Western Science

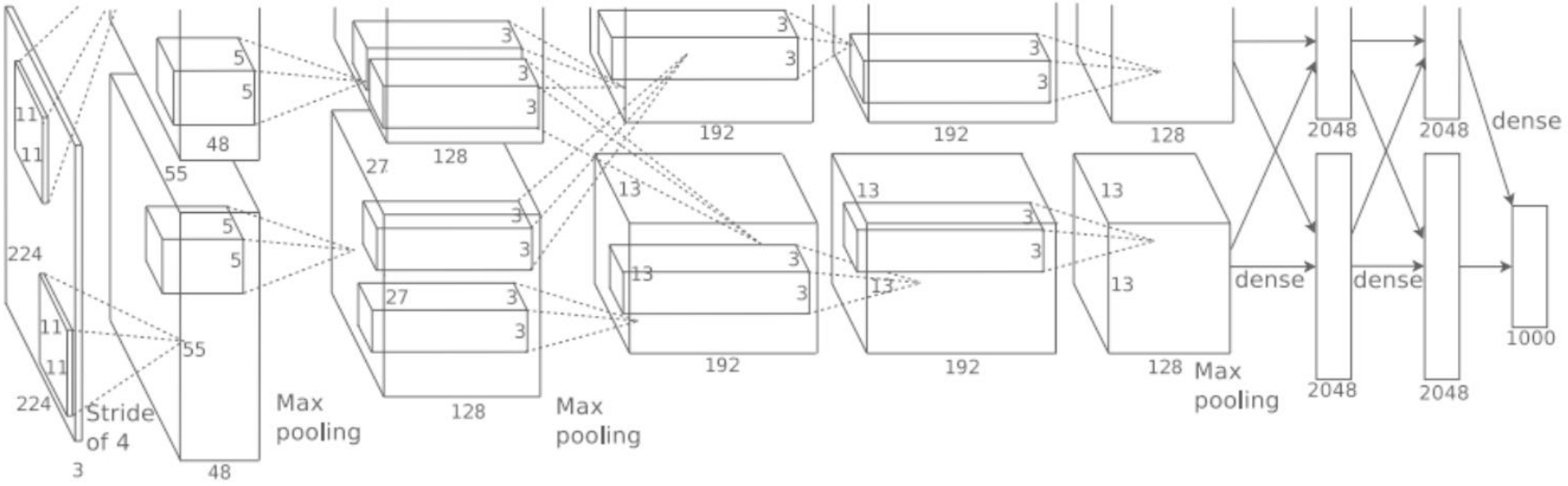
Artificial Intelligence II

Part 2: Lecture 5

Yalda Mohsenzadeh

Winter 2023

Slides are adapted from Antonio Torralba and Philip Isola (MIT)

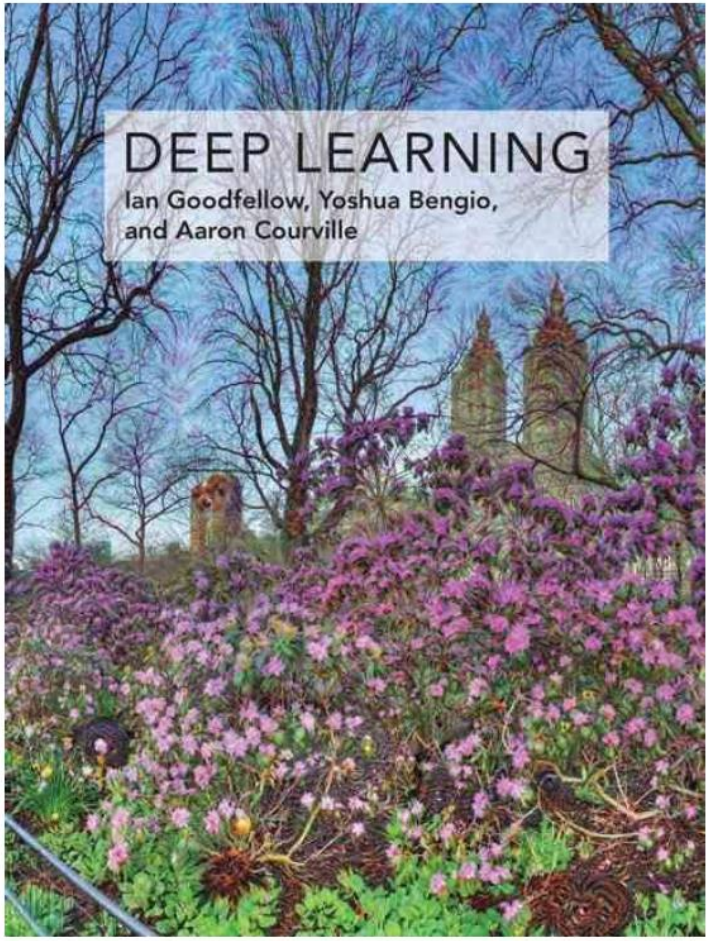


Computer Vision

Neural Networks

Neural Networks

- Brief history
- Basic formulation (hierarchical processing)
- Optimization via gradient descent
- Layer types (Linear, Pointwise, non-linearity)
- Linear classification with perceptron
- Tensor flow
- Regularizers
- Normalization



http://www.deeplearningbook.org/

By Ian Goodfellow, Yoshua Bengio and Aaron Courville

November 2016

Deep Learning

- Modeling the visual world is incredibly complicated. We need high capacity models.

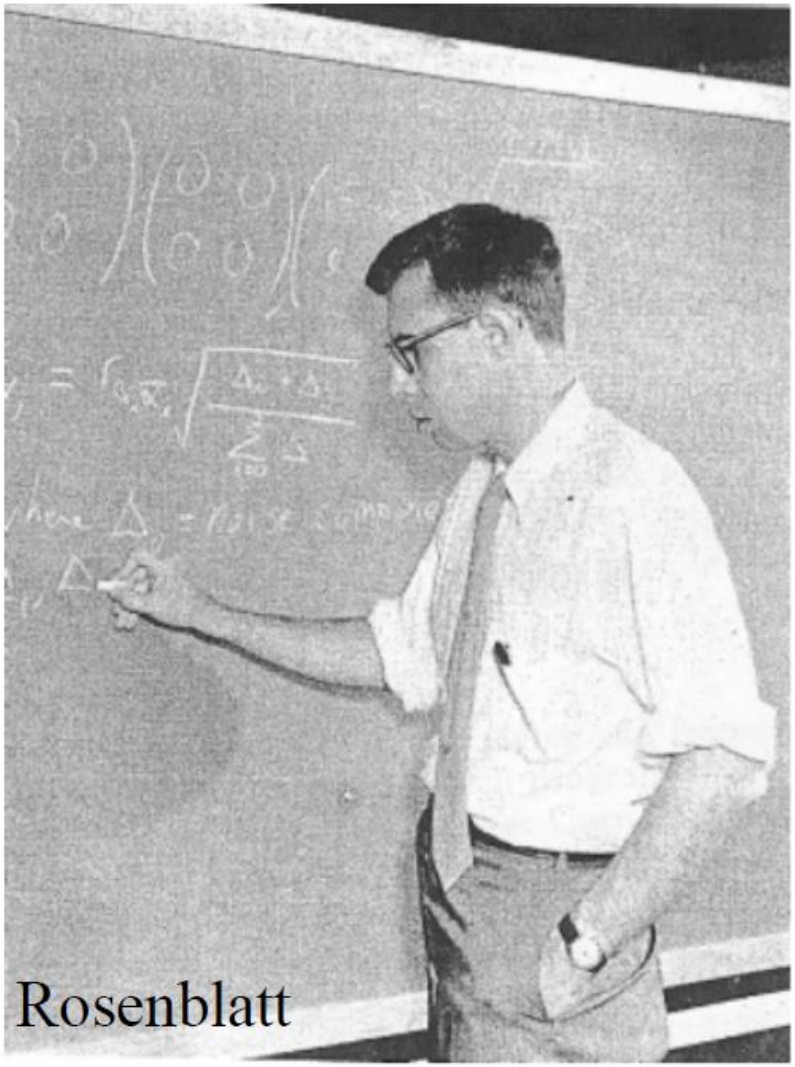
- In the past, we didn't have enough data to fit these models. But now we do!

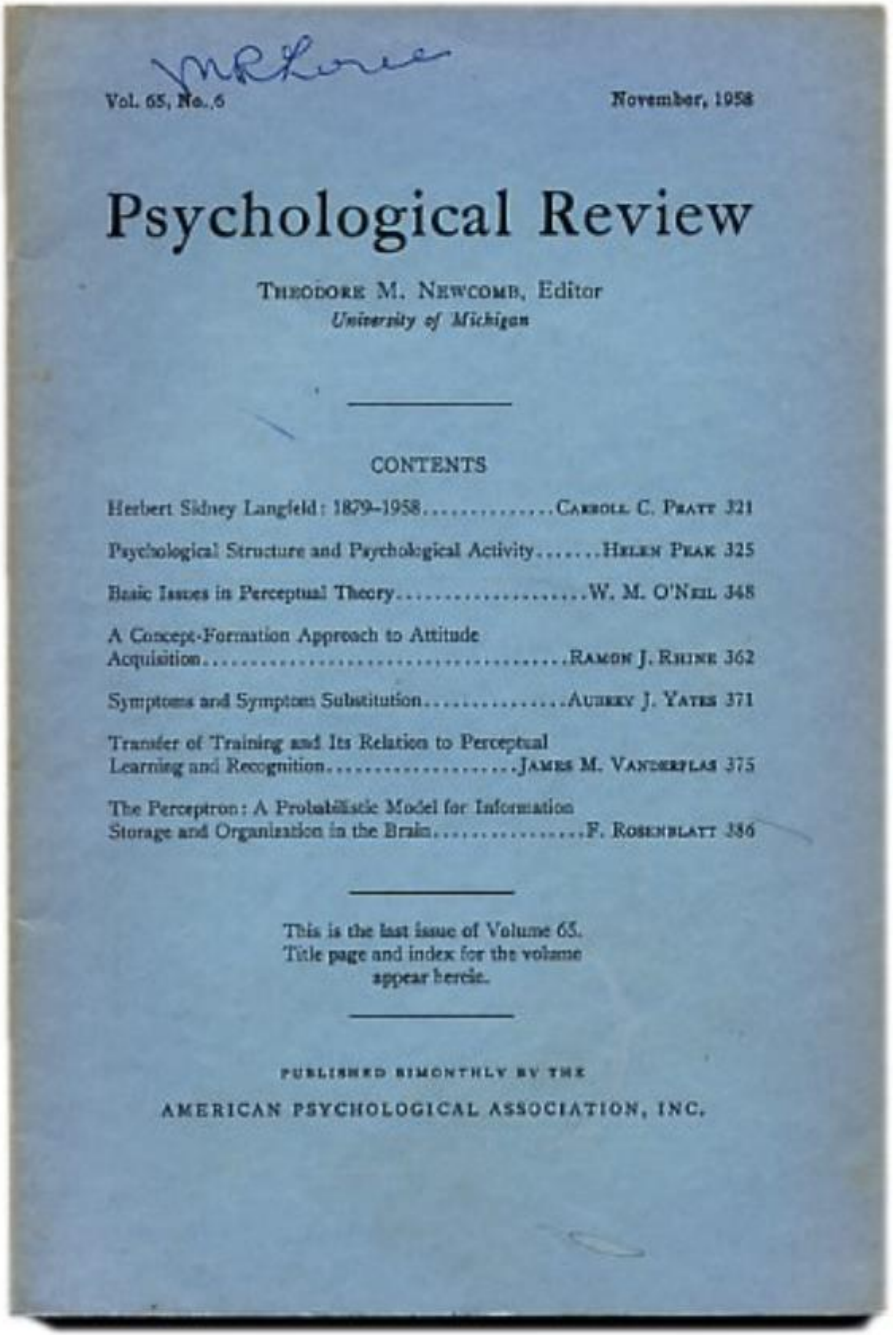
- We want a class of **high capacity models** that are **easy to optimize**.

- **Deep Neural Networks**!

A brief history of neural networks

Perceptrons, 1958





Vol. 65, No. 6 November, 1958

Psychological Review

THEODORE M. NEWCOMB, Editor
University of Michigan

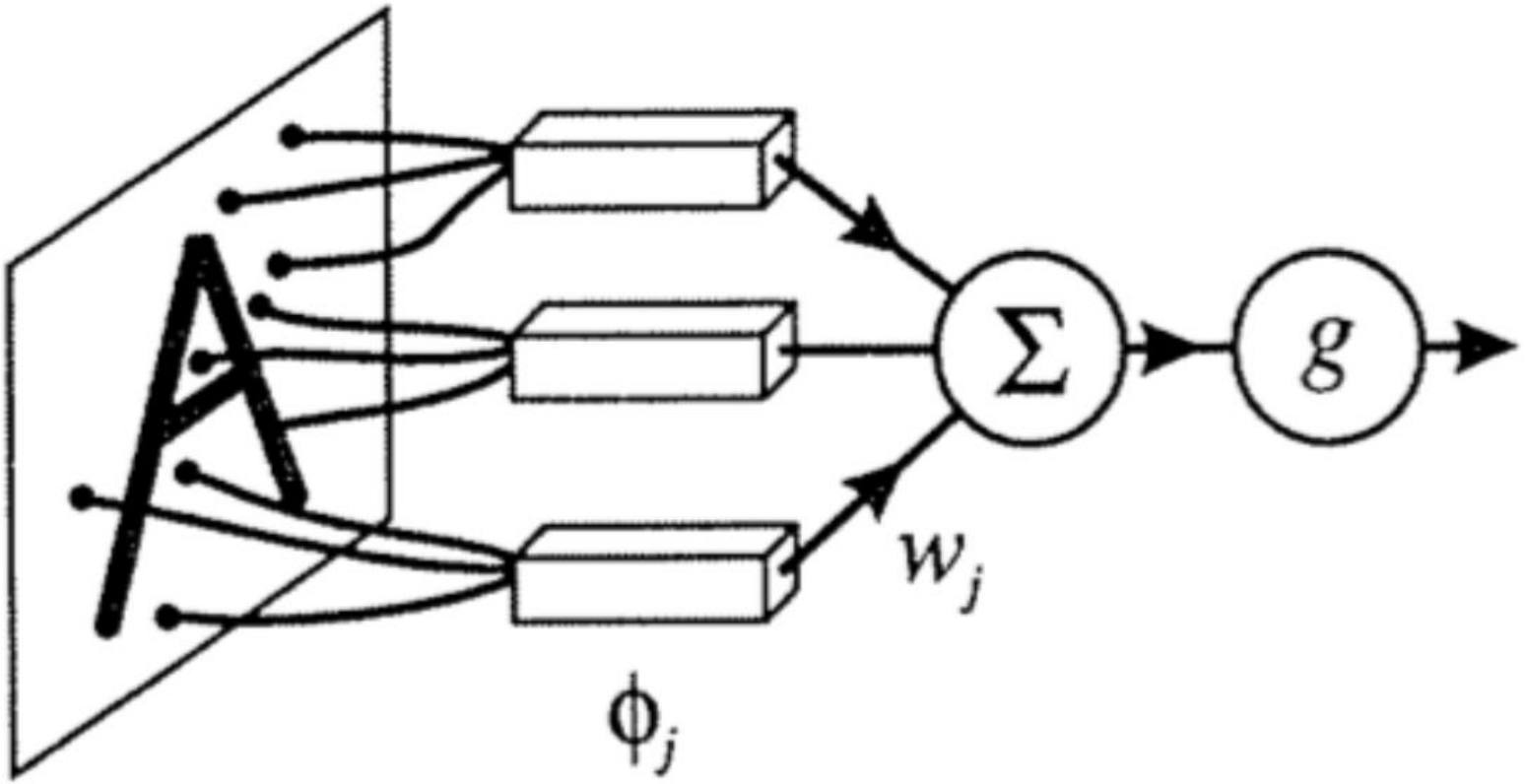
CONTENTS

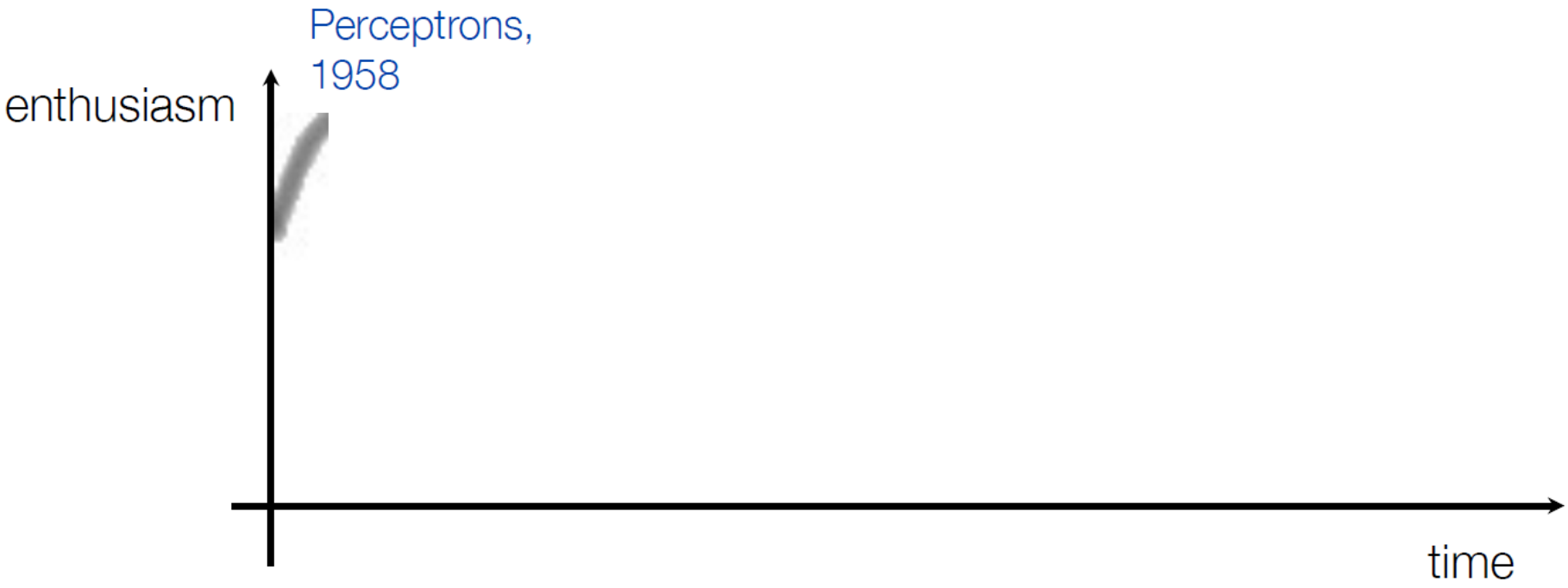
This is the last issue of Volume 65.
Title page and index for the volume
appear herein.

PUBLISHED BIMONTHLY BY THE
AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.

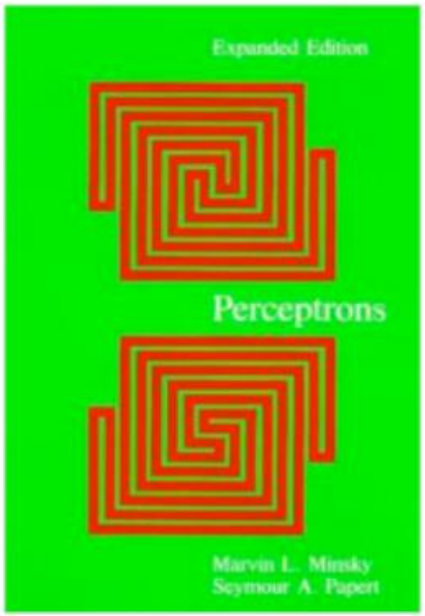
http://www.ecse.rpi.edu/homepages/nagy/PDF_chrono/2011_Nagy_Pace_FR.pdf. Photo by George Nagy

Perceptron, 1958




Perceptrons,
1958
enthusiasm
time

Minsky and Papert, Perceptrons, 1972



Perceptrons, expanded edition

An Introduction to Computational Geometry

By Marvin Minsky and Seymour A. Papert

Overview

Perceptrons - the first systematic study of parallelism in computation - has remained a classical work on threshold automata networks for nearly two decades. It marked a historical turn in artificial intelligence, and it is required reading for anyone who wants to understand the connectionist counterrevolution that is going on today.

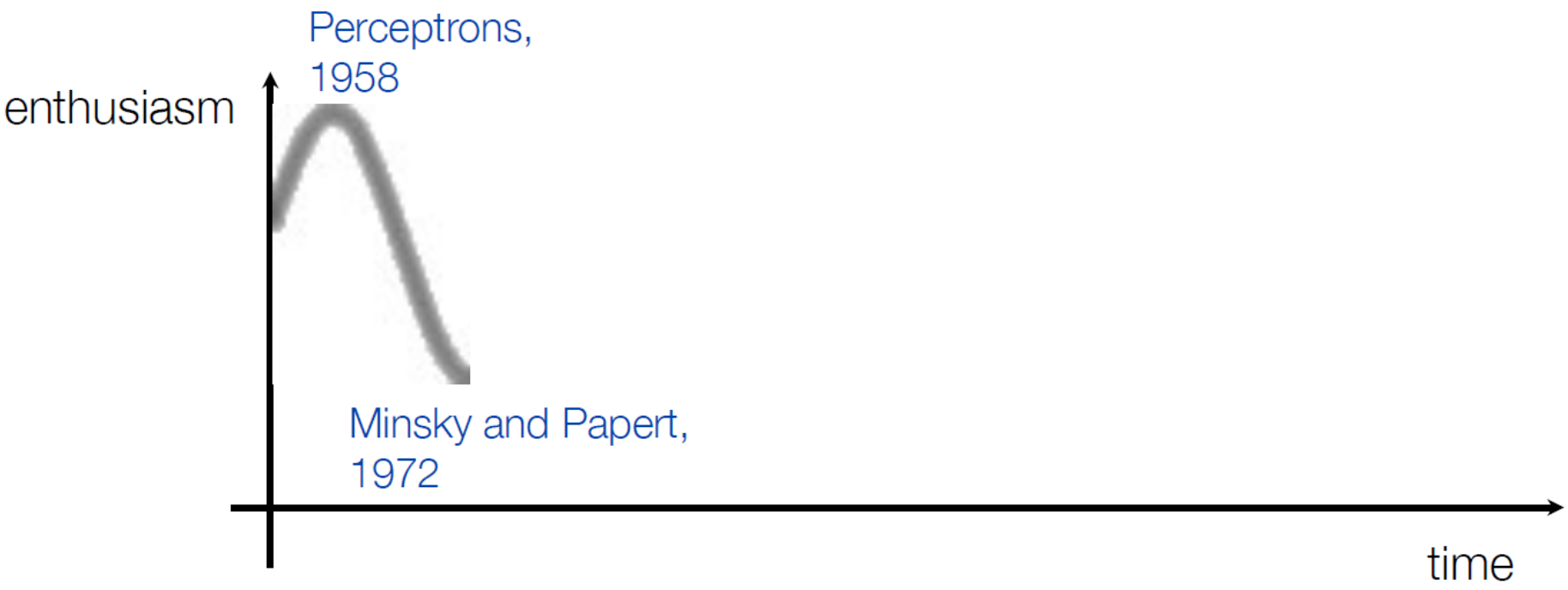
Artificial-intelligence research, which for a time concentrated on the programming of ton Neumann computers, is swinging back to the idea that intelligence might emerge from the activity of networks of neuronlike entities. Minsky and Papert's book was the first example of a mathematical analysis carried far enough to show the exact limitations of a class of computing machines that could seriously be considered as models of the brain. Now the new developments in mathematical tools, the recent interest of physicists in the theory of disordered matter, the new insights into and psychological models of how the brain works, and the evolution of fast computers that can simulate networks of automata have given *Perceptrons* new importance.

Witnessing the swing of the intellectual pendulum, Minsky and Papert have added a new chapter in which they discuss the current state of parallel computers, review developments since the appearance of the 1972 edition, and identify new research directions related to connectionism. They note a central theoretical challenge facing connectionism: the challenge to reach a deeper understanding of how "objects" or "agents" with individuality can emerge in a network. Progress in this area would link connectionism with what the authors have called "society theories of mind."

