Small quizz !

True or False:

The brain has more connections (synapses) than ChatGPT?

Philosophie des modèles de langage: The Bigger, The Better

GPT (2018)

Taille du cerveau

100 millions de paramètres
Cerveau de drosophile



Taille du jeu d'entraînement

1 milliards de mots
0.1 mot / personne sur Terre

GPT3 (2020)

Taille du cerveau

200 milliards de paramètres
Cerveau de pieuvre



Taille du jeu d'entraînement

250 milliards de mots
30 mots / personne sur Terre

GPT4 (2024)

Taille du cerveau

2000 milliards de paramètres
Cerveau de rat



Taille du jeu d'entraînement

10 mille milliards de mots
5 pages / personne sur Terre

Philosophie des modèles de langage: The Bigger, The Better

GPT (2018)

Taille du cerveau

100 millions de paramètres
Cerveau de drosophile



Taille du jeu d'entraînement

1 milliards de mots
0.1 mot / personne sur Terre

GPT3 (2020)

Taille du cerveau

200 milliards de paramètres
Cerveau de pieuvre



Taille du jeu d'entraînement

250 milliards de mots
30 mots / personne sur Terre

GPT4 (2024)

Taille du cerveau

2000 milliards de paramètres
Cerveau de rat

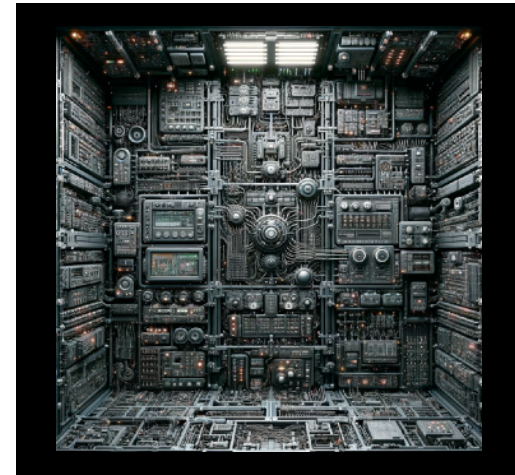


Taille du jeu d'entraînement

10 mille milliards de mots
5 pages / personne sur Terre

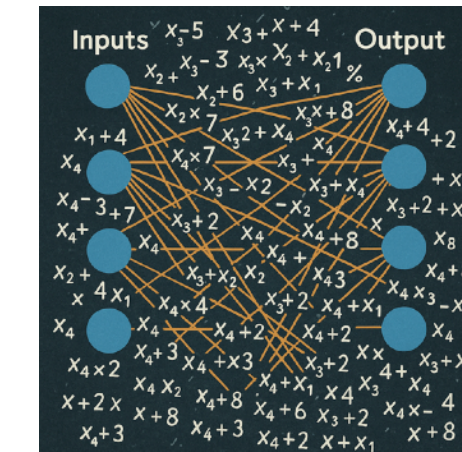
*“We have but one internet... You could even say that data is the **fossil fuel** of AI. It was created somehow, and now we use it, and we've achieved peak data, and there'll be no more.” (Ilya Sutskever, 2024)*

Why don't we understand?



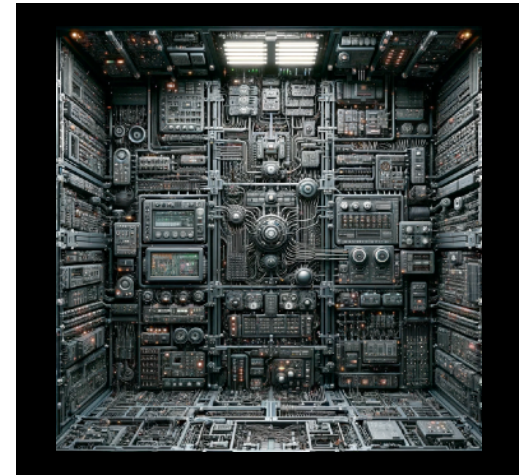
‘Lagopède
Alpin’

$$f_w(x) = W_L \phi(W_{L-1} \phi(\cdots \phi(W_1 x + b_1) \cdots) + b_{L-1}) + b_L.$$



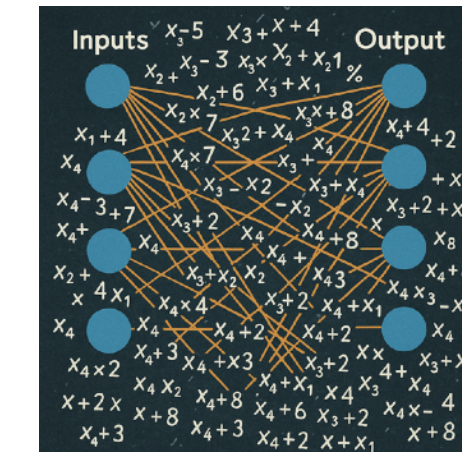
Just because we know exactly what happens at the
‘local scale’ doesn’t mean we understand the behaviour at
the macroscopic level.

Why don't we understand?

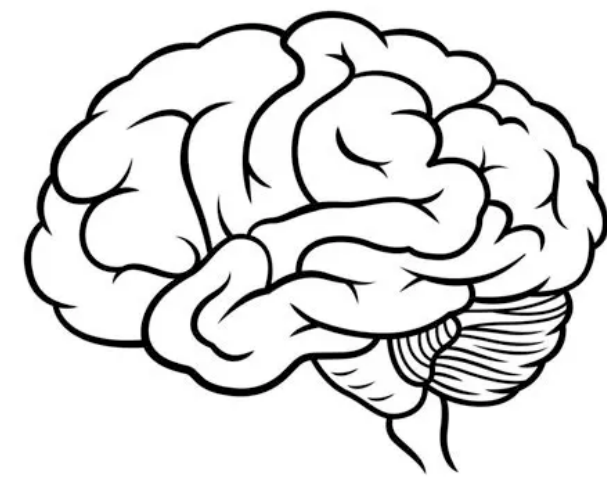


‘Lagopedè
Alpin’

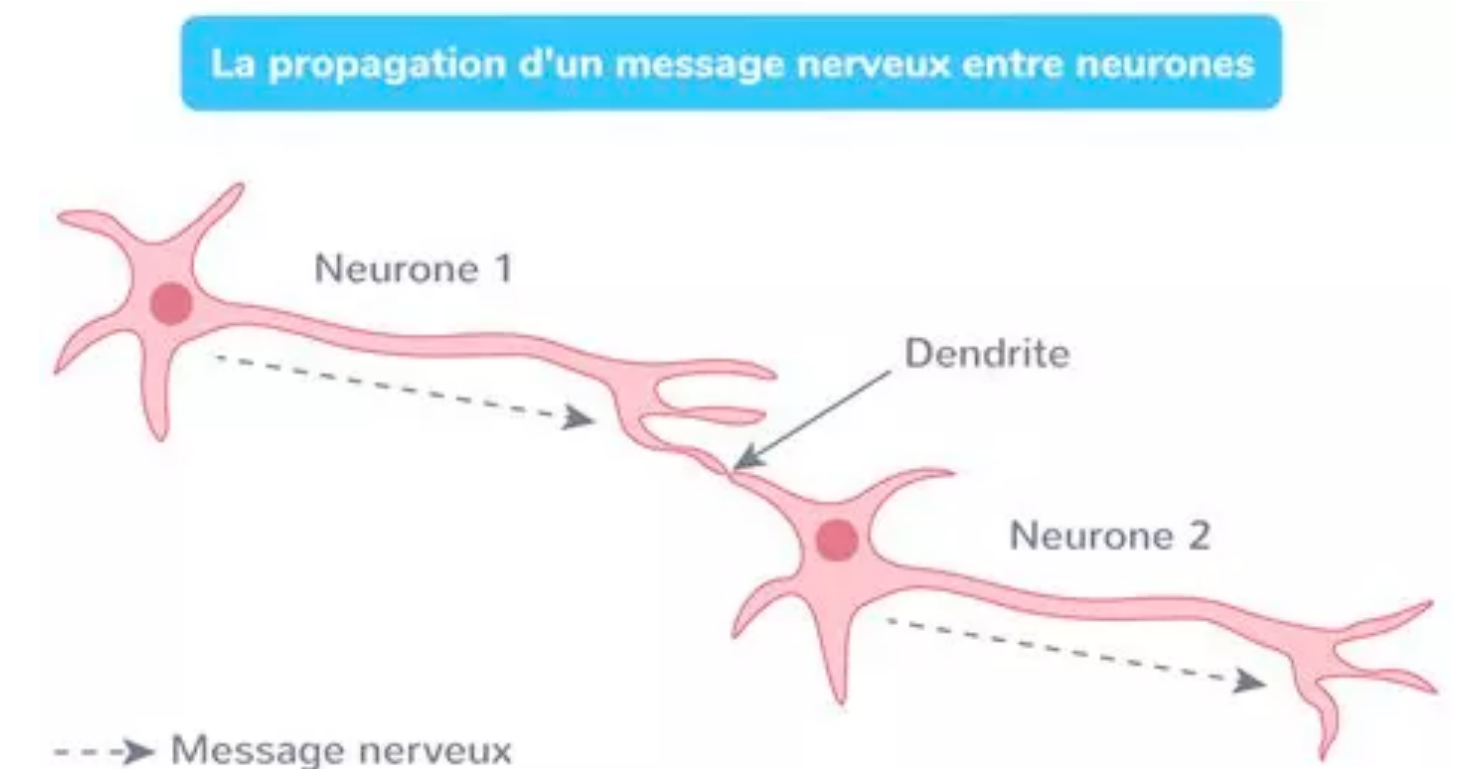
$$f_w(x) = W_L \phi(W_{L-1} \phi(\cdots \phi(W_1 x + b_1) \cdots) + b_{L-1}) + b_L.$$