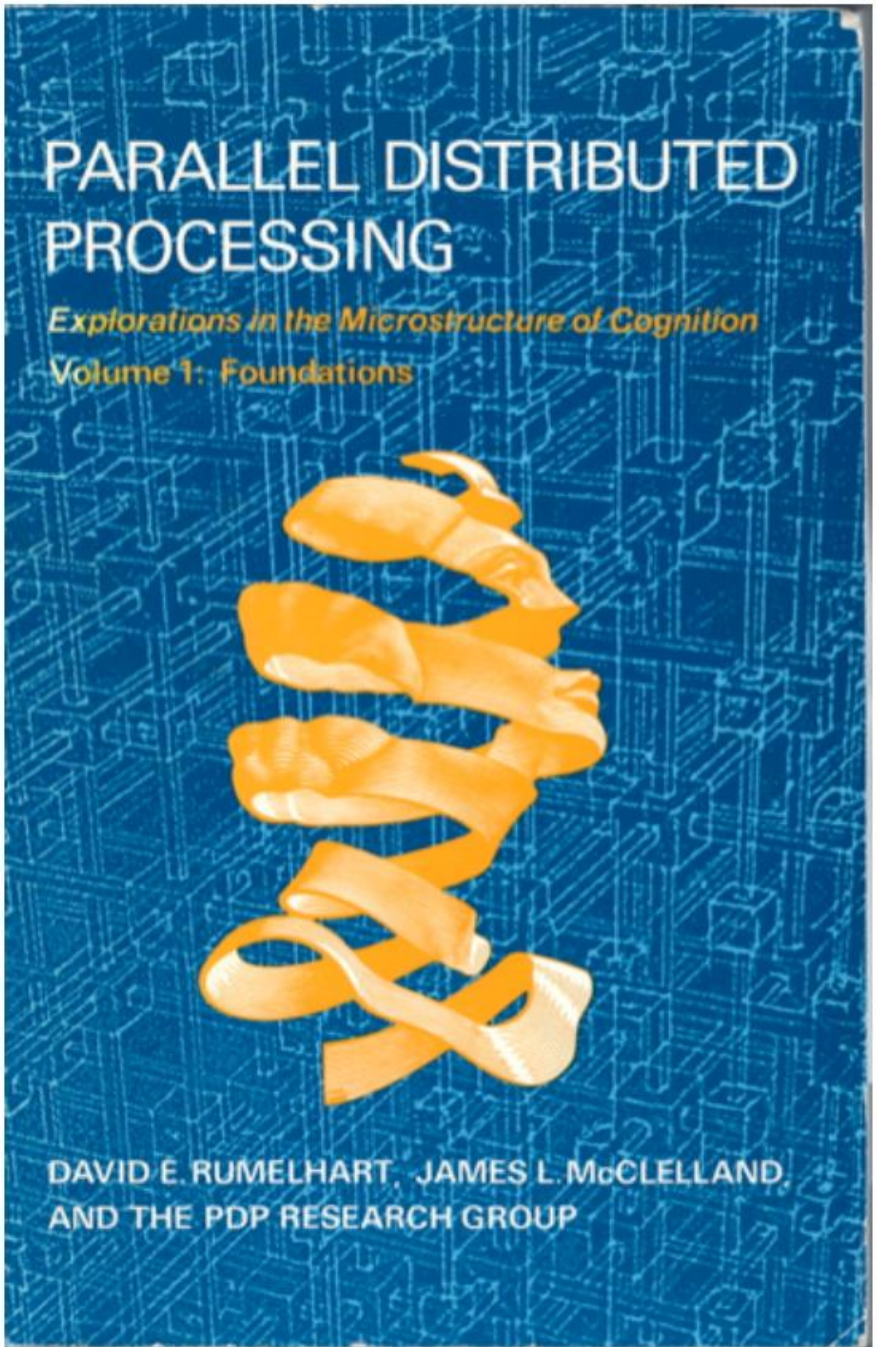
FOR BUYING OPTIONS, START HERE

Select Shipping Destination

Paperback | $35.00 Short | £24.95 | ISBN: 9780262631112 | 308 pp. | 6 x 8.9 in | December 1987

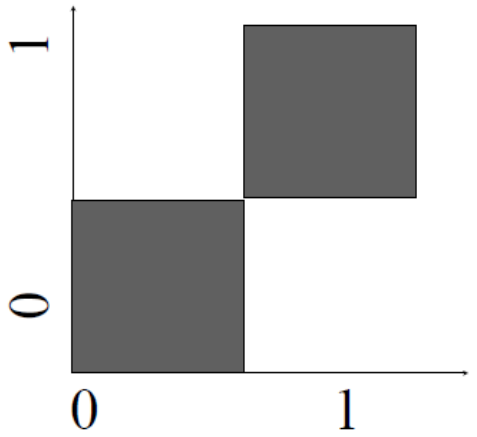

Perceptrons,
1958
enthusiasm
Minsky and Papert,
1972
time

Parallel Distributed Processing (PDP), 1986

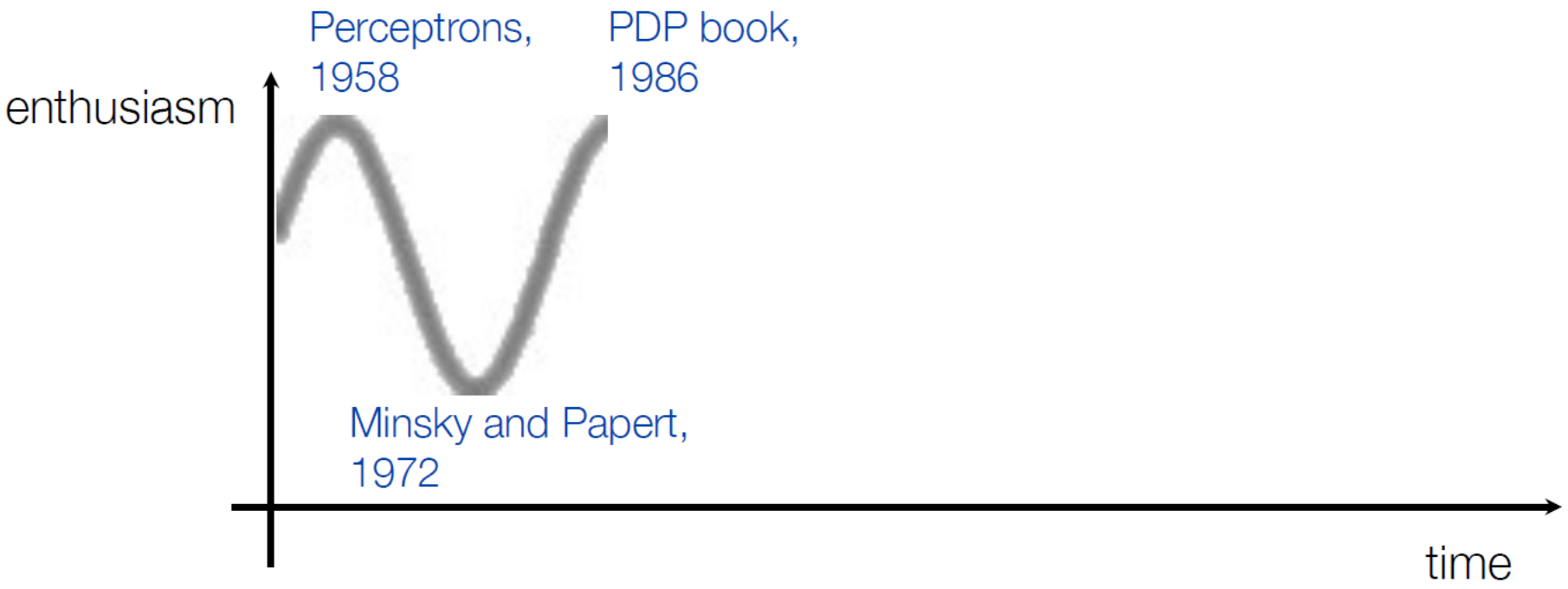


XOR problem

Inputs		Output
0	0	0
1	0	1
0	1	1
1	1	0



PDP authors pointed to the backpropagation algorithm as a breakthrough, allowing multi-layer neural networks to be trained. Among the functions that a multi-layer network can represent but a single-layer network cannot: the XOR function.


Perceptrons,
1958
PDP book,
1986
enthusiasm
Minsky and Papert,
1972
time

LeCun conv nets, 1998

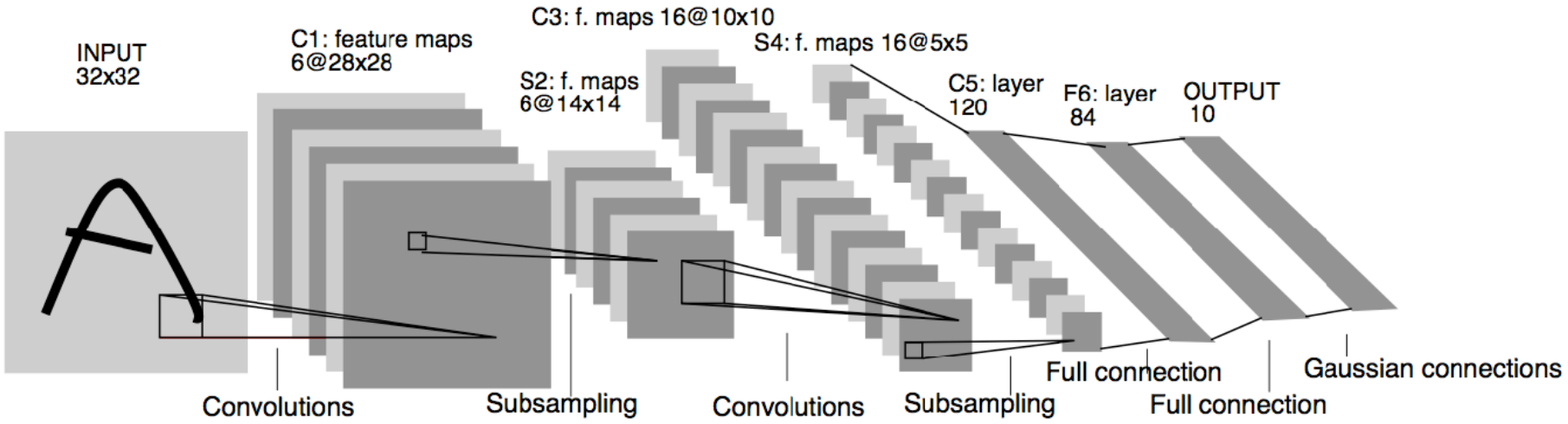


Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

Demos:
http://yann.lecun.com/exdb/lenet/index.html

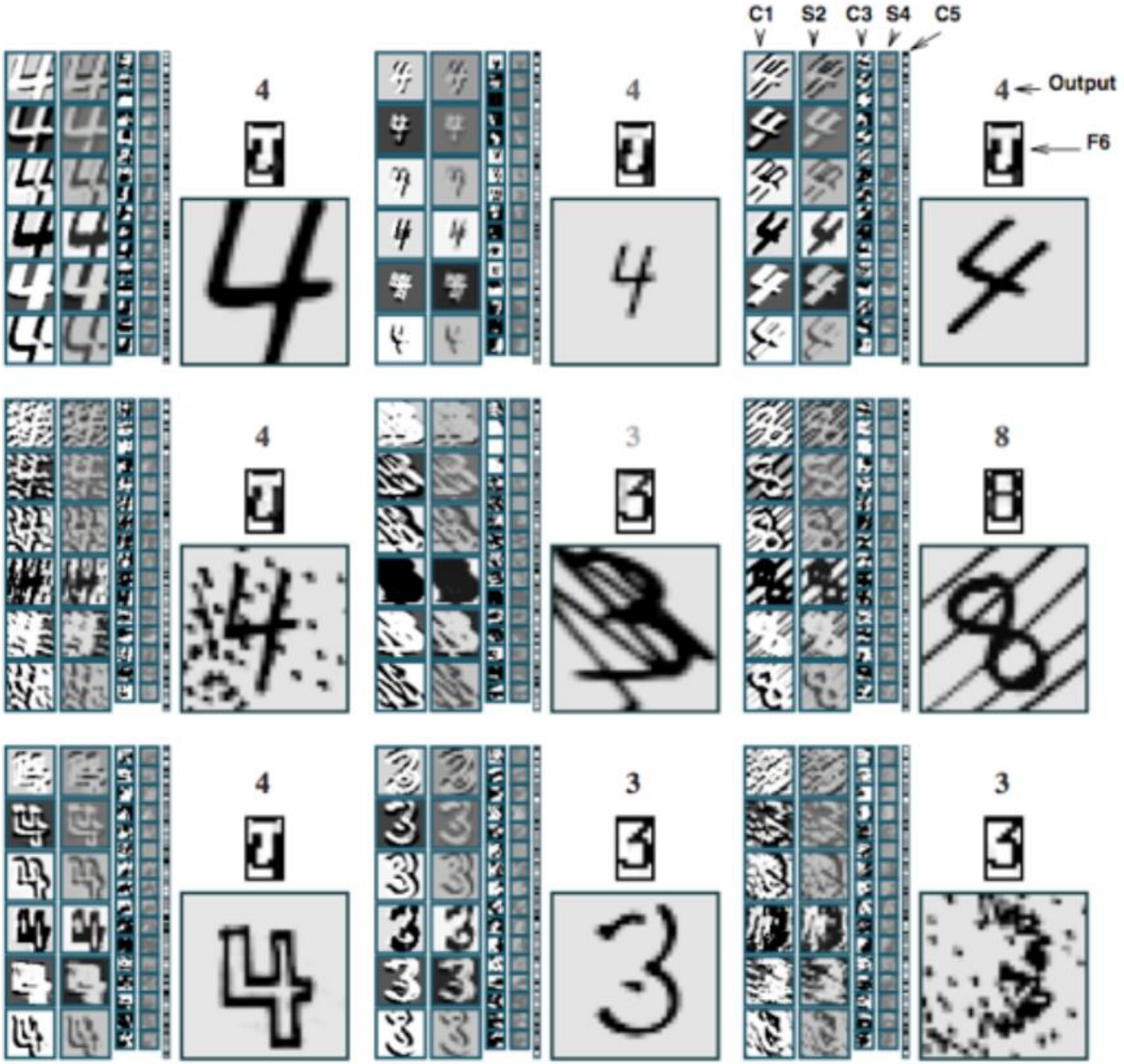
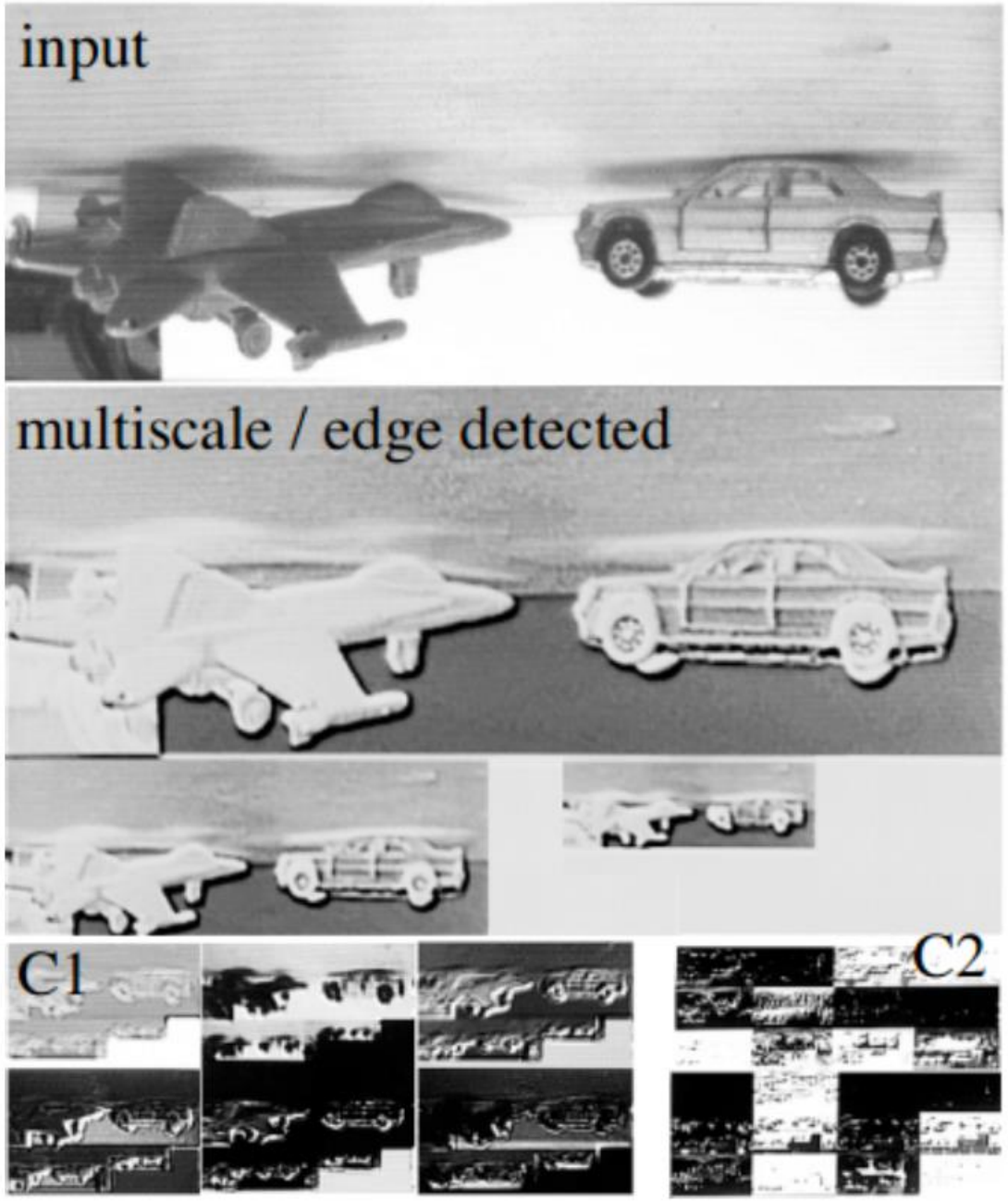


Fig. 13. Examples of unusual, distorted, and noisy characters correctly recognized by LeNet-5. The grey-level of the output label represents the penalty (lighter for higher penalties).



Neural networks to recognize handwritten digits?

Yes

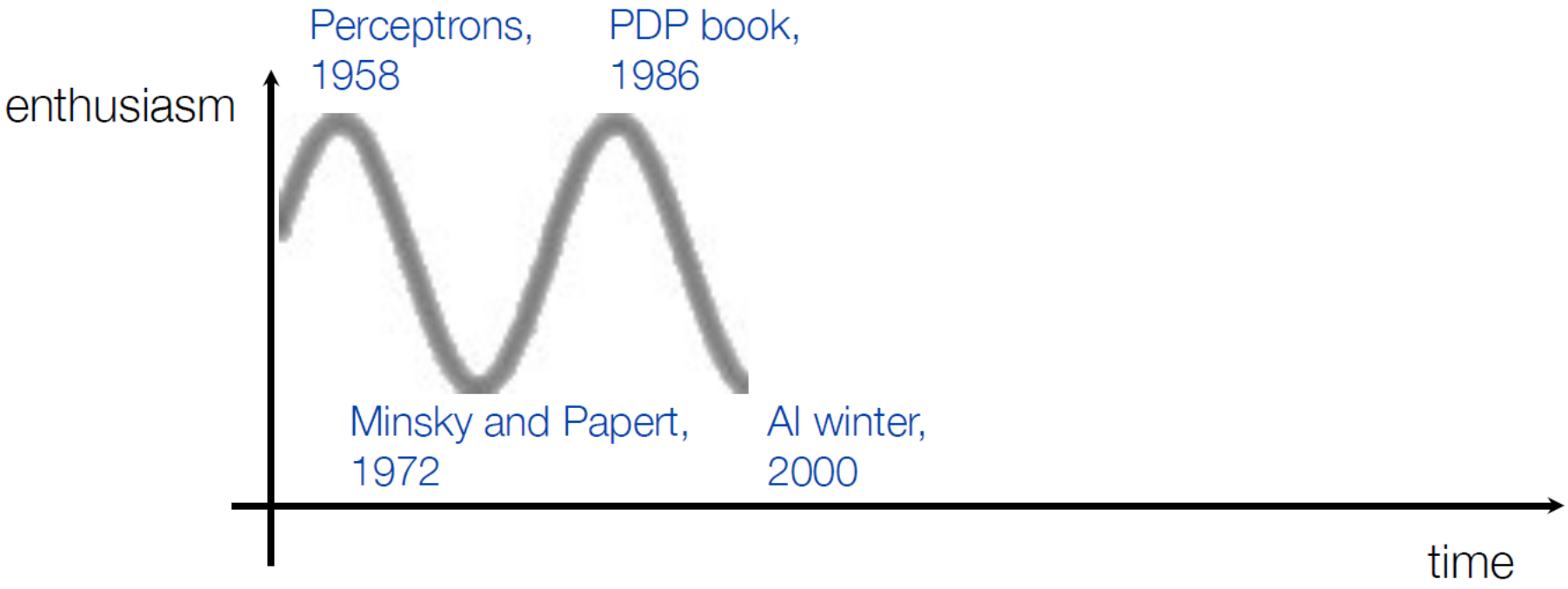
Neural networks for tougher problems?

Not really

http://pub.clement.farabet.net/ecvw09.pdf

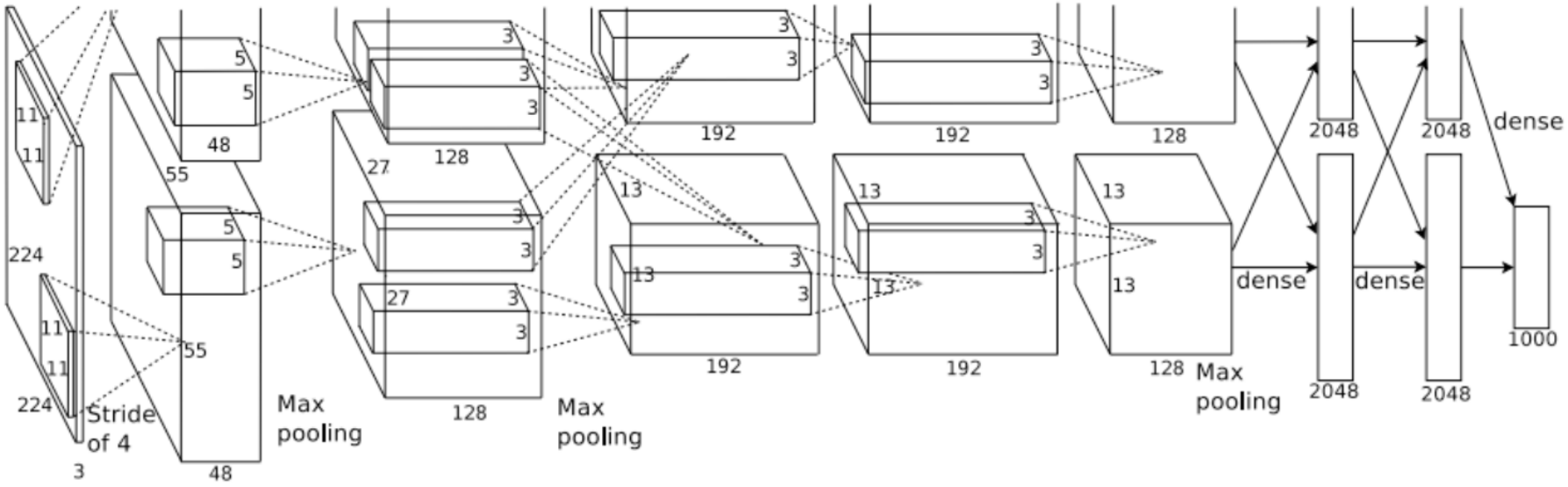
Neural Information Processing Systems 2000

- Neural Information Processing Systems, is the premier conference on machine learning. Evolved from an interdisciplinary conference to a machine learning conference.
- For the 2000 conference:
  - title words predictive of paper acceptance: “Belief Propagation” and “Gaussian”.
  - title words predictive of paper rejection: “Neural” and “Network”.