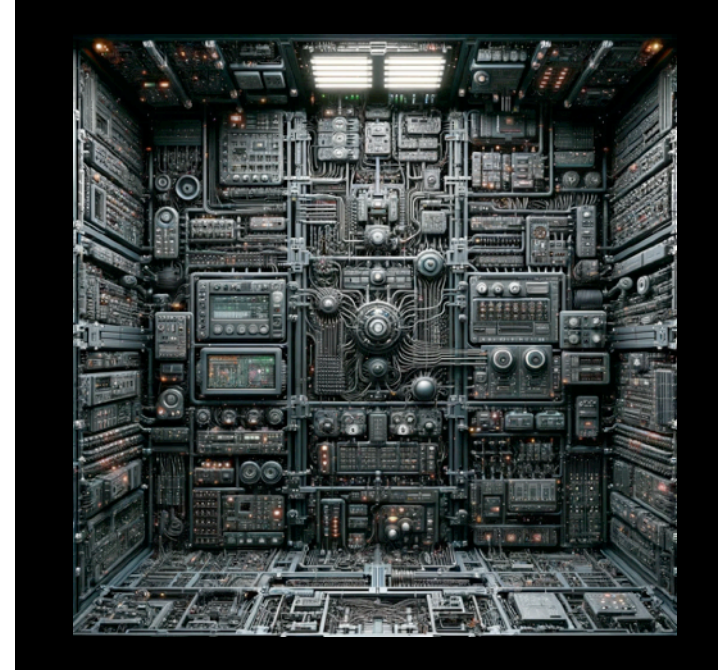
Just because we know exactly what happens at the
‘local scale’ doesn’t mean we understand the behaviour at
the macroscopic level.



‘Lagopedè
Alpin’



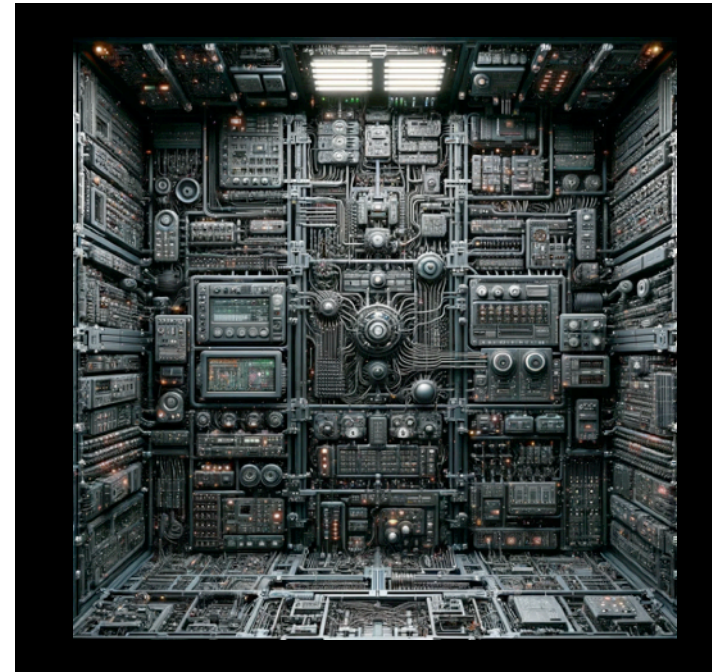
What are the issues?



‘Lagopède
Alpin’

It works brilliantly, but we don't understand why!

What are the issues?



‘Lagopedè
Alpin’

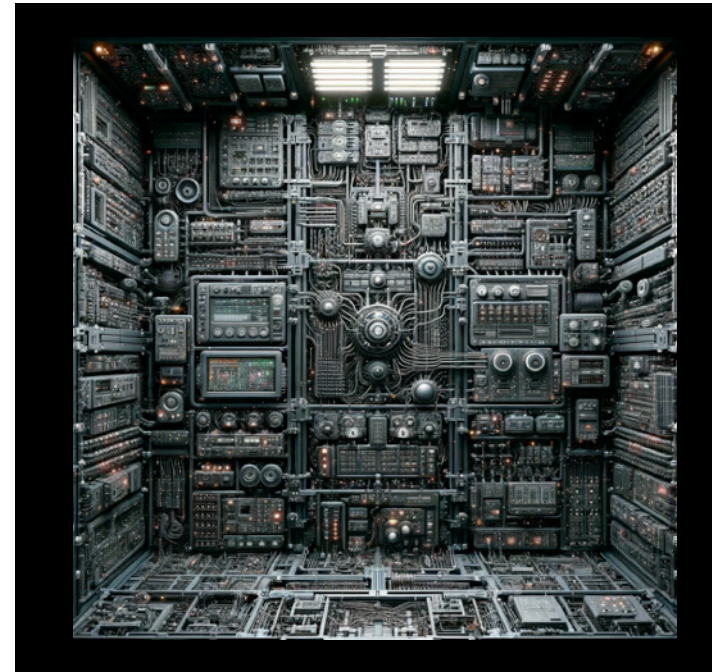
It works brilliantly, but we don't understand why!

How does it make a decision? With which criteria?

When does it make a mistake? Can we fool it?

Is it biased? Can we extract sensible data?

What are the issues?



‘Lagopède
Alpin’

It works brilliantly, but we don't understand why!

How does it make a decision? With which criteria?

When does it make a mistake? Can we fool it?

Is it biased? Can we extract sensible data?

Issues: security, ethics, reliability, sustainability...

The models (currently) do not perceive the world as we do!



“panda”

57.7% confidence

+ .007 ×



noise

=



“gibbon”

99.3% confidence

Recent successes challenge many of our conceptions:
Machines now navigate territories reserved to conscious beings

A large, empty rectangular box with rounded corners and a thin red border, occupying the top third of the slide.A large, empty rectangular box with rounded corners and a thin red border, occupying the middle third of the slide.A large, empty rectangular box with rounded corners and a thin red border, occupying the bottom third of the slide.

Recent successes challenge many of our conceptions:
Machines now navigate territories reserved to conscious beings

Does ChatGPT *understand* things?

Recent successes challenge many of our conceptions:
Machines now navigate territories reserved to conscious beings

Does ChatGPT *understand* things? Is it *intelligent*?

Recent successes challenge many of our conceptions:
Machines now navigate territories reserved to conscious beings

Does ChatGPT *understand* things?

Is it *intelligent*?

Is it *creative*?

Recent successes challenge many of our conceptions:
Machines now navigate territories reserved to conscious beings

Does ChatGPT *understand* things?

Is it *intelligent*?

Is it *creative*?

Could it be *conscious*?

Recent successes challenge many of our conceptions:
Machines now navigate territories reserved to conscious beings

Does ChatGPT *understand* things? Is it *intelligent*?
Is it *creative*? Could it be *conscious*?
How do we even define these words?