Perceptrons,
1958
PDP book,
1986
enthusiasm
Minsky and Papert,
1972
AI winter,
2000
time

Krizhevsky, Sutskever, and Hinton, NeurIPS 2012

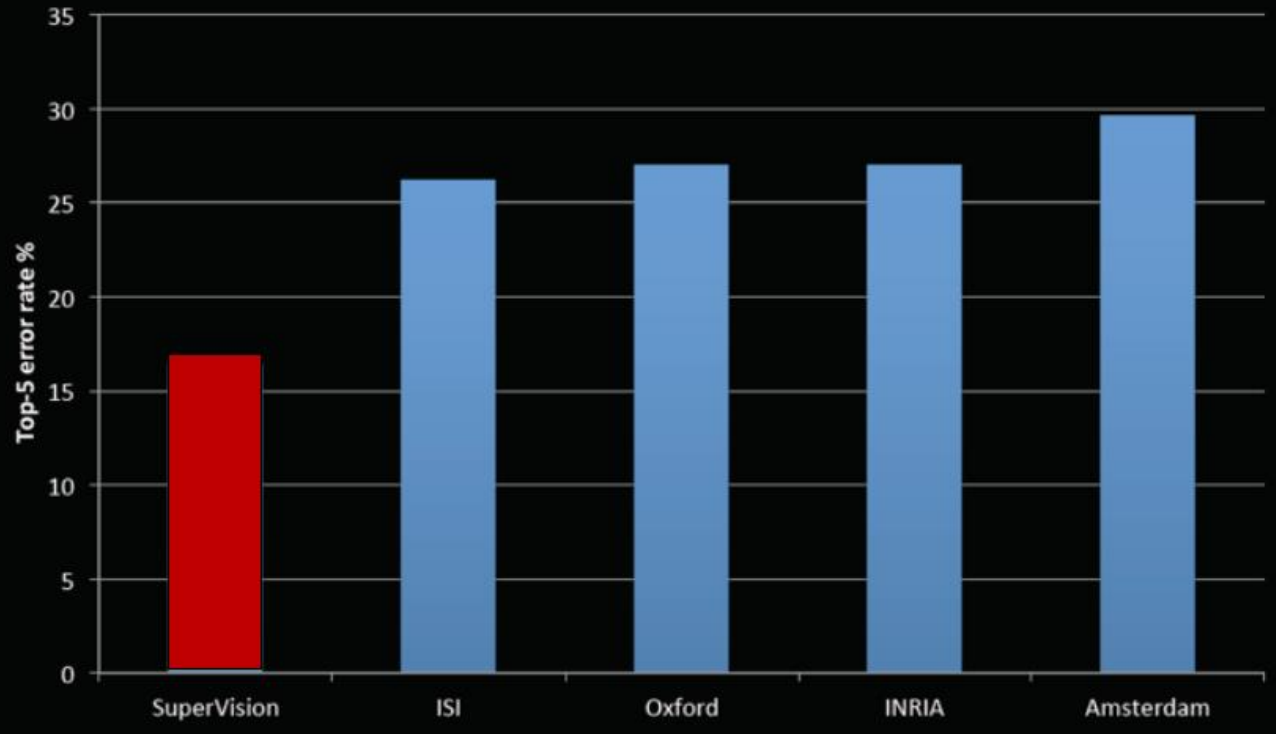
“Alexnet”



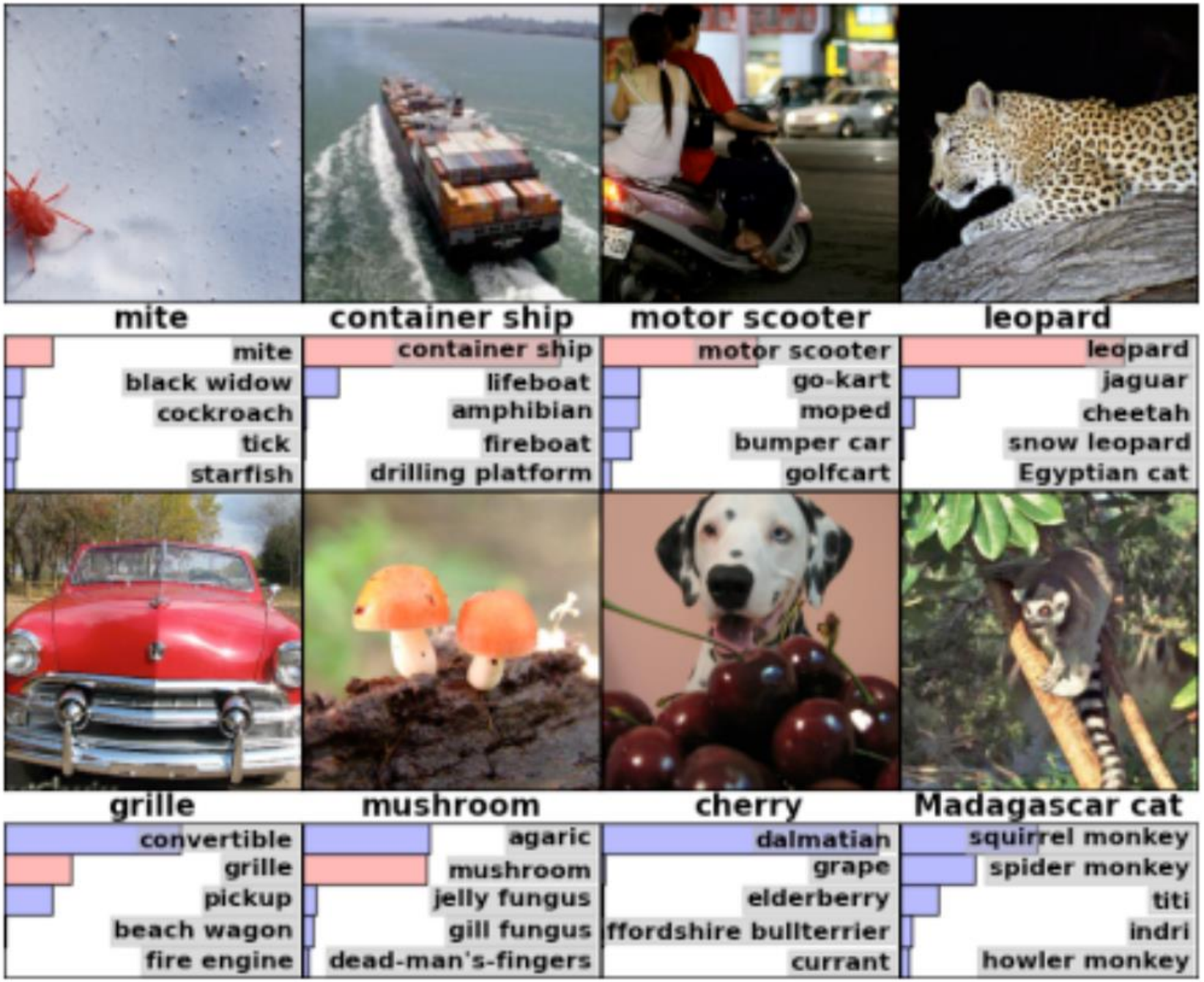
Slide from Rob Fergus, NYU

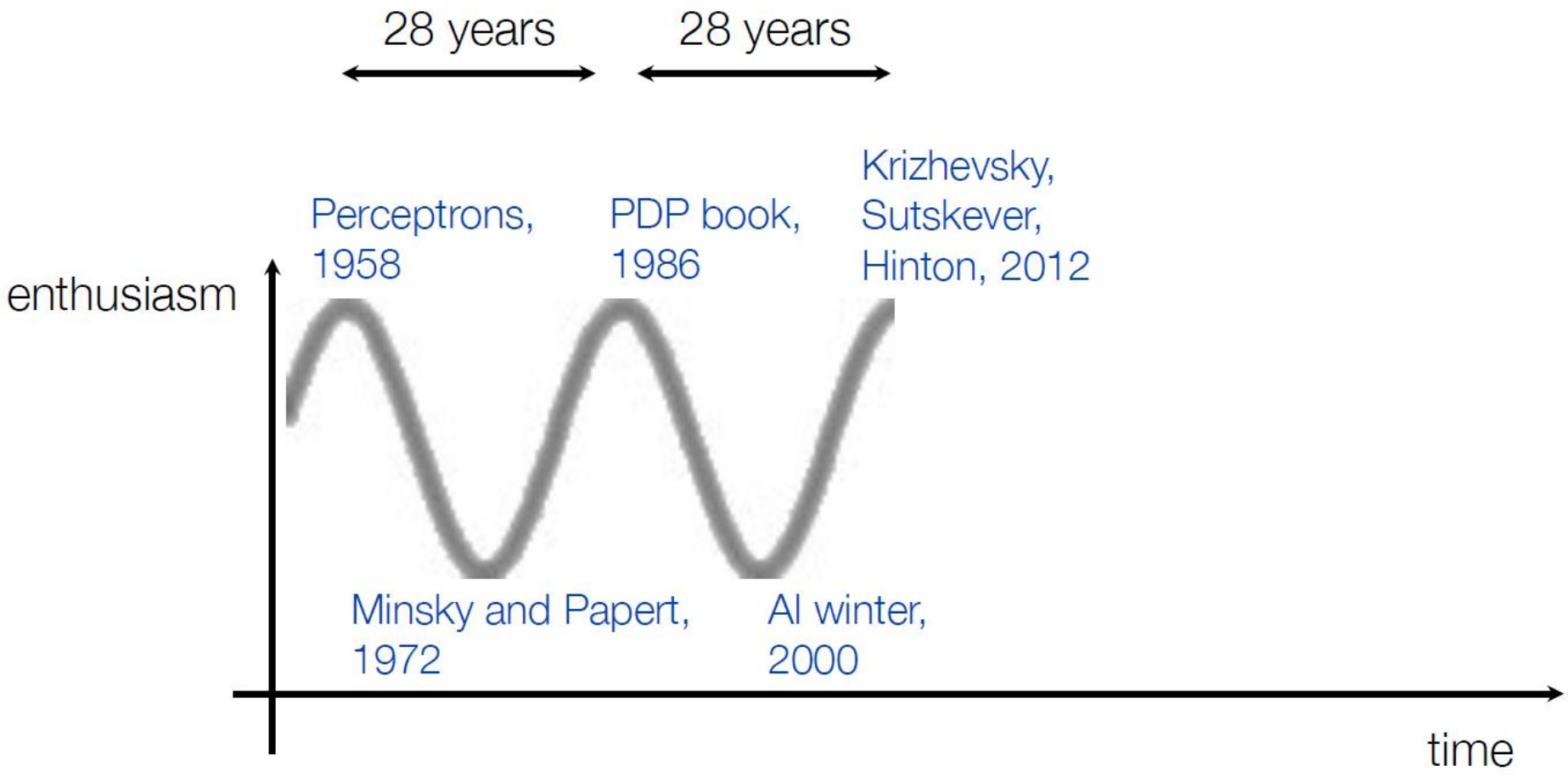
ImageNet Classification 2012

- Krizhevsky et al. -- 16.4% error (top-5)
- Next best (non-convnet) – 26.2% error

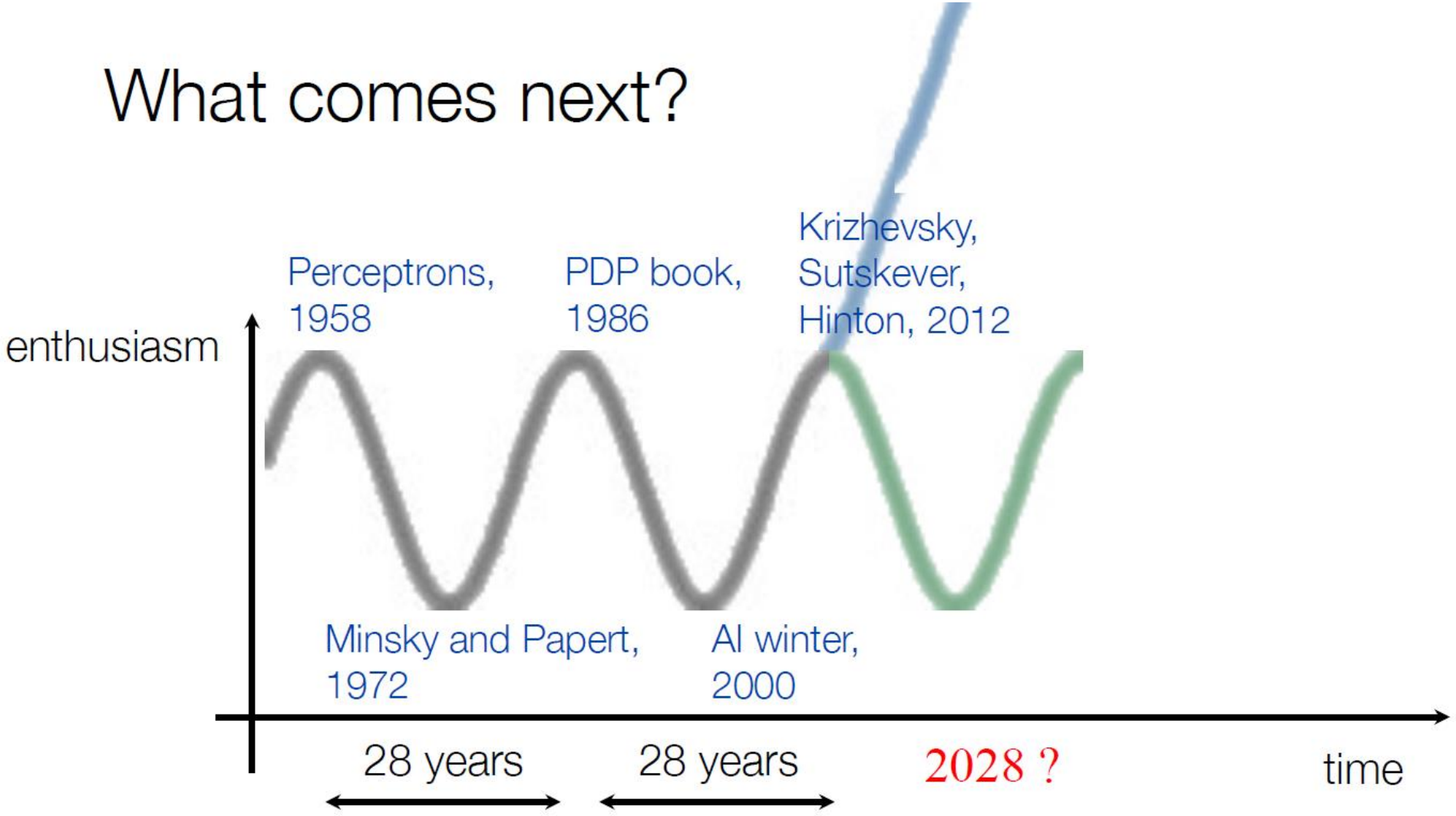


Krizhevsky, Sutskever, and Hinton, NeurIPS 2012

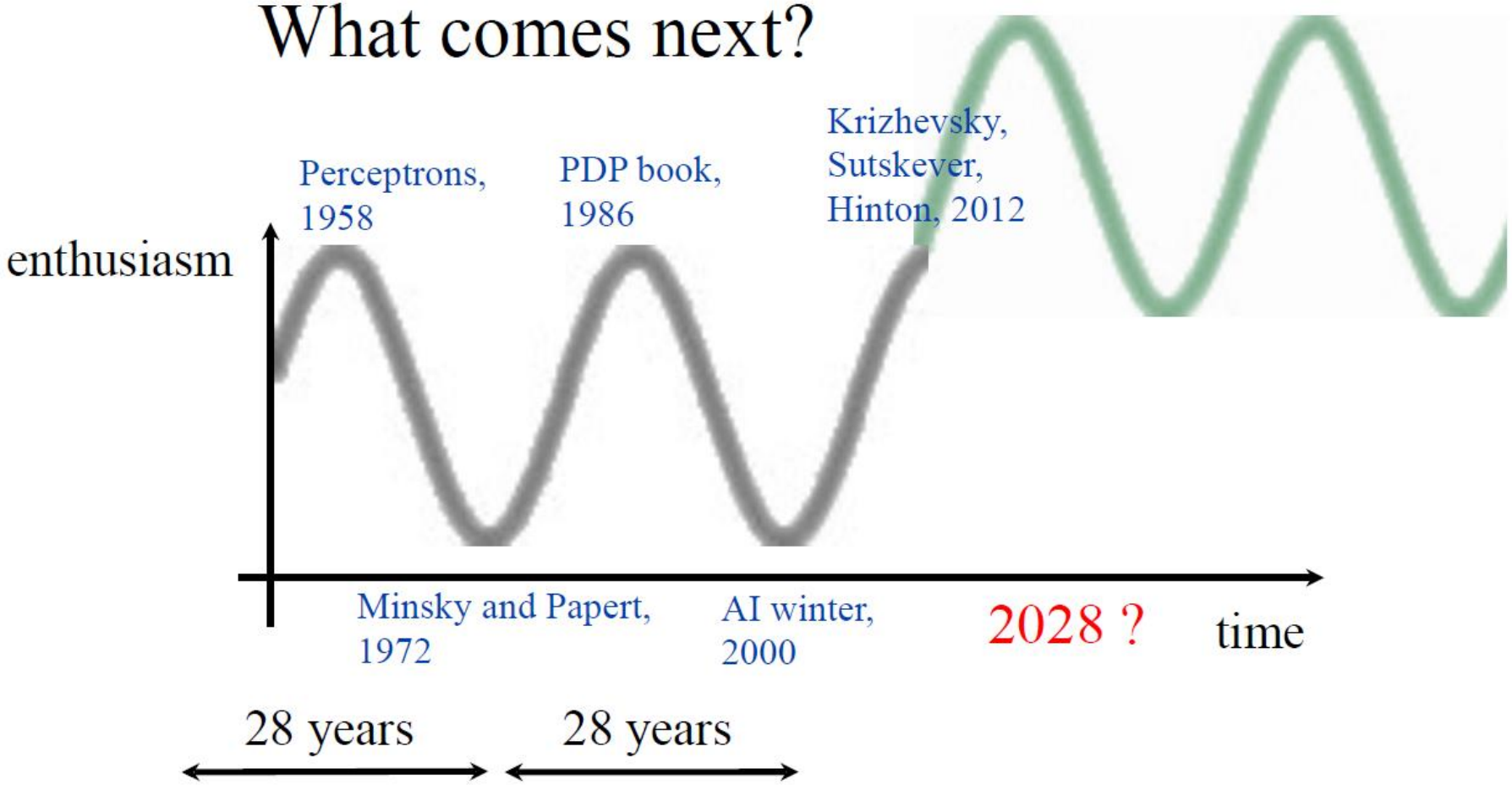



28 years
28 years
Perceptrons, 1958
PDP book, 1986
Krizhevsky, Sutskever, Hinton, 2012
enthusiasm
Minsky and Papert, 1972
AI winter, 2000
time

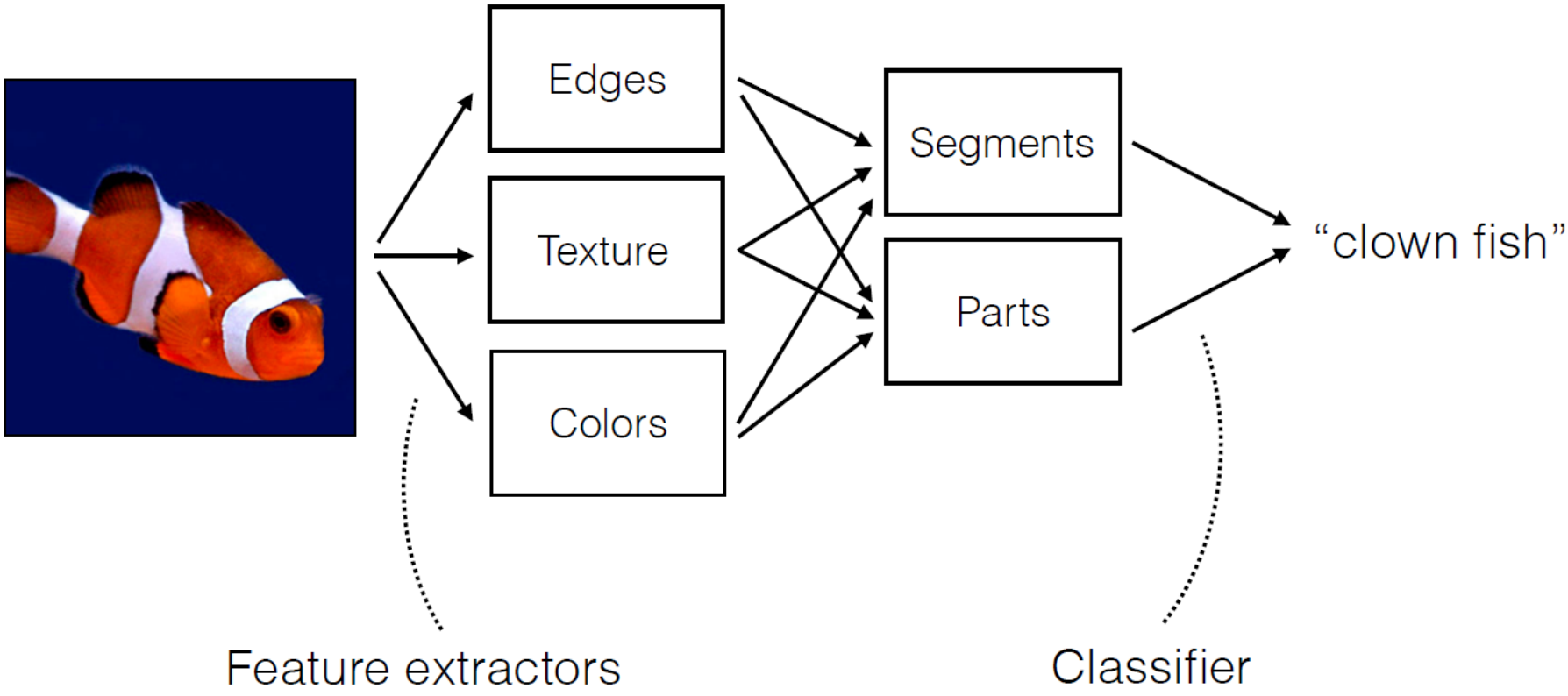
What comes next?



What comes next?



Object Recognition



Object Recognition

Edges

Texture

Colors

Segments

Parts

Learned

"clown fish"

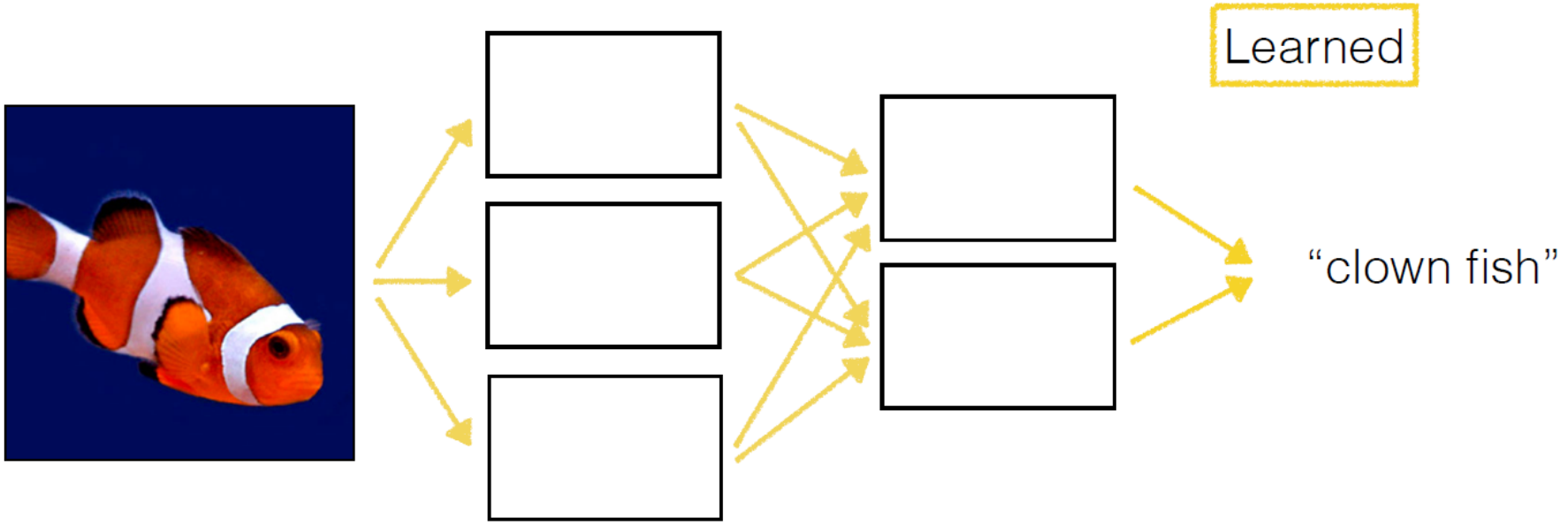
$\phi_k(x)$

Feature extractors

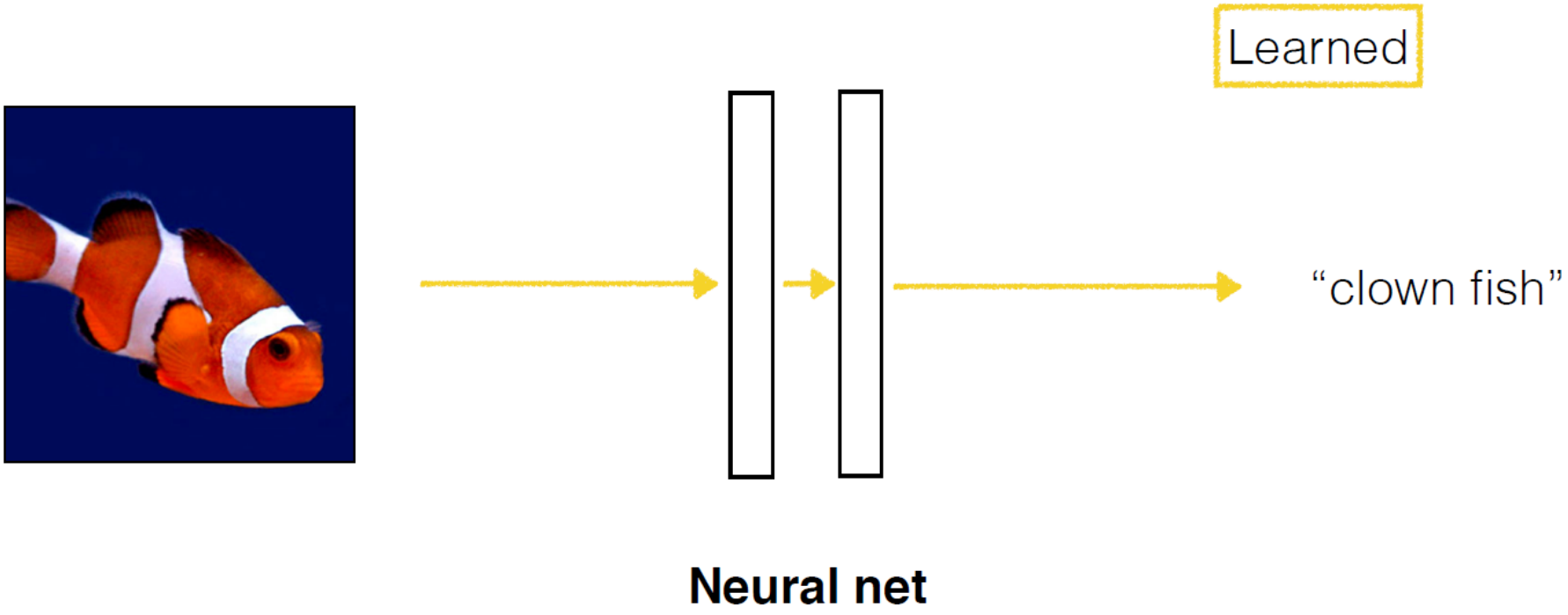
$$f_\theta(x) = \sum_{k=1}^{K} \theta_k \phi_k(x)$$

Classifier

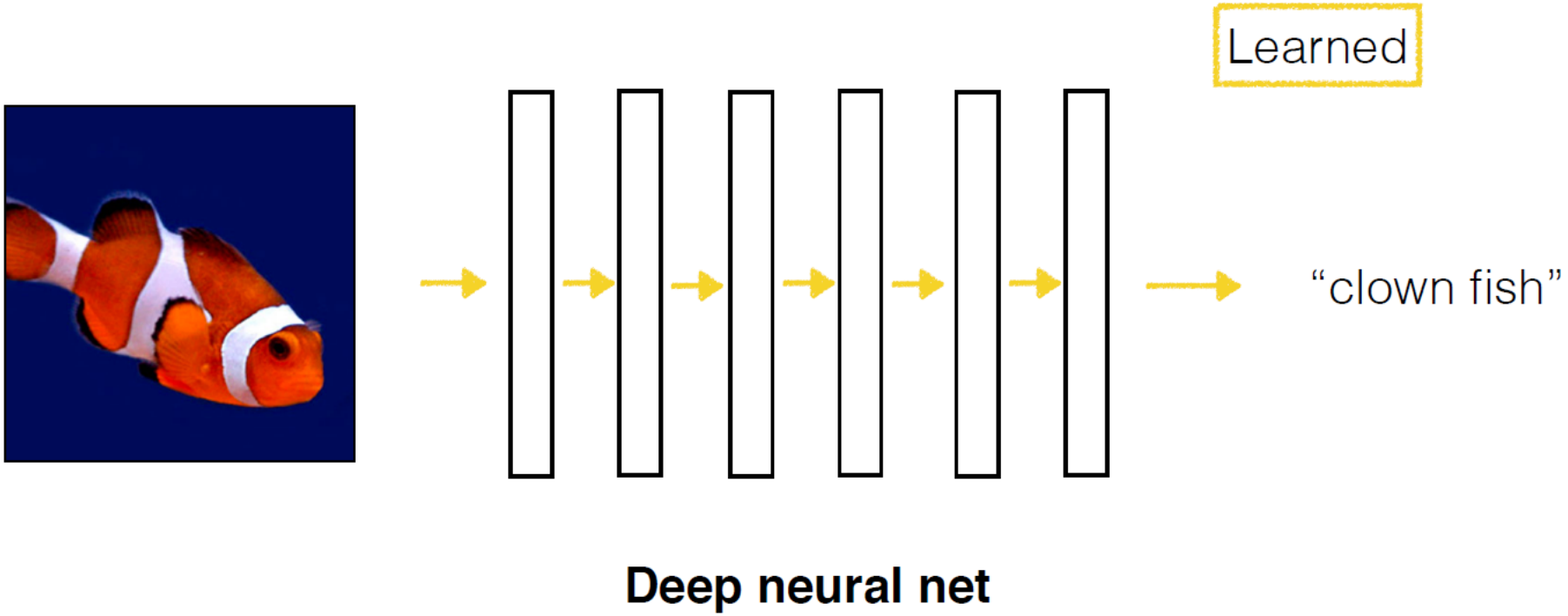
Object Recognition



Object Recognition



Object Recognition

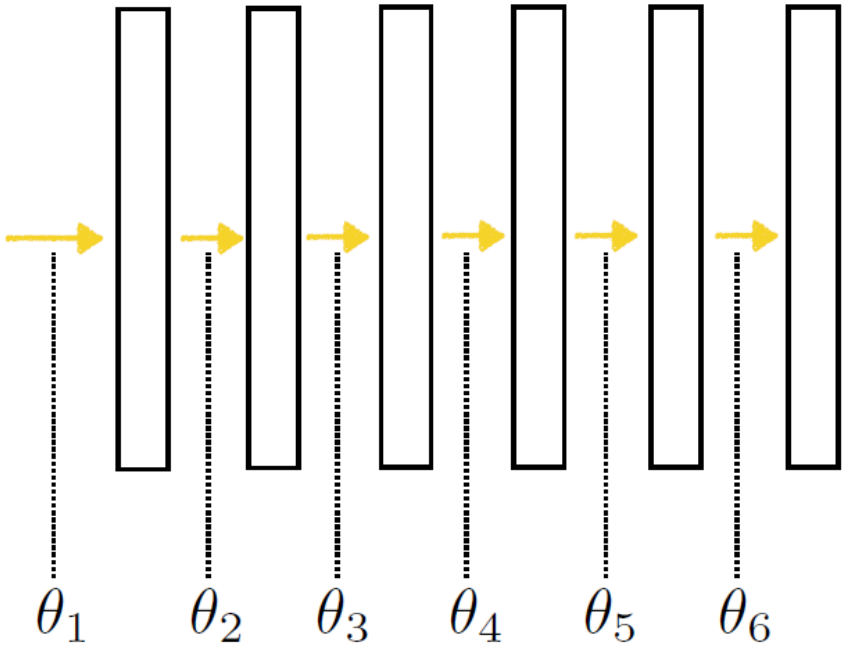


Deep learning

Learned

$\mathbf{y}_i$
"clown fish"

$\mathbf{x}_i$

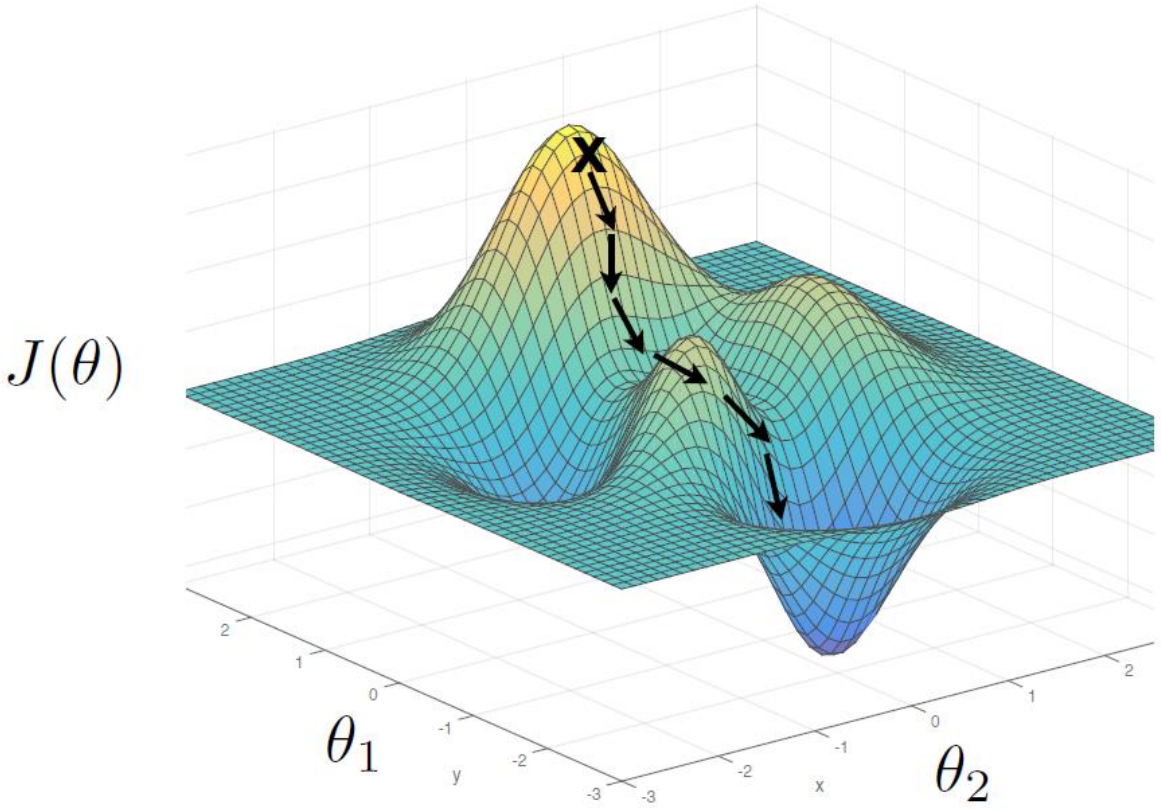


$$\theta^* = \arg\min_\theta \sum_{i=1}^{N} \mathcal{L}(f_\theta(\mathbf{x}_i), \mathbf{y}_i)$$

Gradient descent

$$\theta^* = \arg\min_{\theta} \underbrace{\sum_{i=1}^{N} \mathcal{L}(f_\theta(\mathbf{x}_i), \mathbf{y}_i)}_{J(\theta)}$$

Gradient Descent



$$\theta^* = \arg\min_{\theta} J(\theta)$$

Gradient Descent

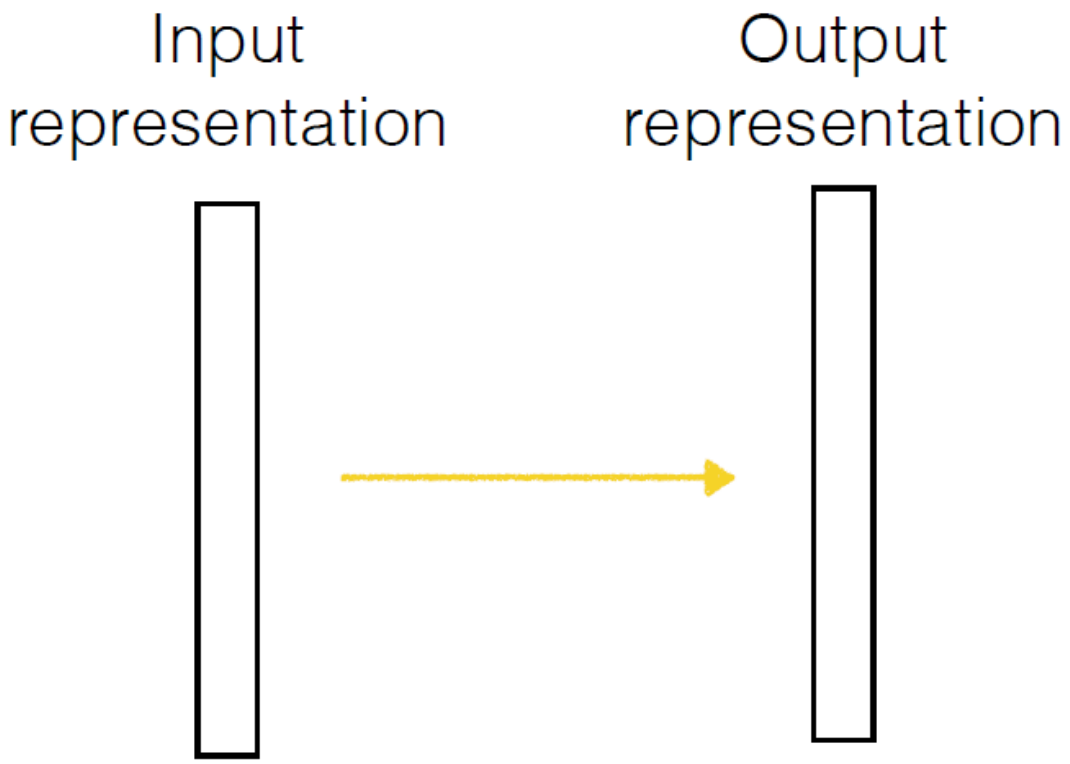
$$\theta^* = \arg\min_{\theta} \underbrace{\sum_{i=1}^{N} \mathcal{L}(f_\theta(\mathbf{x}_i), \mathbf{y}_i)}_{J(\theta)}$$

One iteration of gradient descent:

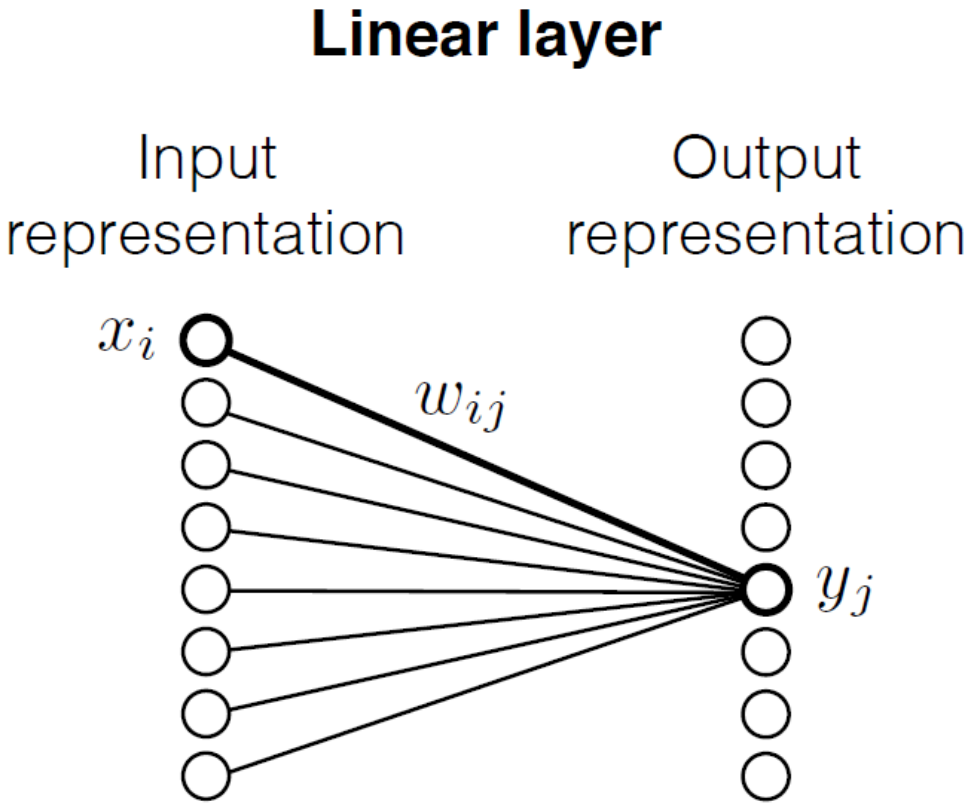
$$\theta^{t+1} = \theta^t - \eta_t \left. \frac{\partial J(\theta)}{\partial \theta} \right|_{\theta=\theta^t}$$

learning rate

Computation in a neural network

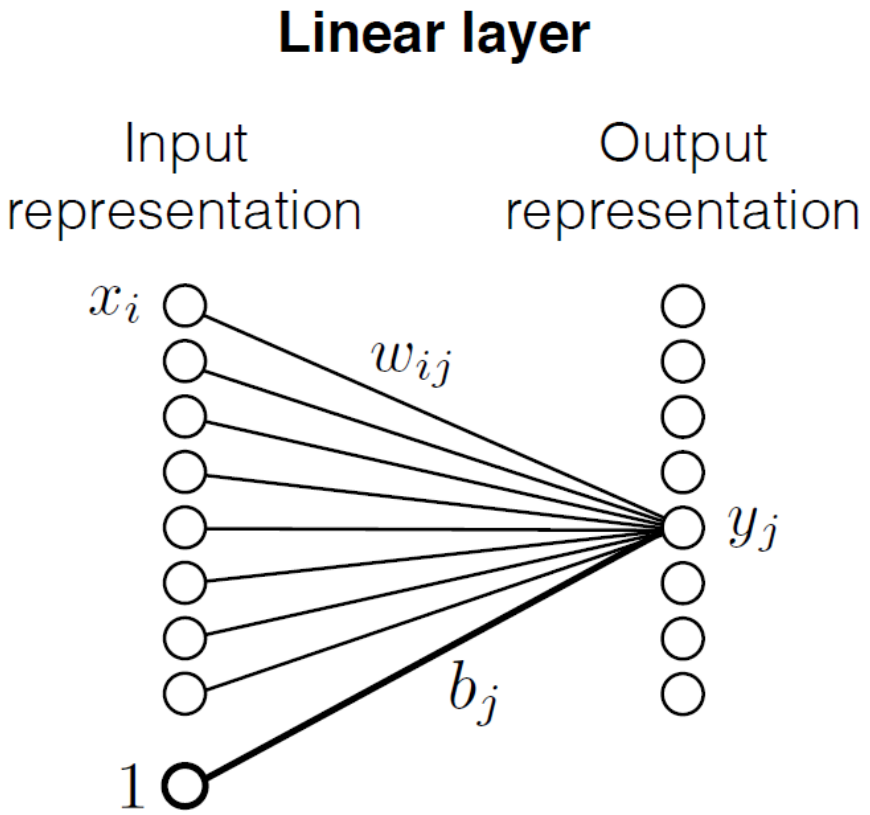


Computation in a neural network



$$y_j = \sum_i w_{ij} x_i$$

Computation in a neural network

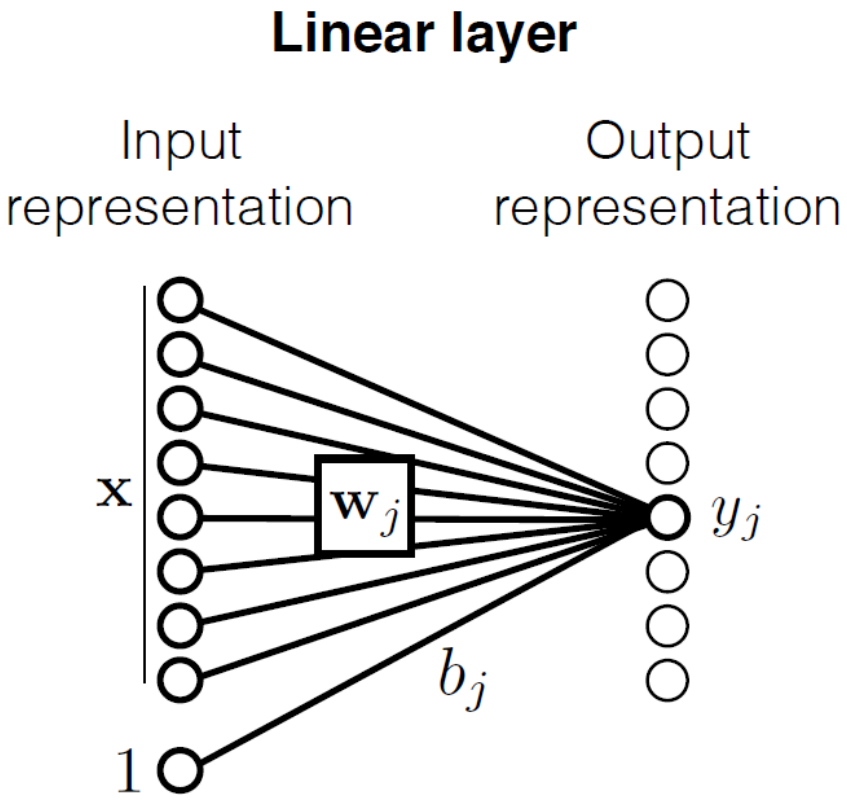


$$y_j = \sum_i w_{ij} x_i + b_j$$

weights

bias

Computation in a neural network



$$y_j = \mathbf{x}^T \mathbf{w}_j + b_j$$

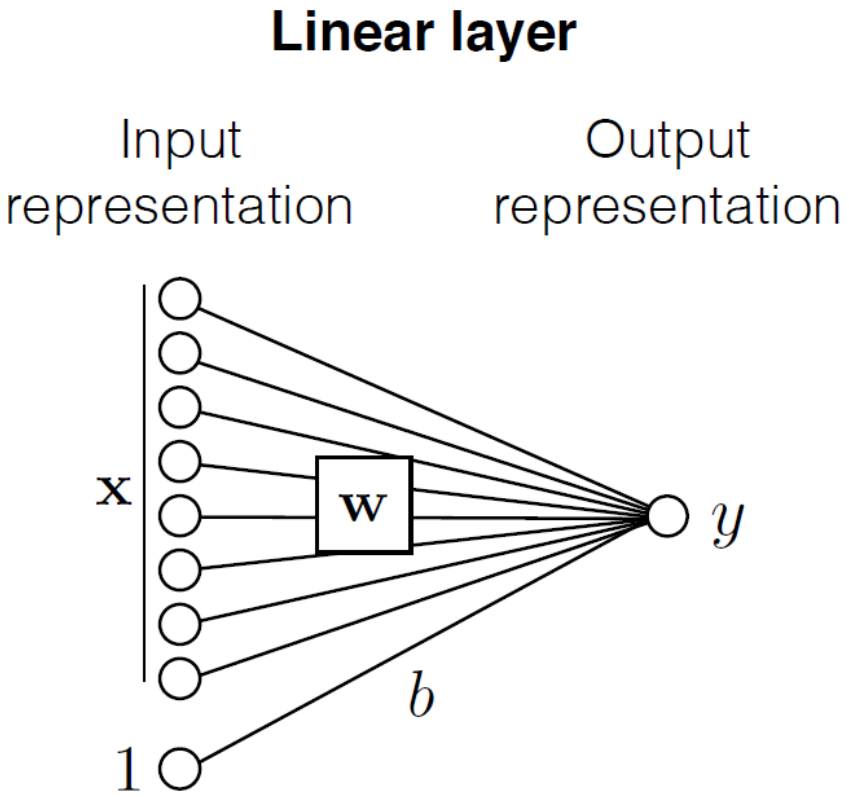
weights

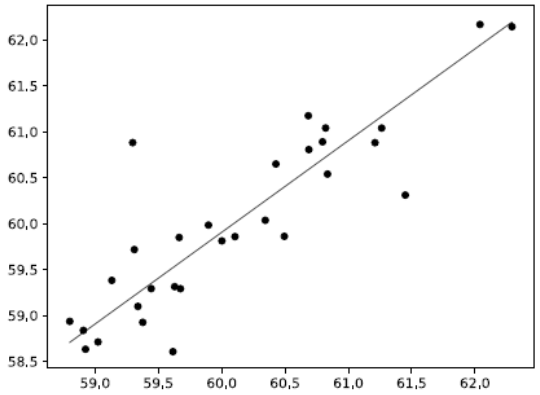
bias

$$\theta = \{\mathbf{W}, \mathbf{b}\}$$

parameters of the model

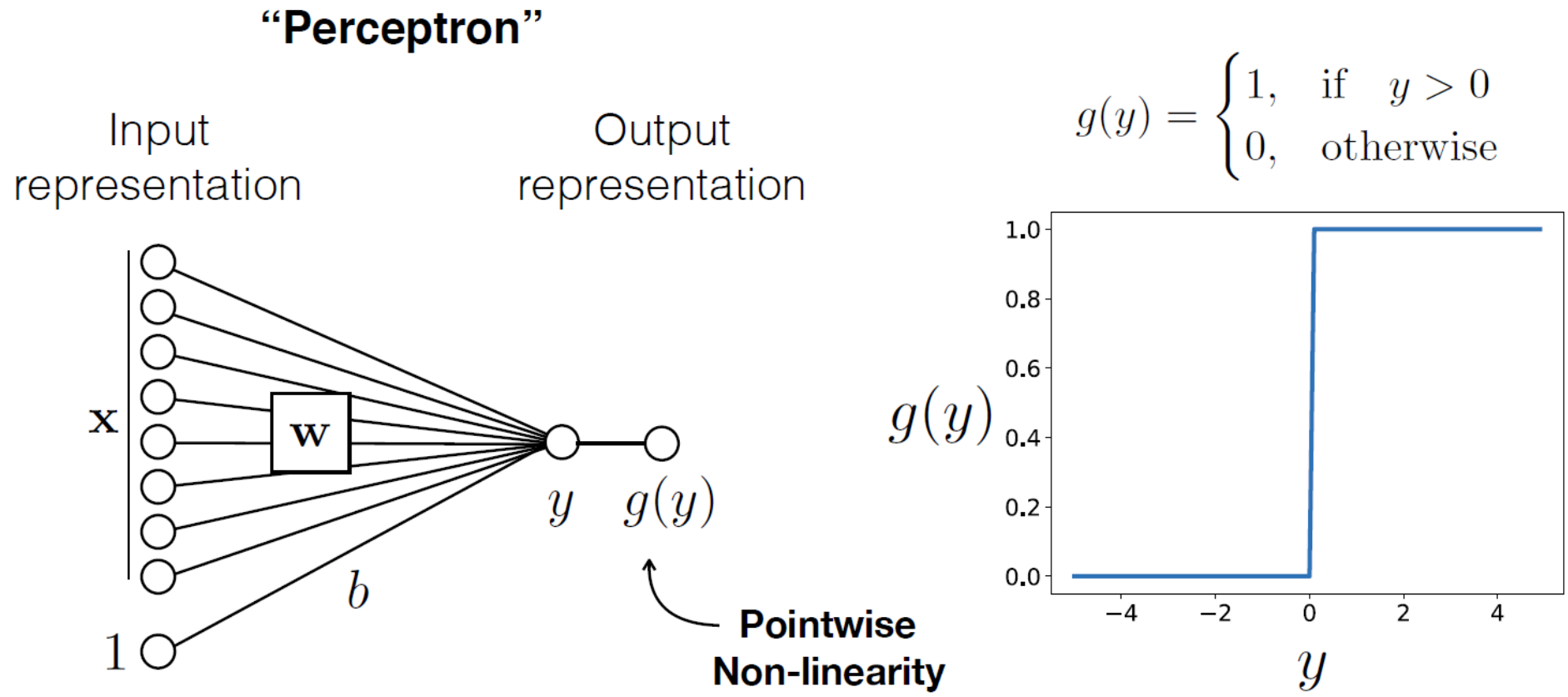
Example: linear regression with a neural net