Recent successes challenge many of our conceptions:
Machines now navigate territories reserved to conscious beings

Does ChatGPT *understand* things? Is it *intelligent*?
Is it *creative*? Could it be *conscious*?
How do we even define these words?

What about us, humans:

Recent successes challenge many of our conceptions:
Machines now navigate territories reserved to conscious beings

Does ChatGPT *understand* things? Is it *intelligent*?
Is it *creative*? Could it be *conscious*?
How do we even define these words?

What about us, humans:

Is consciousness / intelligence only computational?

Recent successes challenge many of our conceptions:
Machines now navigate territories reserved to conscious beings

Does ChatGPT *understand* things? Is it *intelligent*?
Is it *creative*? Could it be *conscious*?
How do we even define these words?

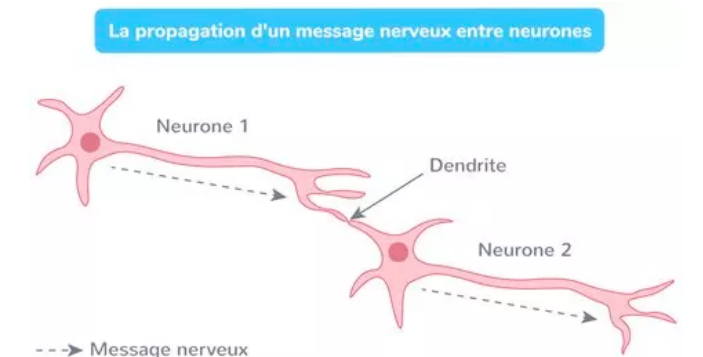
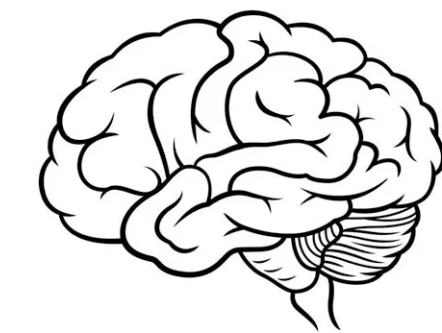
What about us, humans:

Is consciousness / intelligence only computational?
Is creativeness about interpolation or extrapolation?

Recent successes challenge many of our conceptions:
Machines now navigate territories reserved to conscious beings

Does ChatGPT *understand* things? Is it *intelligent*?
Is it *creative*? Could it be *conscious*?
How do we even define these words?

What about us, humans:



Is consciousness / intelligence only computational?
Is creativeness about interpolation or extrapolation?

A better understanding of these seemingly 'intelligent' machines will
shed light on many aspects of our own.

So what can we do?

As we don't properly understand them,
Current neural network can be depicted as::

As we don't properly understand them,
Current neural network can be depicted as::

Chat-GPT:



Hard to follow what is going on

What theoreticians look at:



What theoreticians look at:



It already has similar properties (it flies!), and much simpler to analyse and understand.

“Nothing is more practical than a good theory”

Practical
deep learning



Transformers

$$\text{Atten}(\mathbf{h})(t) = \mathbf{h}(t) + \sum_{i=1}^h \mathbf{W}_v^i \sum_{s=1}^t \sigma[(\mathbf{W}_q^i \mathbf{h}(t))^\top \mathbf{W}_k^i \mathbf{h}(s)] \mathbf{W}_v^i \mathbf{h}(s)$$
$$\text{FF}(\mathbf{h})(t) = f(\mathbf{h})(t).$$

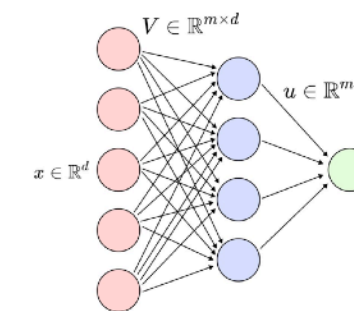
Intuition
through
experiments



Theoretical
deep learning



Simplified
network



Understanding
and
Improvements



Future of
deep learning (??)