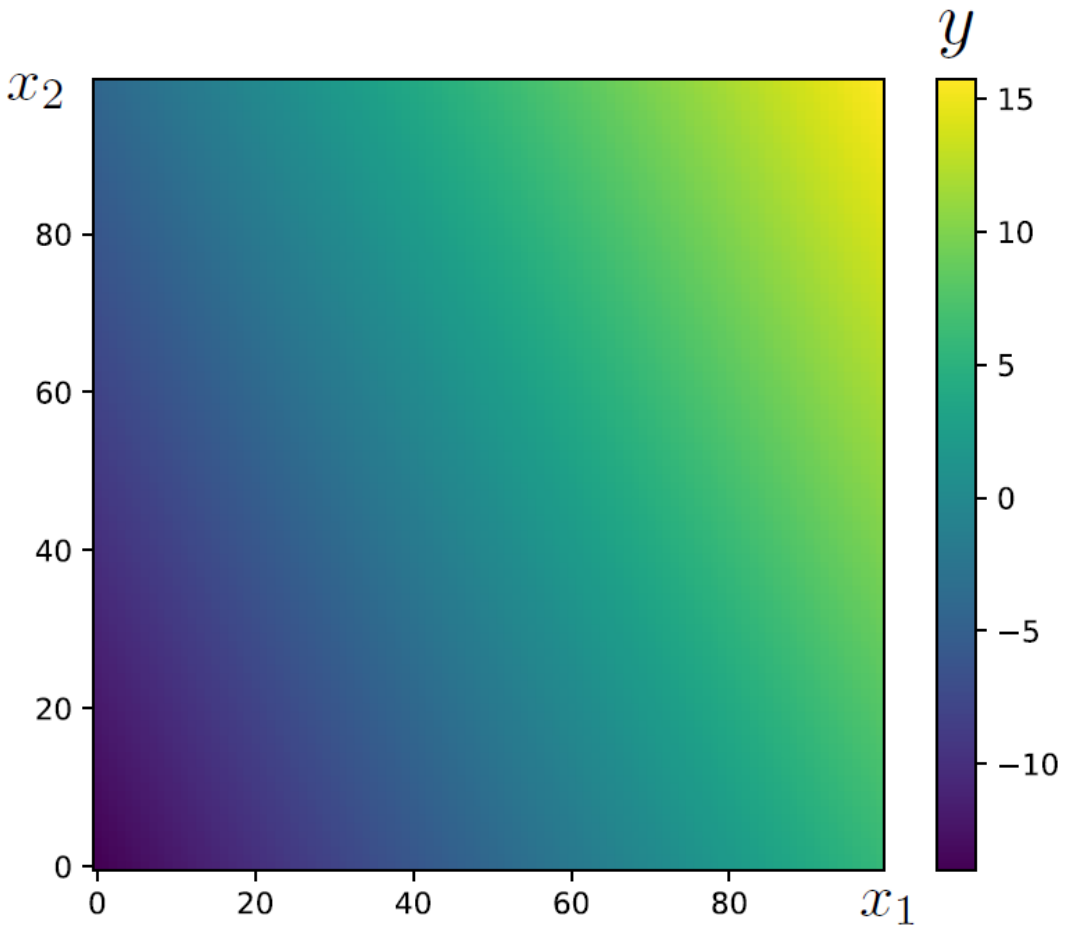


$$f_{\mathbf{w},b}(\mathbf{x}) = \mathbf{x}^T\mathbf{w} + b$$

Computation in a neural network

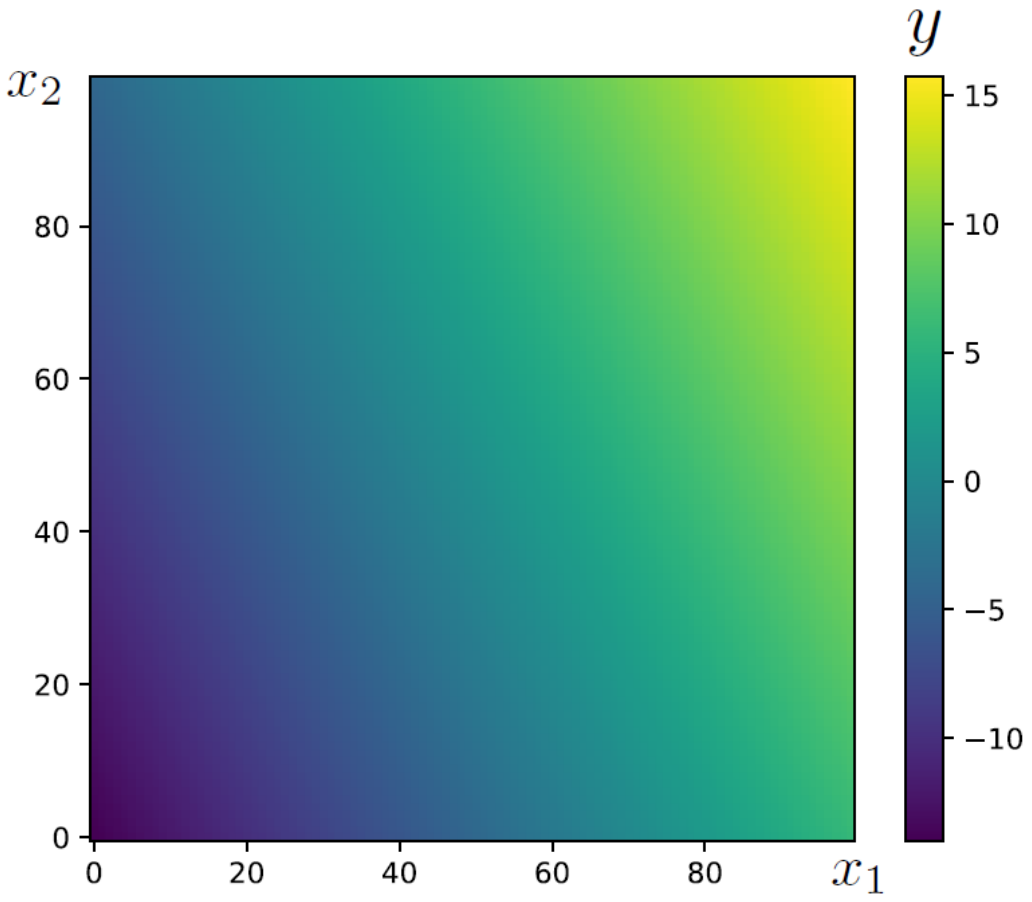


Example: linear classification with a perceptron



$$y = \mathbf{x}^T \mathbf{w} + b$$

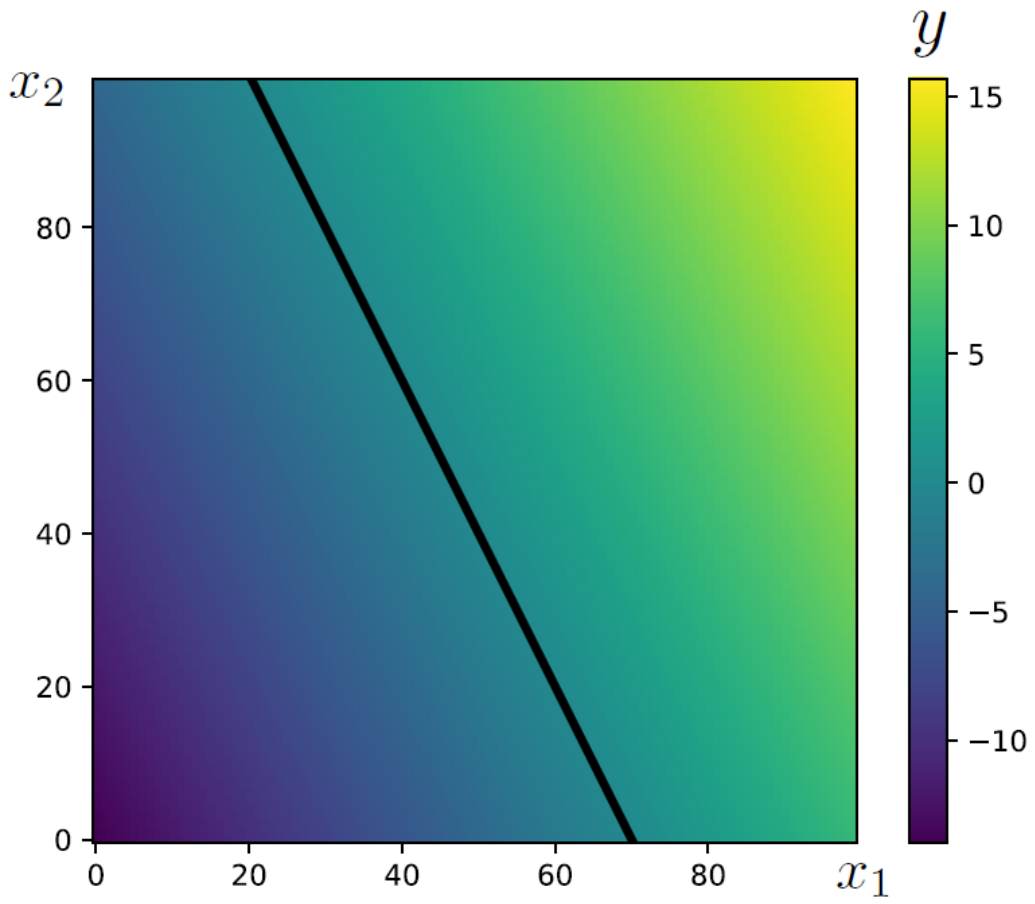
Example: linear classification with a perceptron



$$y = \mathbf{x}^T \mathbf{w} + b$$

$$g(y) = \begin{cases} 1, & \text{if} \quad y > 0 \\ 0, & \text{otherwise} \end{cases}$$

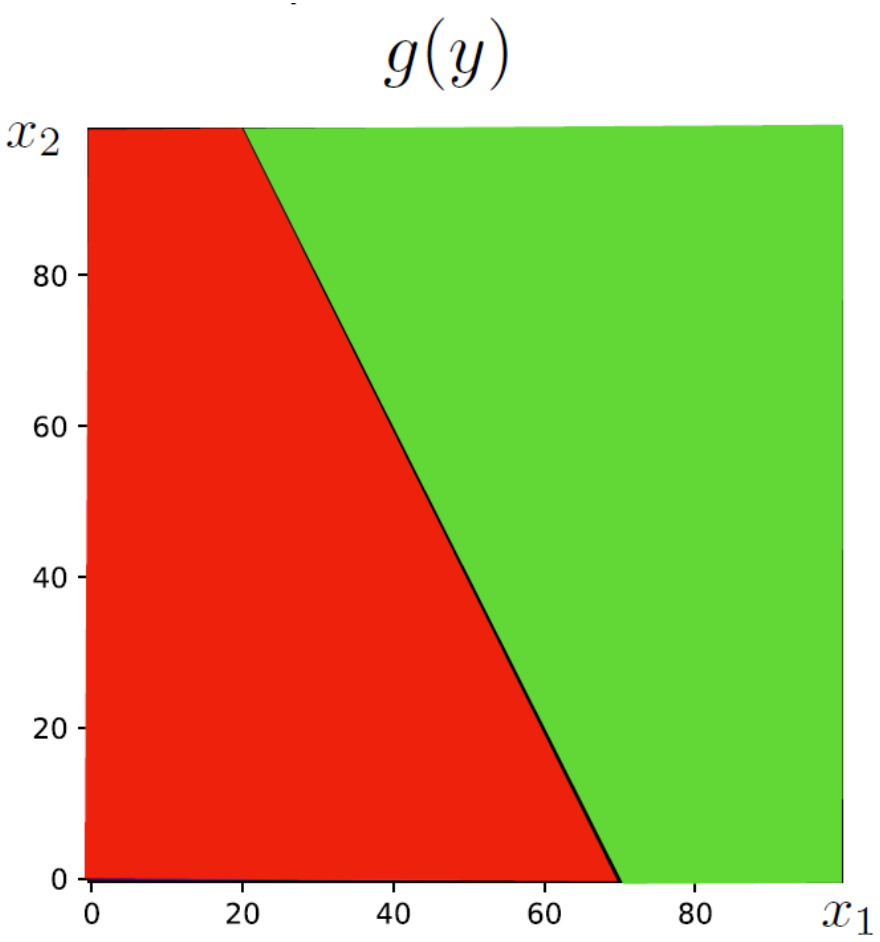
Example: linear classification with a perceptron



$$y = \mathbf{x}^T \mathbf{w} + b$$

$$g(y) = \begin{cases} 1, & \text{if} \quad y > 0 \\ 0, & \text{otherwise} \end{cases}$$

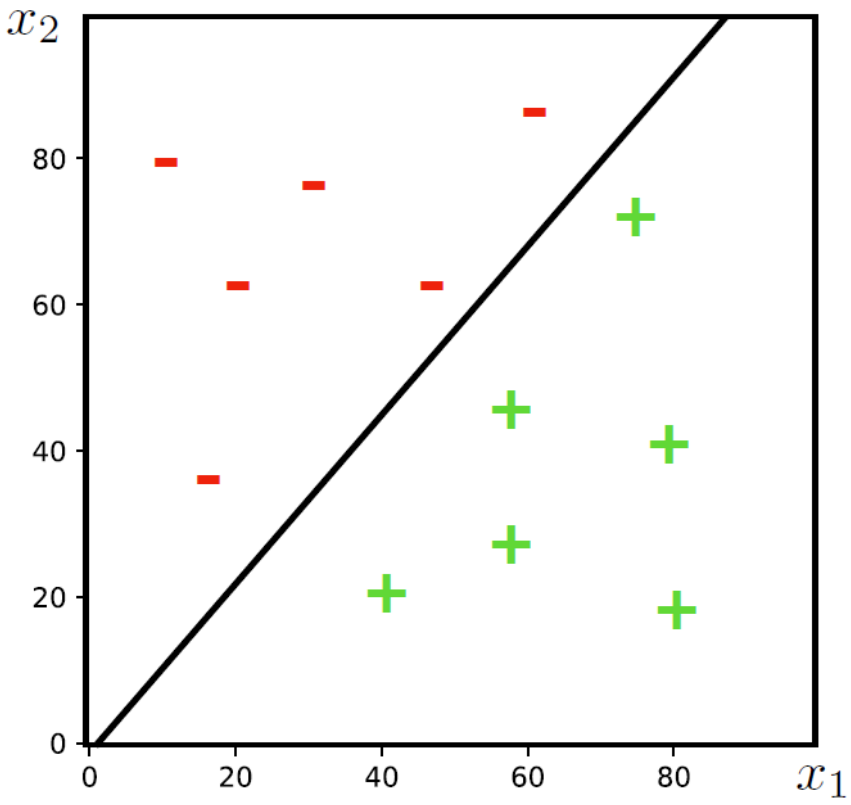
Example: linear classification with a perceptron



$$y = \mathbf{x}^T \mathbf{w} + b$$

$$g(y) = \begin{cases} 1, & \text{if} \quad y > 0 \\ 0, & \text{otherwise} \end{cases}$$

Example: linear classification with a perceptron

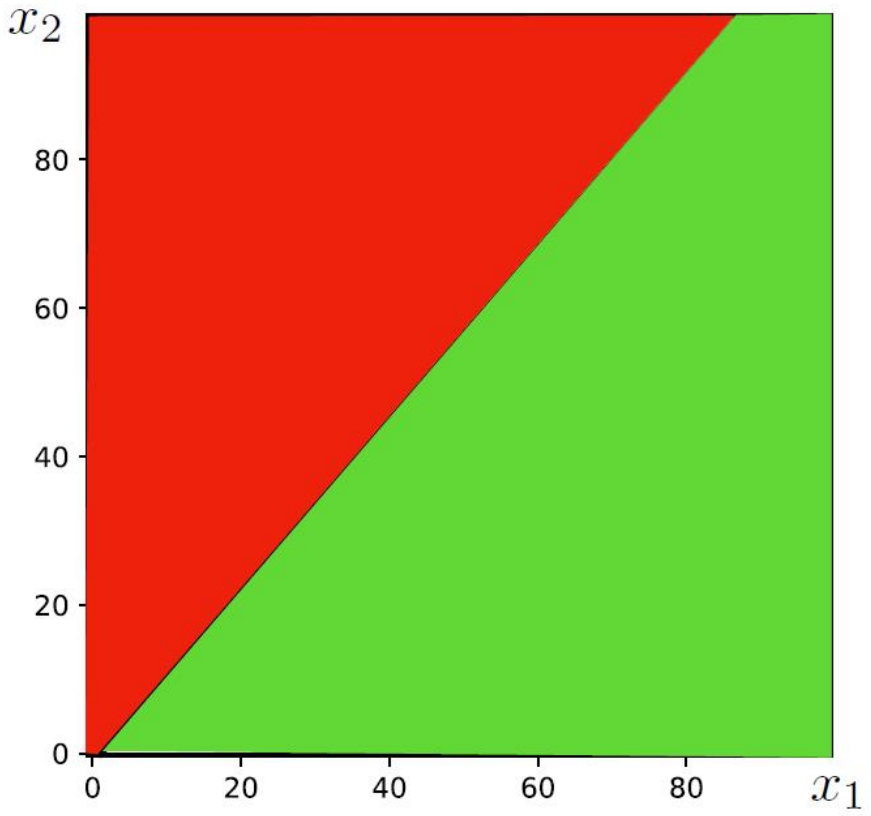


$$\hat{y} = \mathbf{x}^T \mathbf{w} + b$$

$$g(\hat{y}) = \begin{cases} 1, & \text{if} \quad \hat{y} > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbf{w}^*, b^* = \arg\min_{\mathbf{w}, b} \mathcal{L}(g(\hat{y}), y_i)$$

Example: linear classification with a perceptron

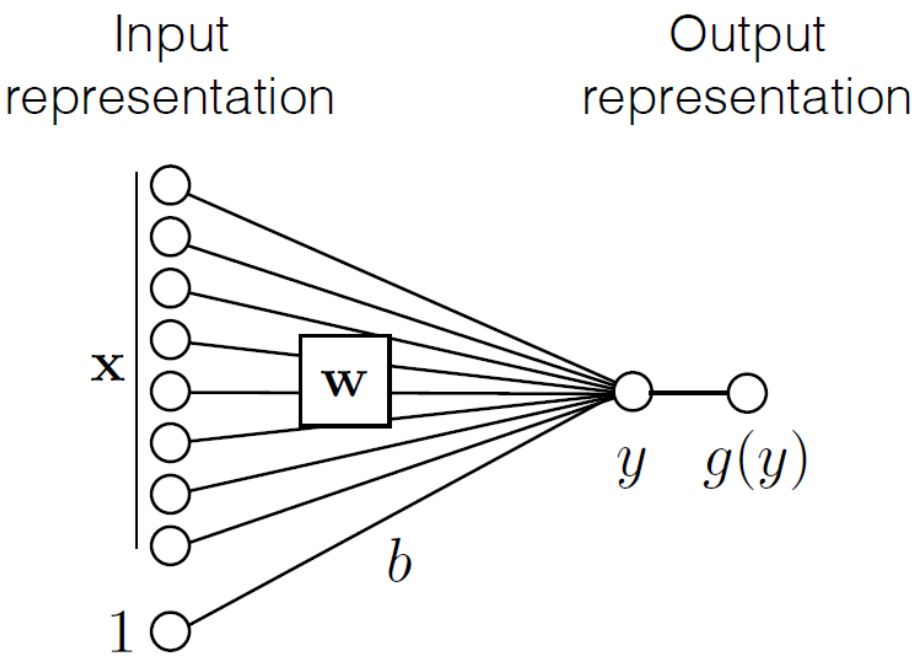


$$\hat{y} = \mathbf{x}^T \mathbf{w} + b$$

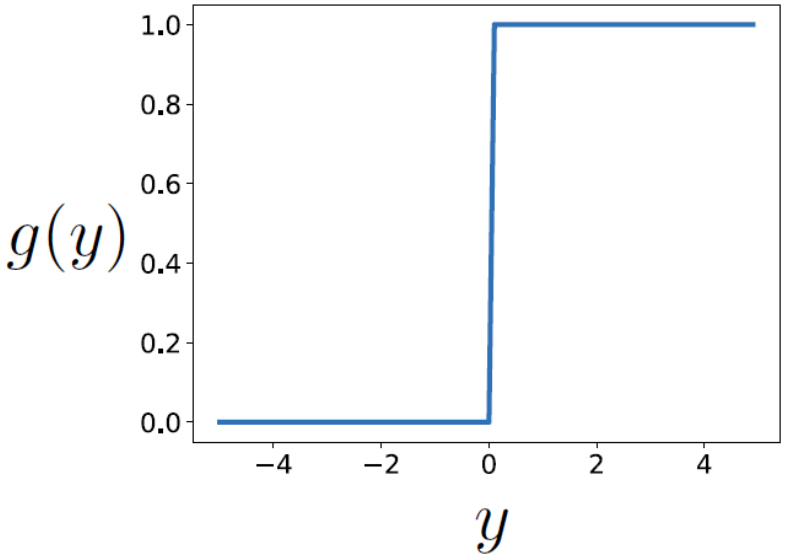
$$g(\hat{y}) = \begin{cases} 1, & \text{if} \quad \hat{y} > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbf{w}^*, b^* = \arg\min_{\mathbf{w},b} \mathcal{L}(g(\hat{y}), y_i)$$

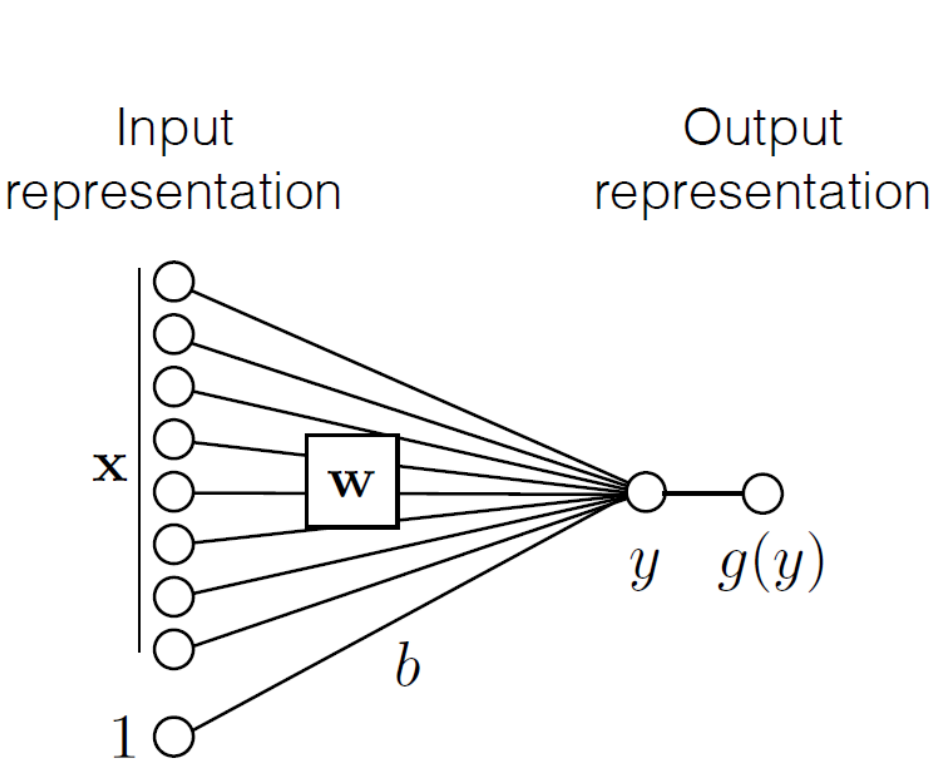
Computation in a neural network



$$g(y) = \begin{cases} 1, & \text{if} \quad y > 0 \\ 0, & \text{otherwise} \end{cases}$$