Architecture of
the future

Transformers

Deep networks

Depth 2 ReLU
networks

Fully connected
linear networks

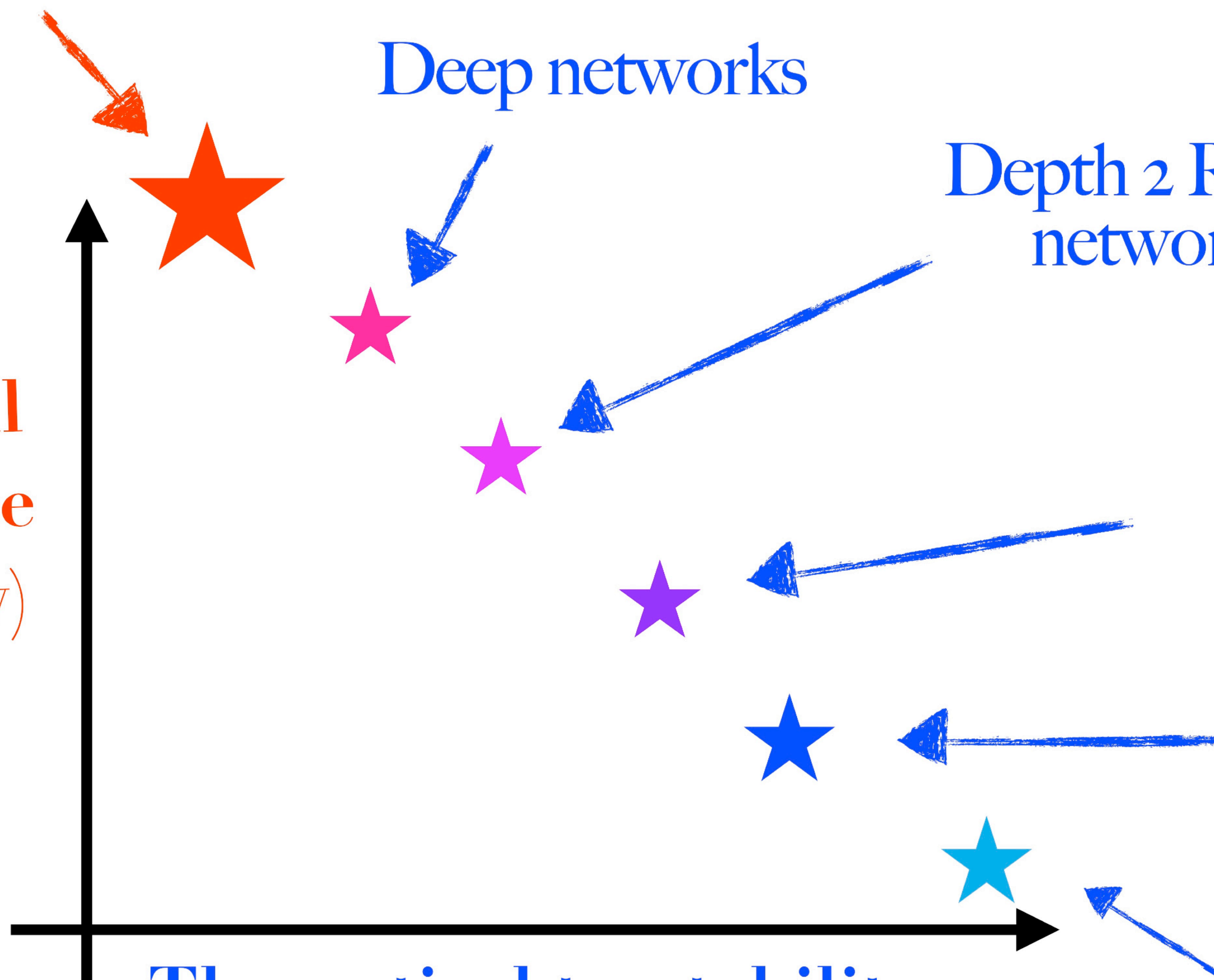
Diagonal
linear networks

Practical
relevance
(for now)

Theoretical tractability
(for now)

$$x \mapsto \langle x, \beta \rangle$$

This course!!!



Up next:
let's do some math!

Online quizz:



<https://scottpesme.github.io/schoen-poll/index.html>

Class webpage

<https://kernel-learning.github.io/>

Grenoble University

Description Schedule Instructors

From Basic Machine Learning models to Advanced
Kernel Learning