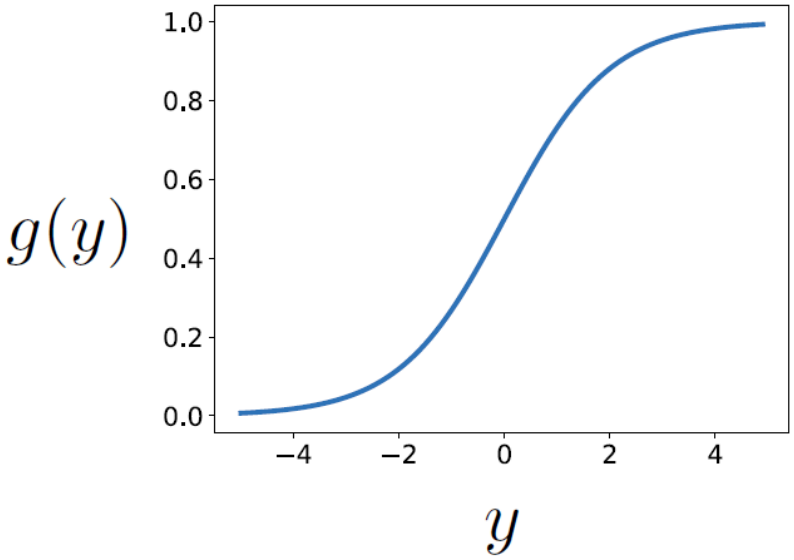


Computation in a neural network - nonlinearity



Sigmoid

$$g(y) = \frac{1}{1 + e^{-y}}$$

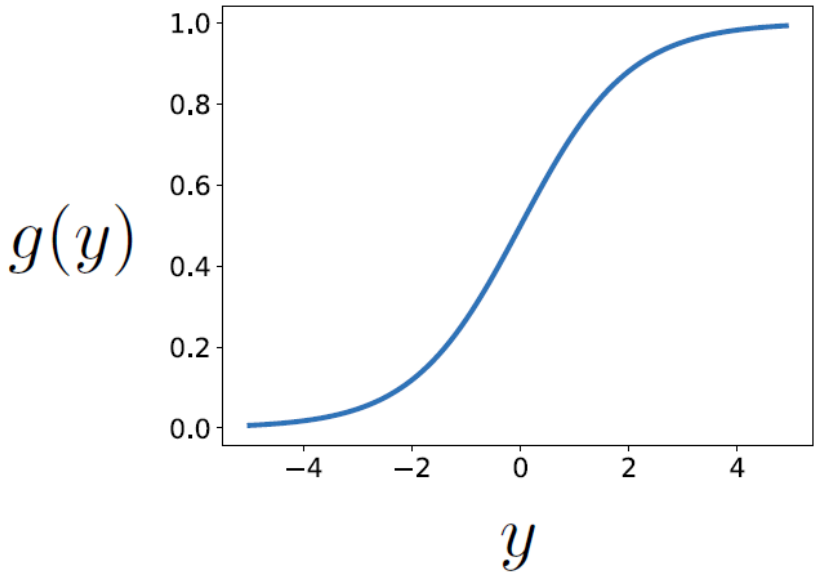


Computation in a neural network - nonlinearity

- Interpretation as firing rate of neuron
- Bounded between [0,1]
- Saturation for large +/- inputs
- Gradients go to zero
- Output centered at 0.5 (poor conditioning)
- Not used in practice

Sigmoid

$$g(y) = \frac{1}{1 + e^{-y}}$$

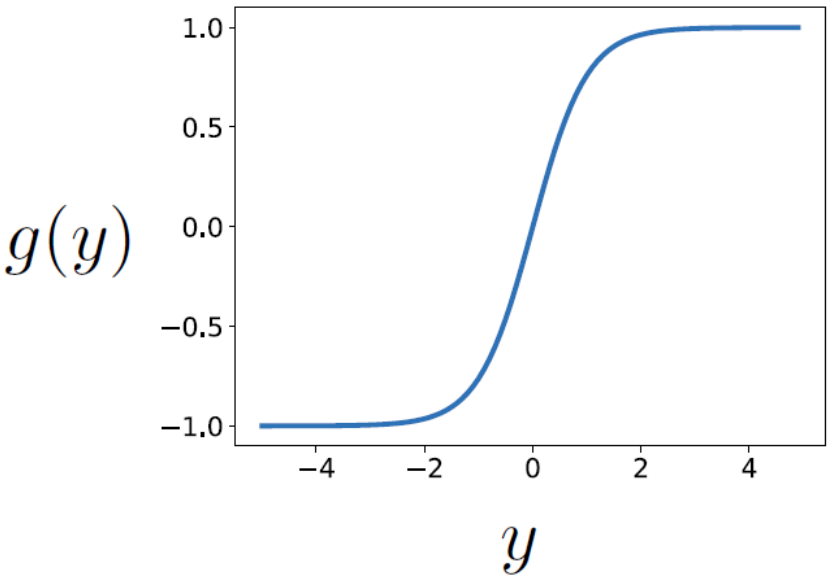


Computation in a neural network - nonlinearity

- Bounded between [-1,+1]
- Saturation for large +/- inputs
- Gradients go to zero
- Output centered at 0
- Preferable to sigmoid
- $\tanh(x) = 2\,sigmoid(2x) - 1$

Tanh

$$g(y) = \frac{e^{y} - e^{-y}}{e^{y} + e^{-y}}$$

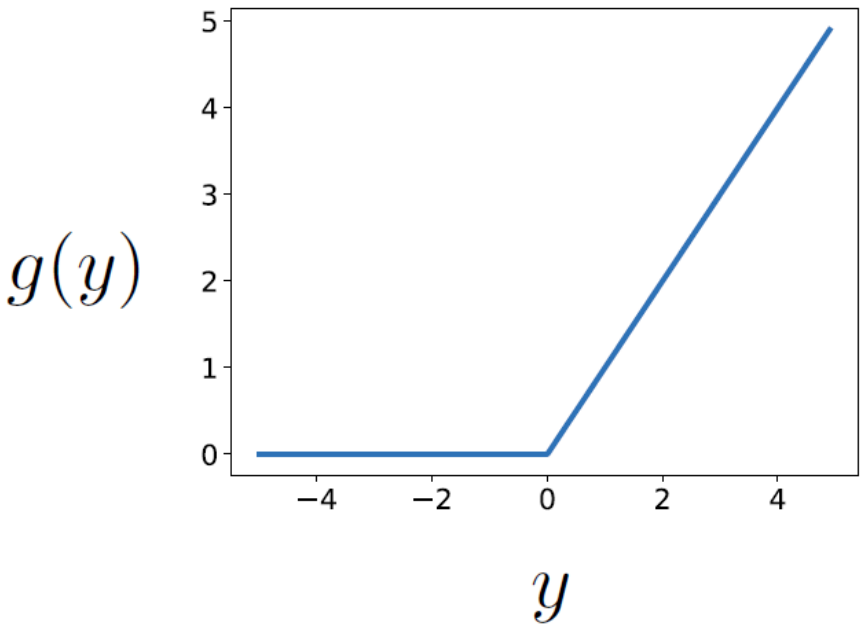


Computation in a neural network - nonlinearity

- unbounded output on positive side
- Efficient to implement
  - $\frac{\partial g}{\partial y} = \begin{cases} 0, if\ y < 0 \\ 1, if\ y \geq 0 \end{cases}$
- Also seems to help convergence
- Drawback: if strongly in negative region, unit is dead forever (no gradient)
- Default choice: widely used in current models.

Rectified linear unit (ReLU)

$$g(y) = \max(0, y)$$

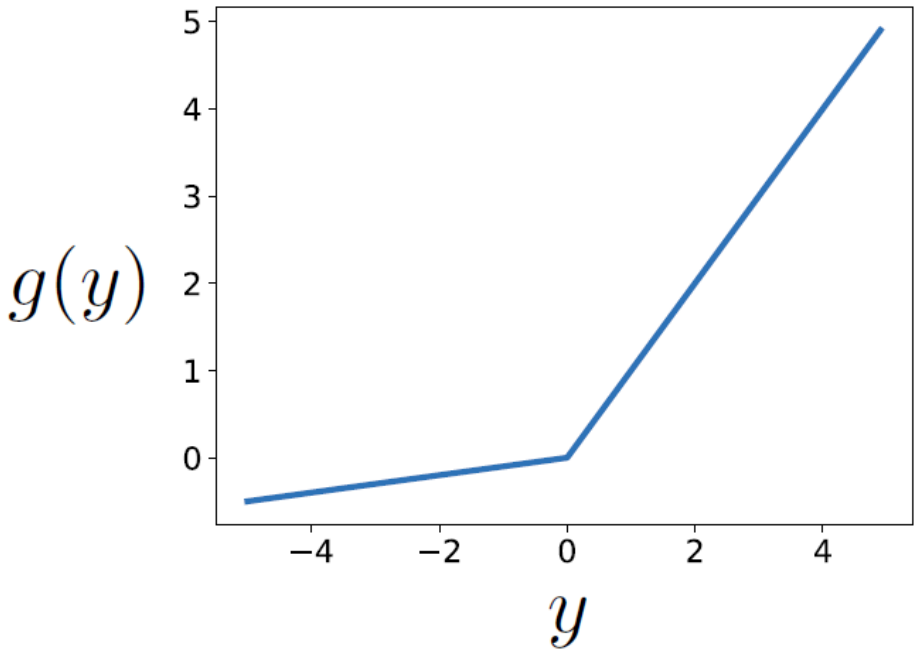


Computation in a neural network - nonlinearity

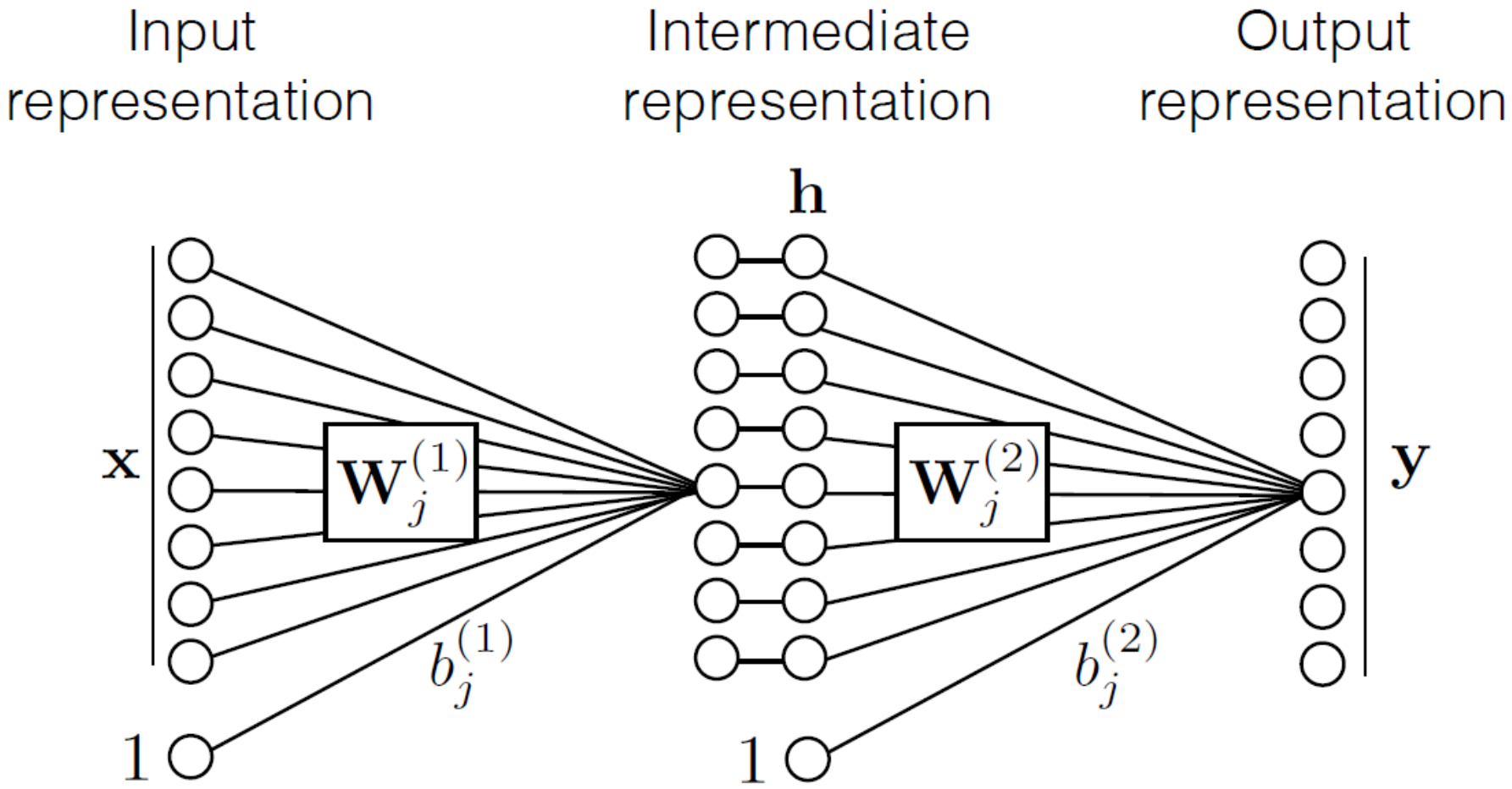
- Where ***a*** is small (e.g. 0.02)
- Efficient to implement
  - $\frac{\partial g}{\partial y} = \begin{cases} -a, & if\ y < 0 \\ 1, & if\ y \geq 0 \end{cases}$
- Also known as probabilistic Relu (PReLU)
- Has non-zero gradients everywhere (unlike ReLu)
- ***a*** can also be learned (see Kaiming He et al. 2015)

Leaky ReLU

$$g(y) = \begin{cases} \max(0, y), & \text{if} \quad y \geq 0 \\ a \min(0, y), & \text{if} \quad y < 0 \end{cases}$$



Stacking layers



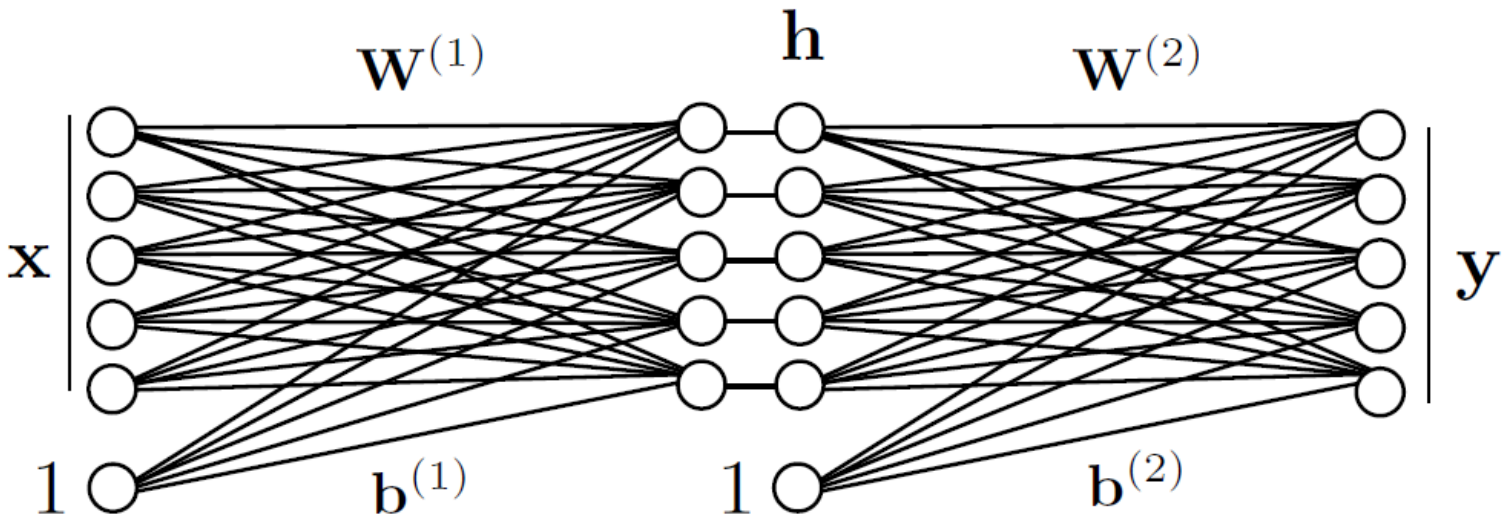
$\mathbf{h}$ = **"hidden units"**

Stacking layers

Input representation

Intermediate representation

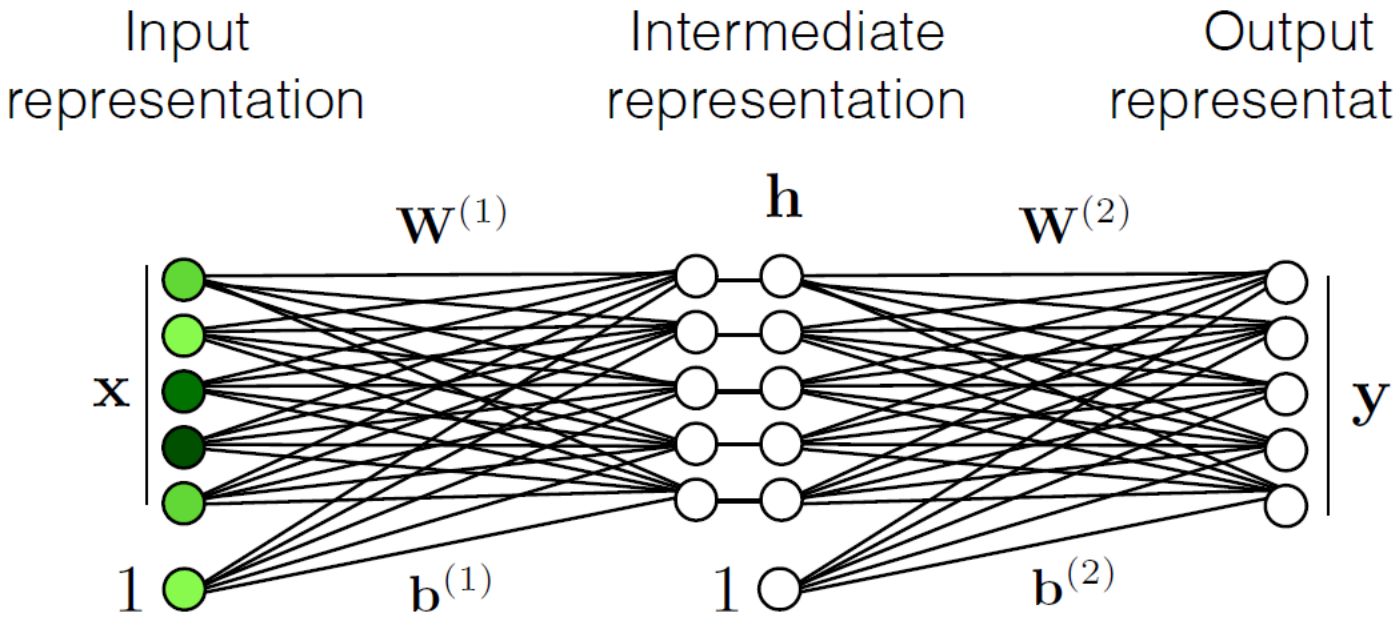
Output representation



$$\mathbf{h} = g(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) \qquad \mathbf{y} = \mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)}$$

$$\theta = \{\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)}, \mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(L)}\}$$

Stacking layers



positive

negative

$$\mathbf{h} = g(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) \qquad \mathbf{y} = \mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)}$$

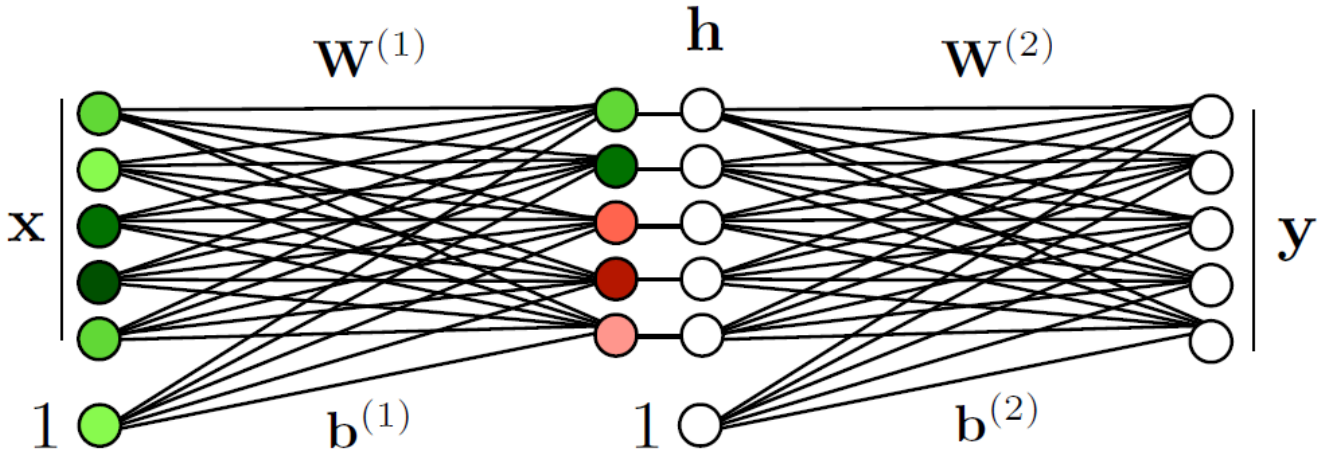
$$\theta = \{\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)}, \mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(L)}\}$$

Stacking layers

Input representation

Intermediate representation

Output representation



positive

negative

$$\mathbf{h} = g(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) \qquad \mathbf{y} = \mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)}$$

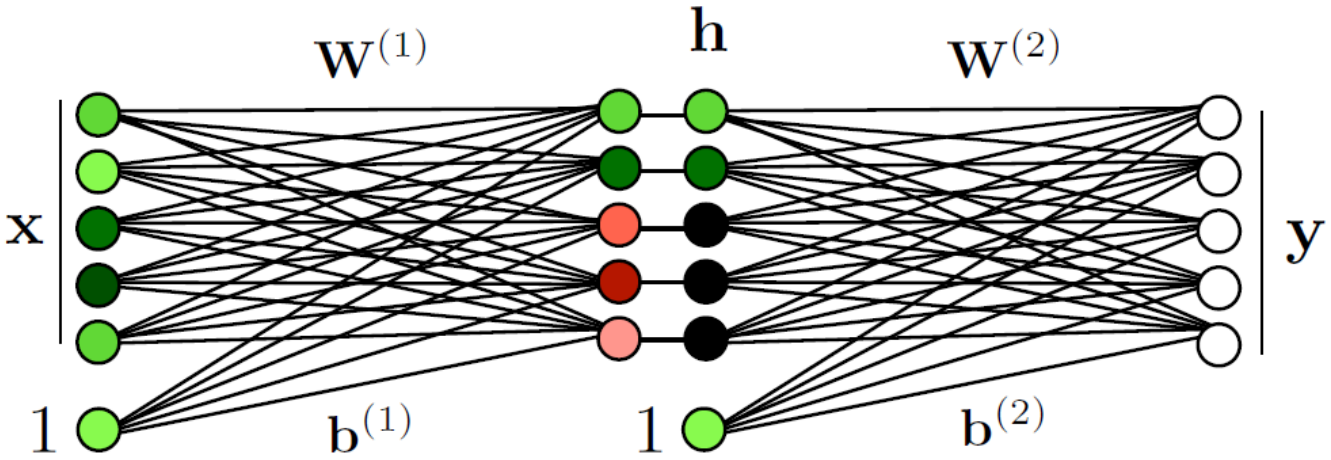
$$\theta = \{\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)}, \mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(L)}\}$$

Stacking layers

Input representation

Intermediate representation

Output representation



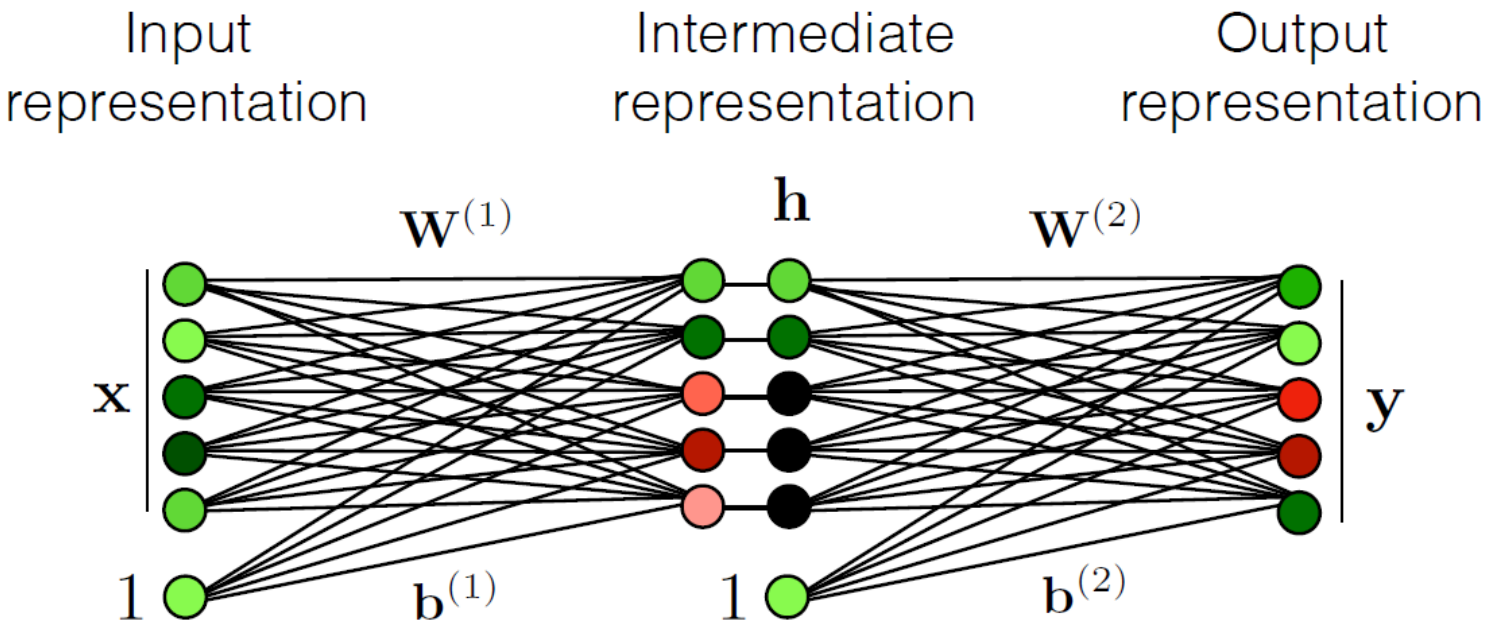
positive

negative

$$\mathbf{h} = g(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) \qquad \mathbf{y} = \mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)}$$

$$\theta = \{\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)}, \mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(L)}\}$$

Stacking layers



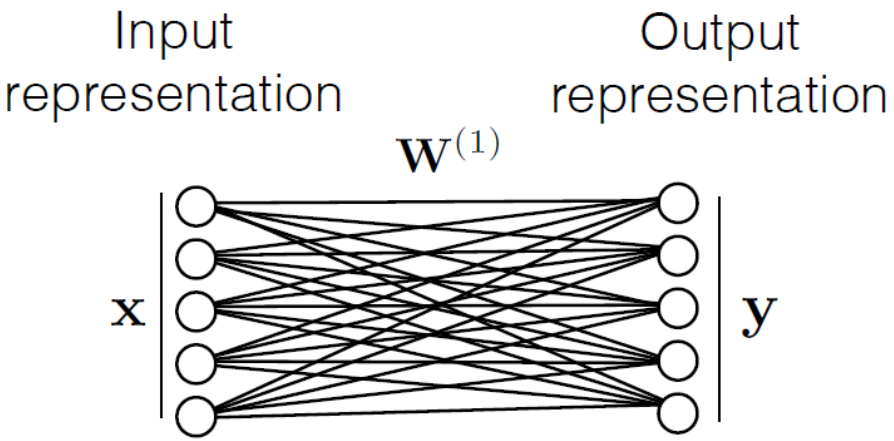
positive

negative

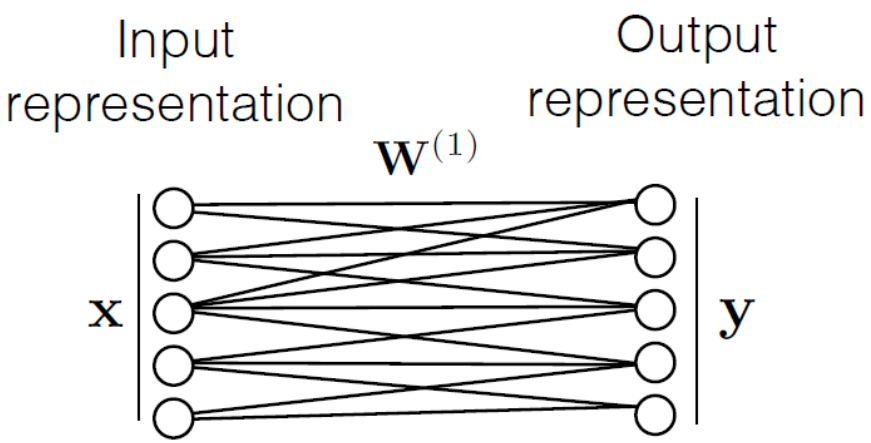
$$\mathbf{h} = g(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) \qquad \mathbf{y} = \mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)}$$

$$\theta = \{\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)}, \mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(L)}\}$$

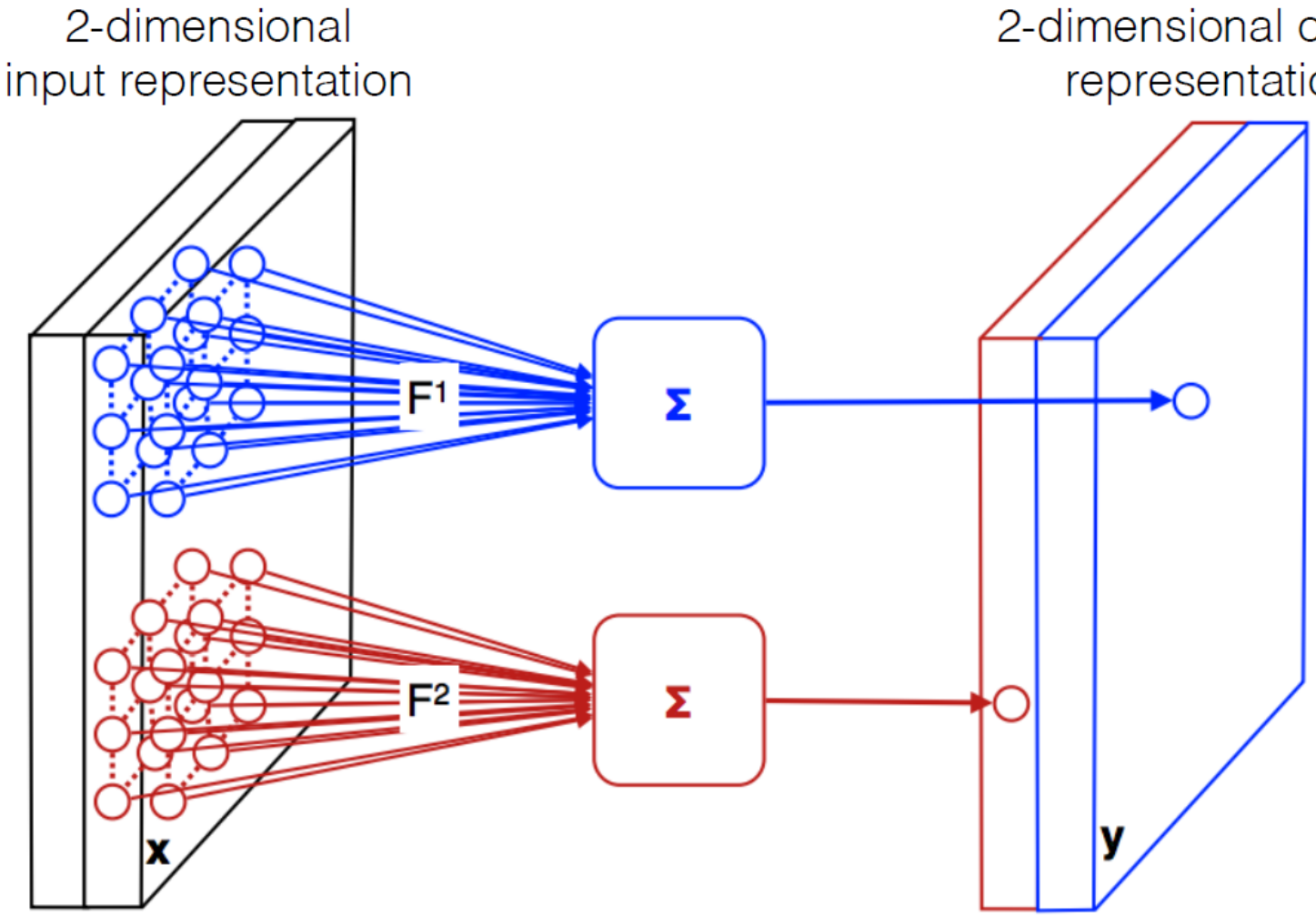
Connectivity patterns



Fully connected layer



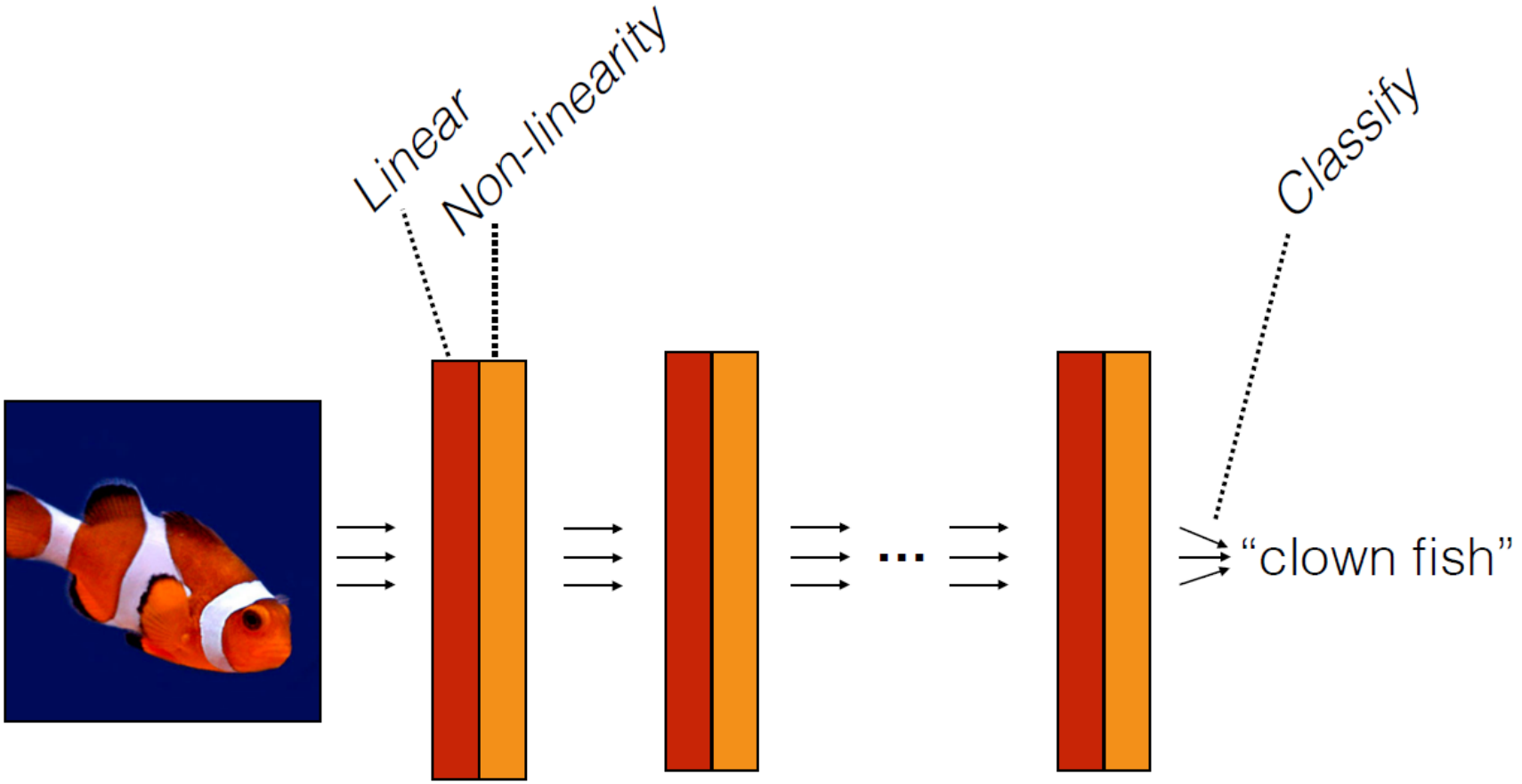
Locally connected layer (Sparse W)



$$\mathbb{R}^{H \times W \times C^{(l)}} \rightarrow \mathbb{R}^{H \times W \times C^{(l+1)}}$$

[Figure from Andrea Vedaldi]

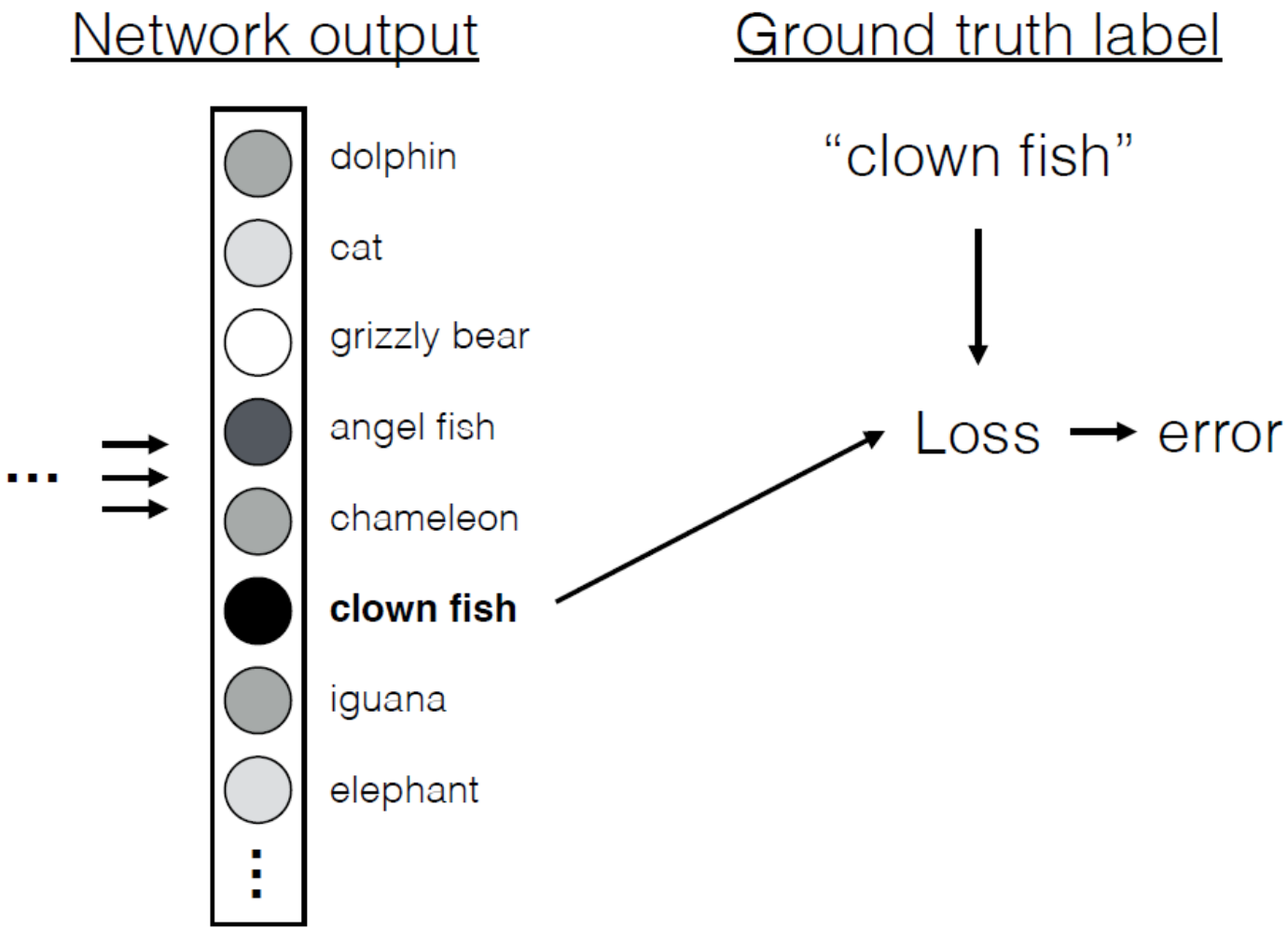
Deep nets



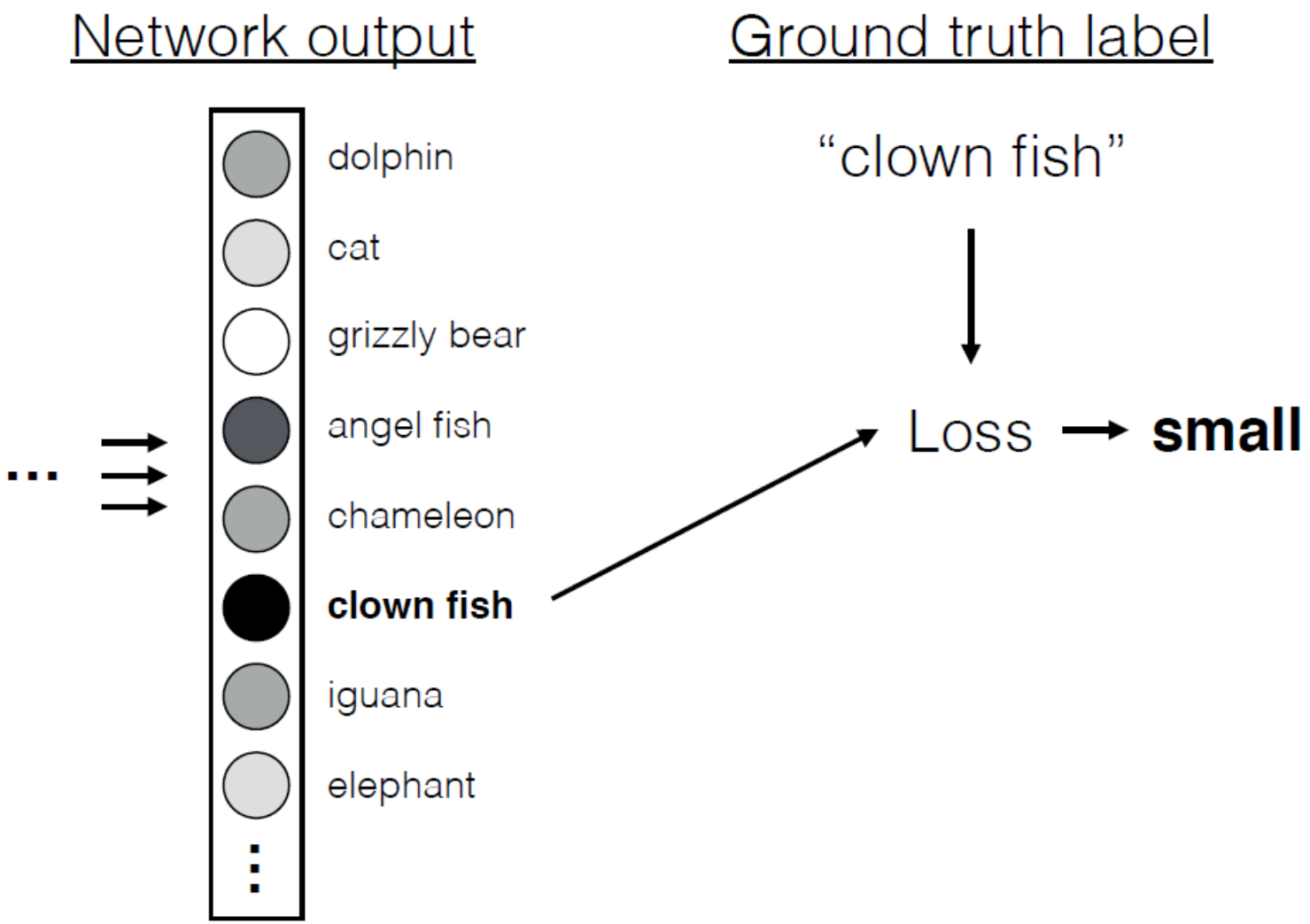
$$f(\mathbf{x}) = f_L(\ldots f_2(f_1(\mathbf{x})))$$

Classifier layer

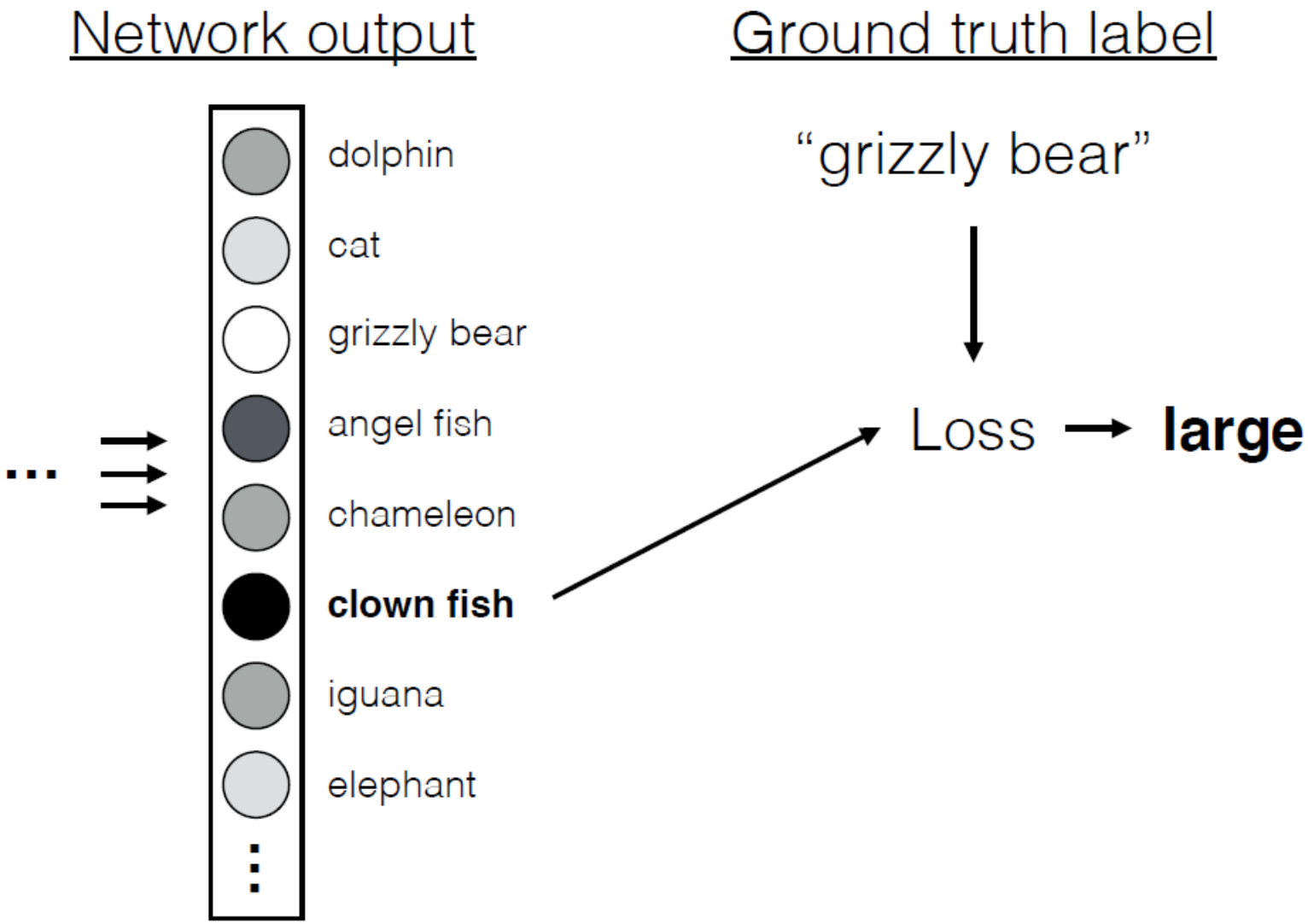
Loss function

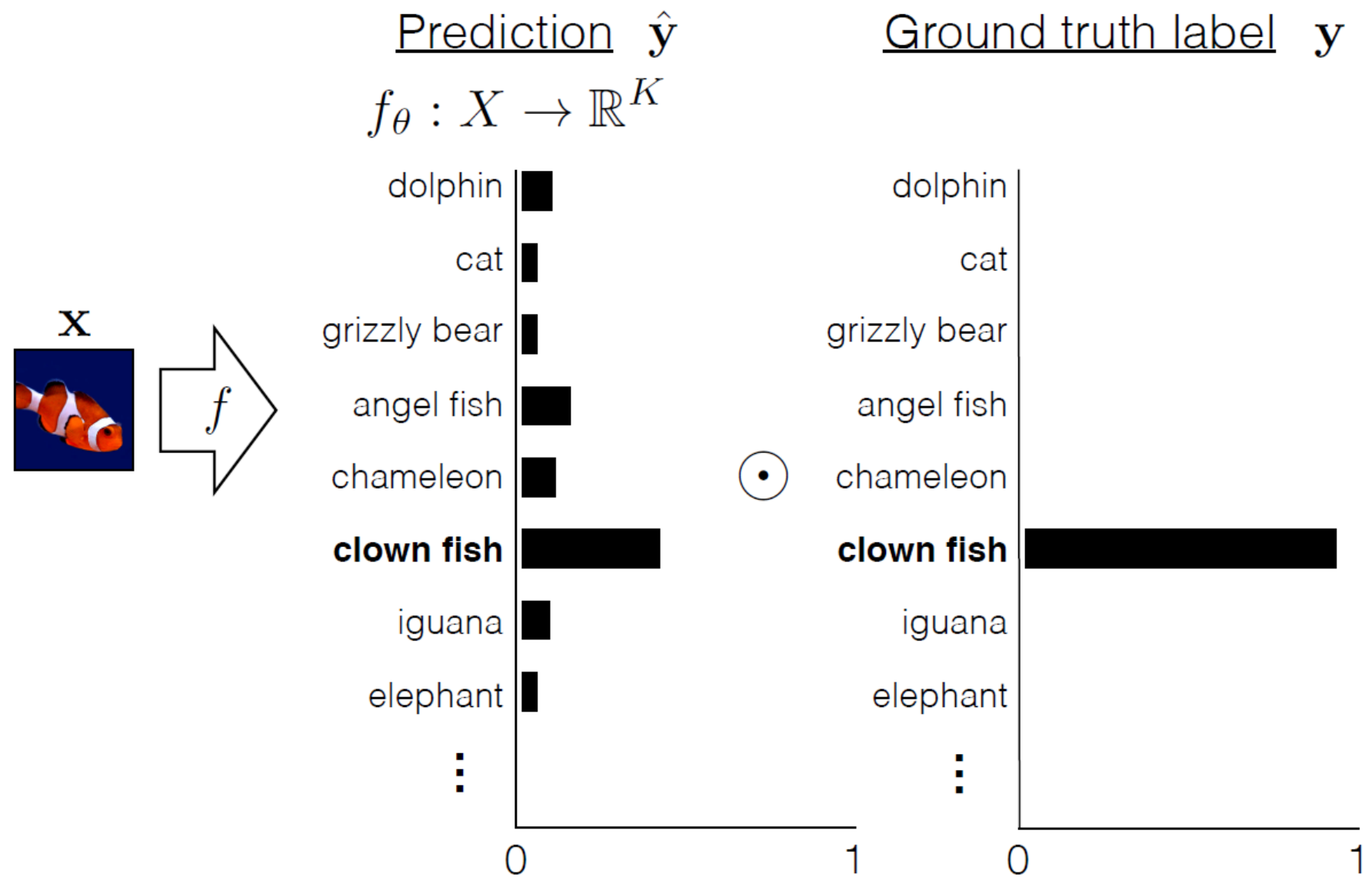


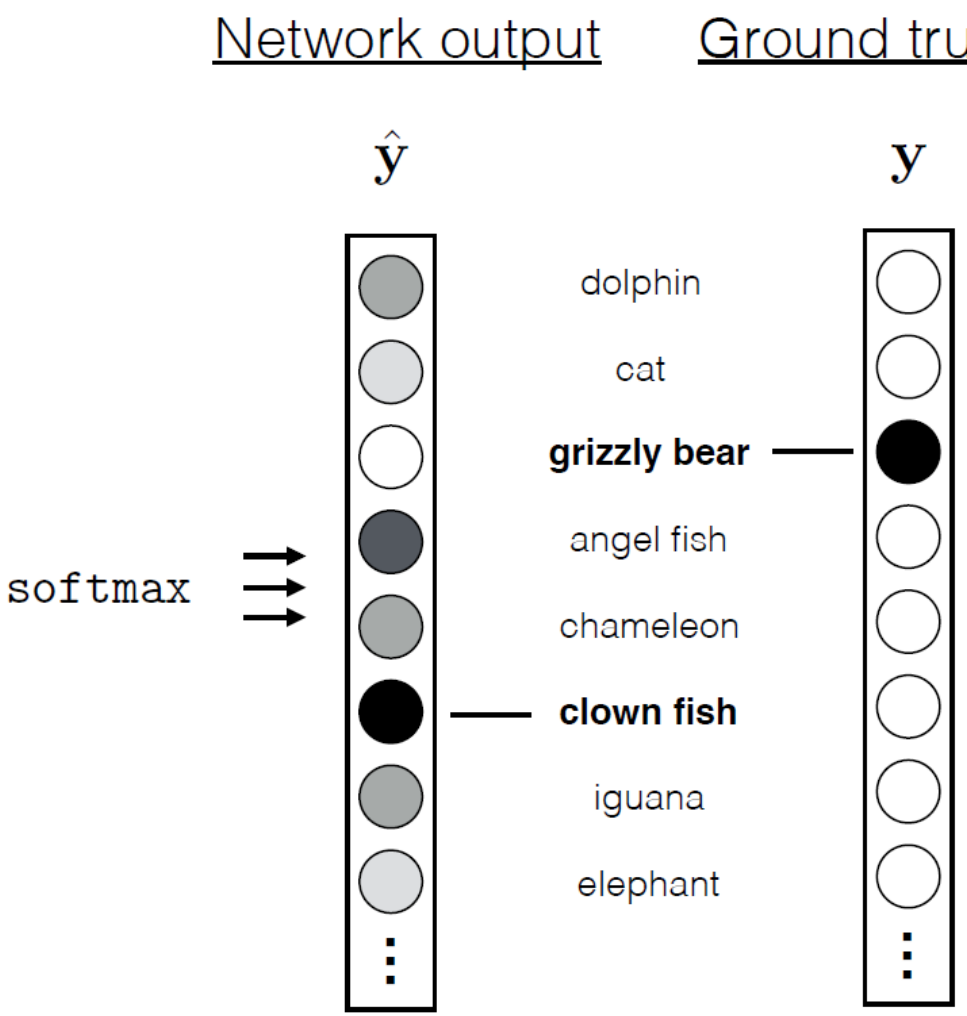
Loss function



Loss function




Prediction $\hat{\mathbf{y}}$
Ground truth label $\mathbf{y}$
$f_\theta : X \to \mathbb{R}^K$
$\mathbf{x}$
f
dolphin
cat
grizzly bear
angel fish
chameleon
clown fish
iguana
elephant
0
1
dolphin
cat
grizzly bear
angel fish
chameleon
clown fish
iguana
elephant
0
1



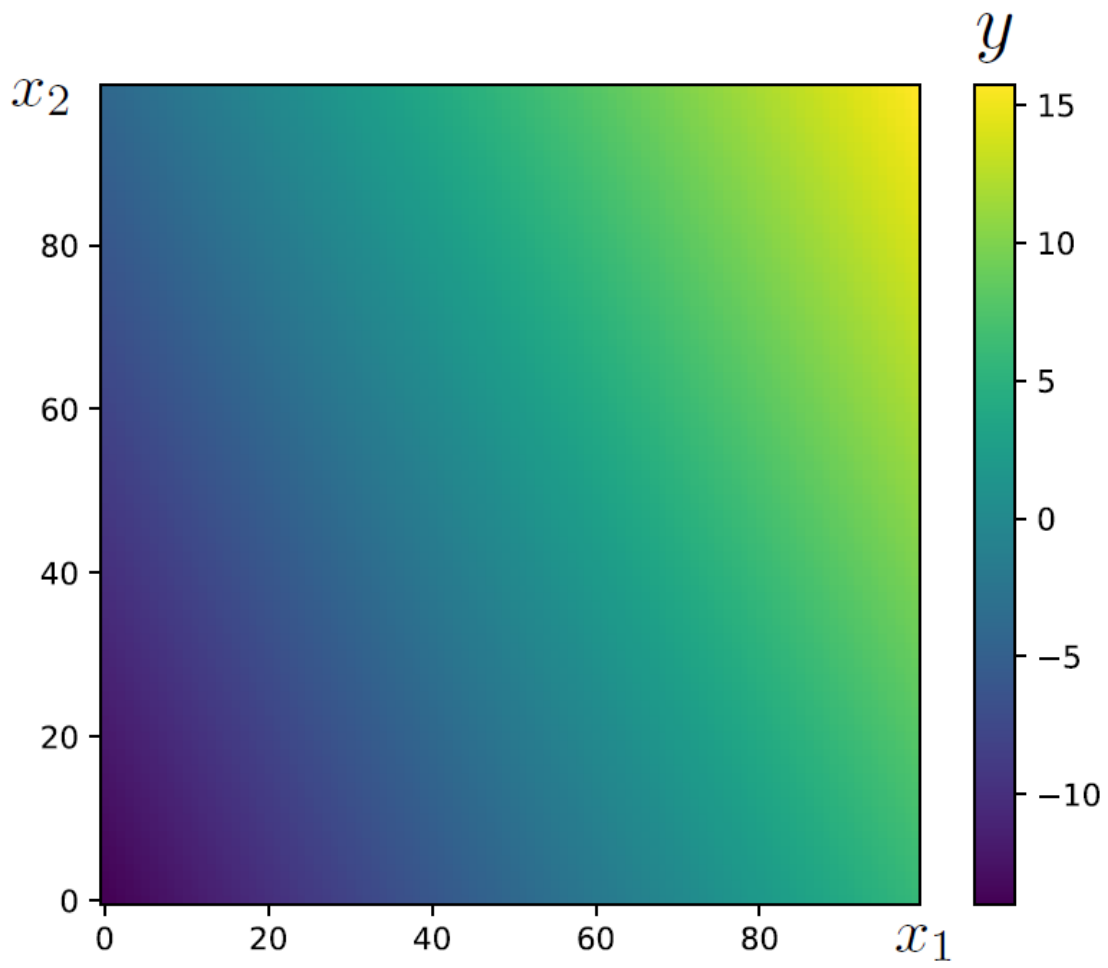
Probability of the observed data under the model

$$H(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{k=1}^{K} y_k \log \hat{y}_k$$

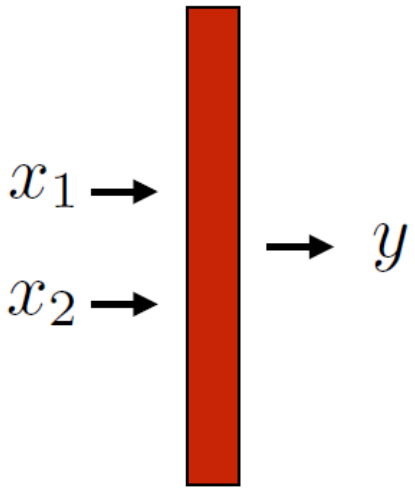
Representational power

- 1 layer? Linear decision surface.
- 2+ layers? In theory, can represent any function. Assuming non-trivial non-linearity.
  - Bengio 2009, http://www.iro.umontreal.ca/~bengioy/papers/ftml.pdf
  - Bengio, Courville, Goodfellow book http://www.deeplearningbook.org/contents/mlp.html
  - Simple proof by M. Neilsen http://neuralnetworksanddeeplearning.com/chap4.html
  - D. Mackay book http://www.inference.phy.cam.ac.uk/mackay/itprnn/ps/482.491.pdf
- But issue is efficiency: very wide two layers vs narrow deep model? In practice, more layers helps.

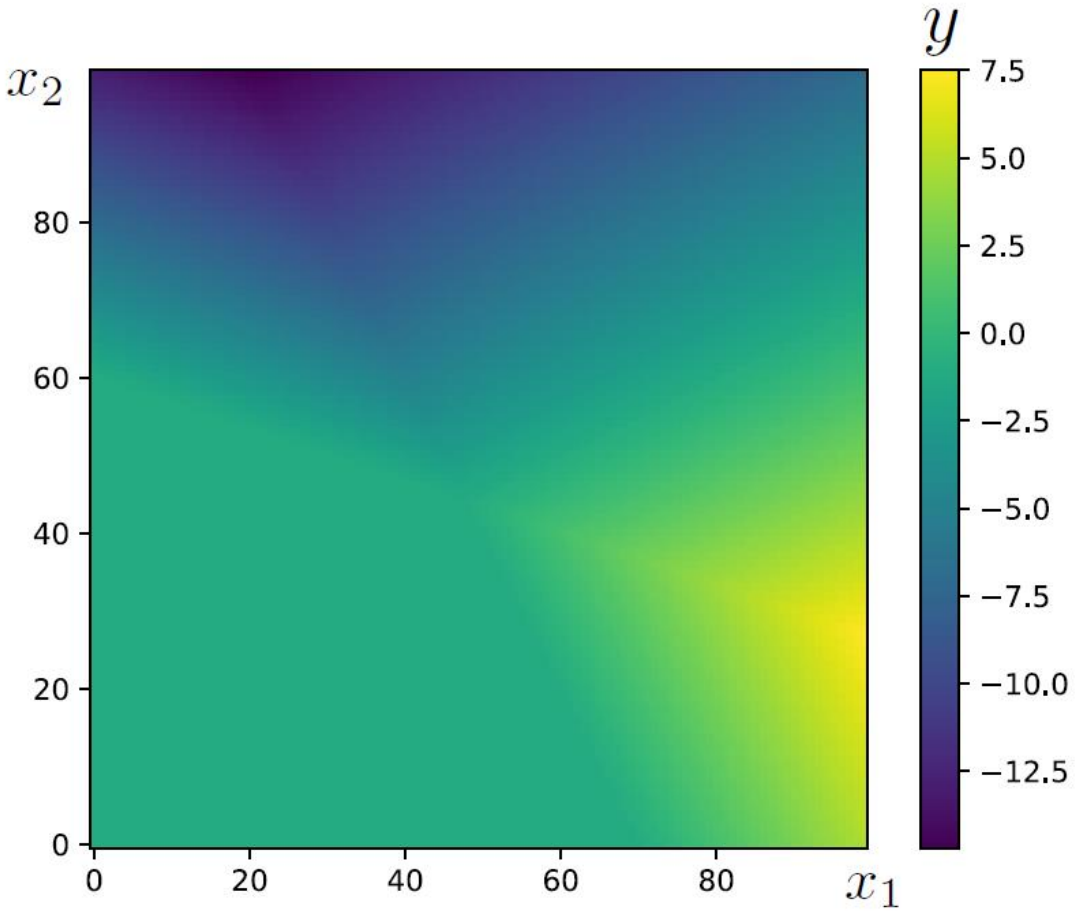
Example: linear classification with a perceptron



$$y = \mathbf{x}^T \mathbf{w} + b$$

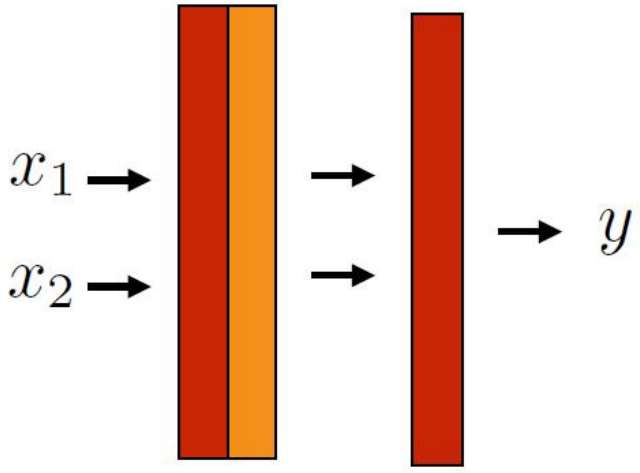


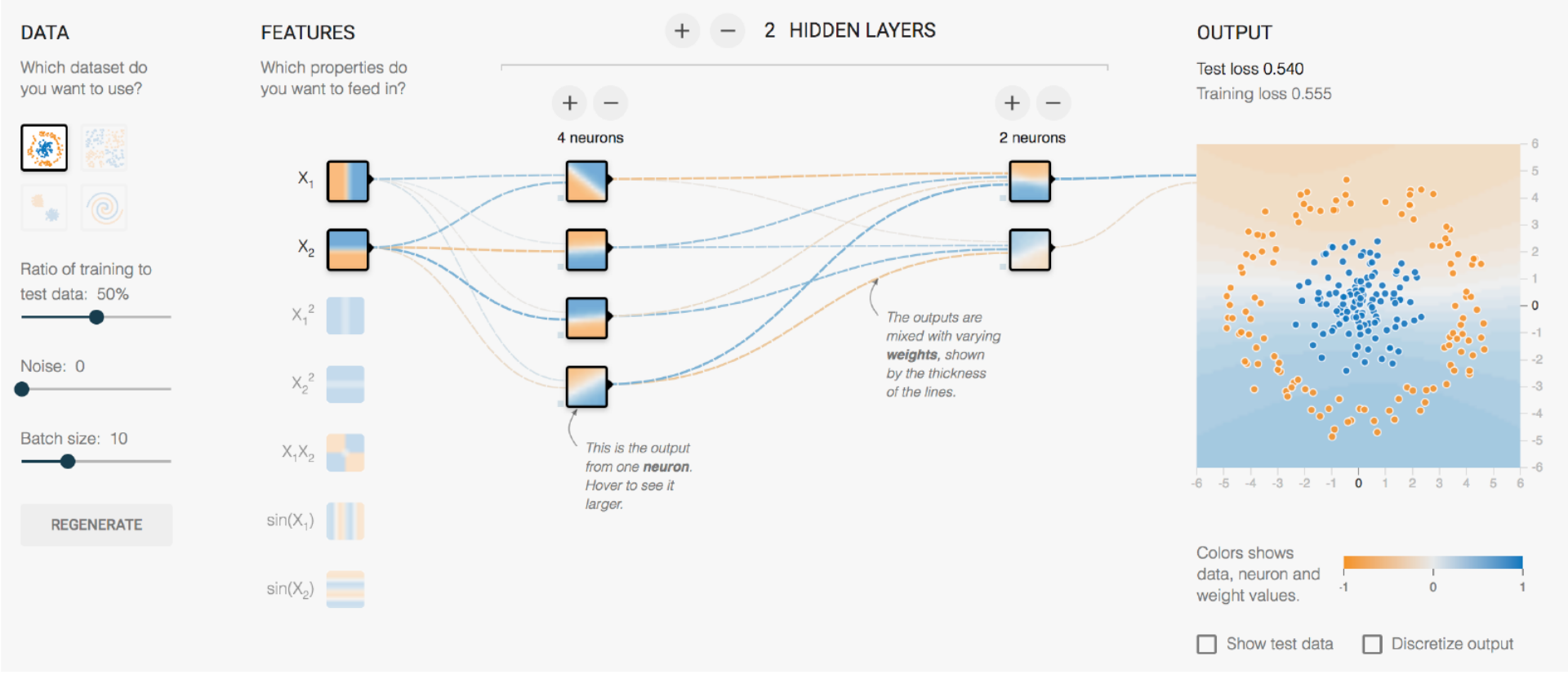
Example: nonlinear classification with a deep net



$$\mathbf{h} = g(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)})$$

$$y = \mathbf{W}^{(2)}\mathbf{h} + b^{(2)}$$

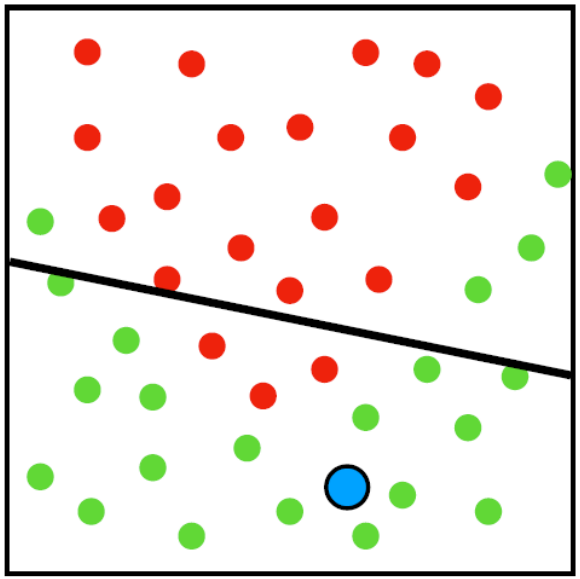




[http://playground.tensorflow.org]

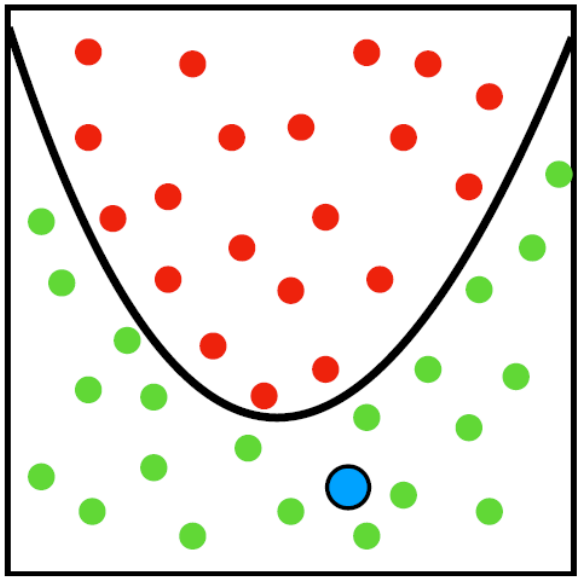
Example: nonlinear classification with a deep net

What class is ● ?



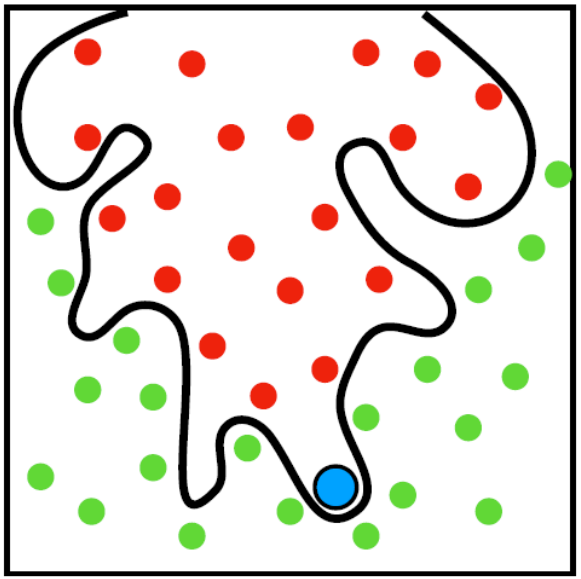
Answer: ●

Underfitting



Answer: ●

Appropriate model



Answer: ●

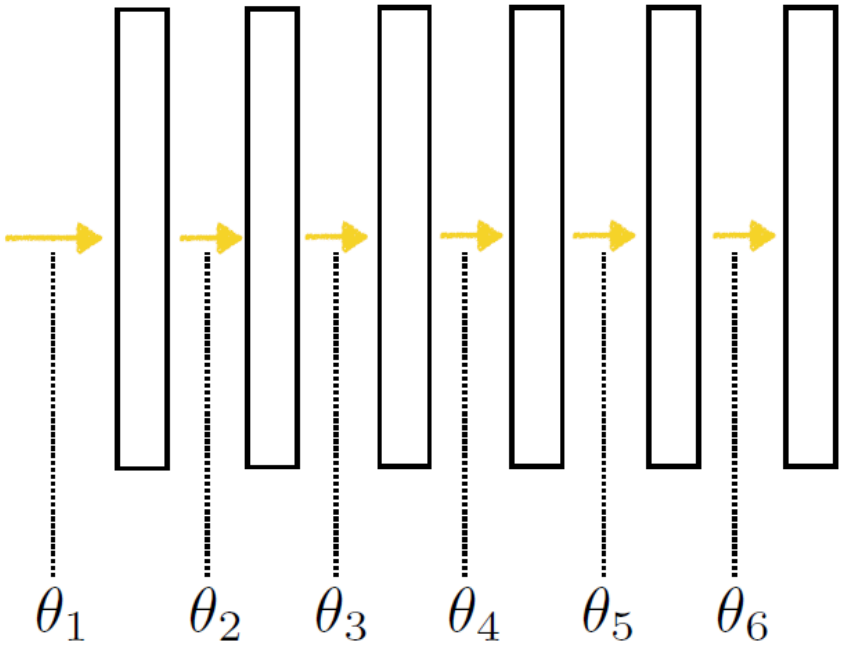
Overfitting

Deep learning

Learned

$\mathbf{y}_1$
"clown fish"

$\mathbf{x}_1$



$$\theta^* = \arg\min_\theta \sum_{i=1}^{N} \mathcal{L}(f_\theta(\mathbf{x}_i), \mathbf{y}_i)$$

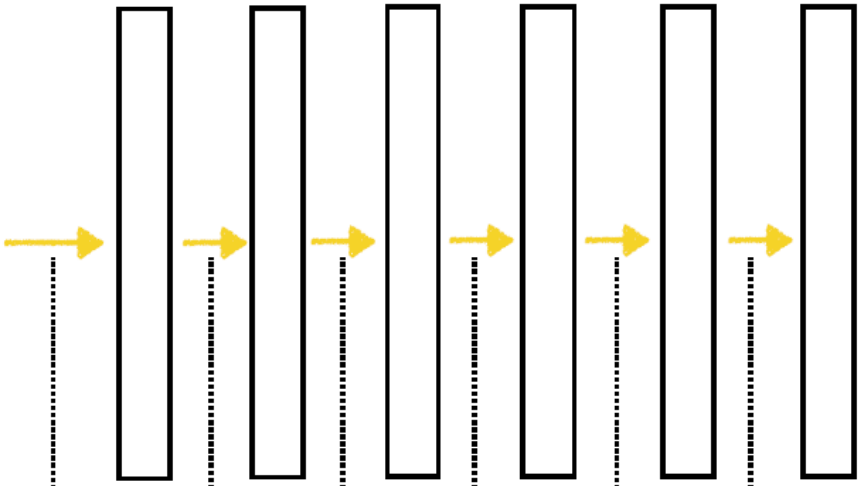
Deep learning

Learned

$\mathbf{y}_2$

"grizzly bear"

$\mathbf{x}_2$



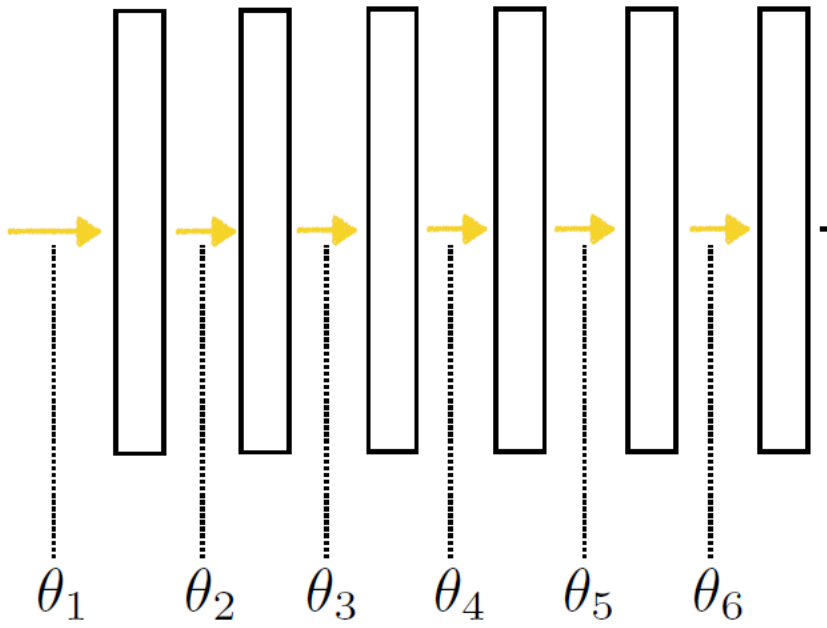
$$\theta^* = \arg\min_\theta \sum_{i=1}^{N} \mathcal{L}(f_\theta(\mathbf{x}_i), \mathbf{y}_i)$$

Deep learning

Learned

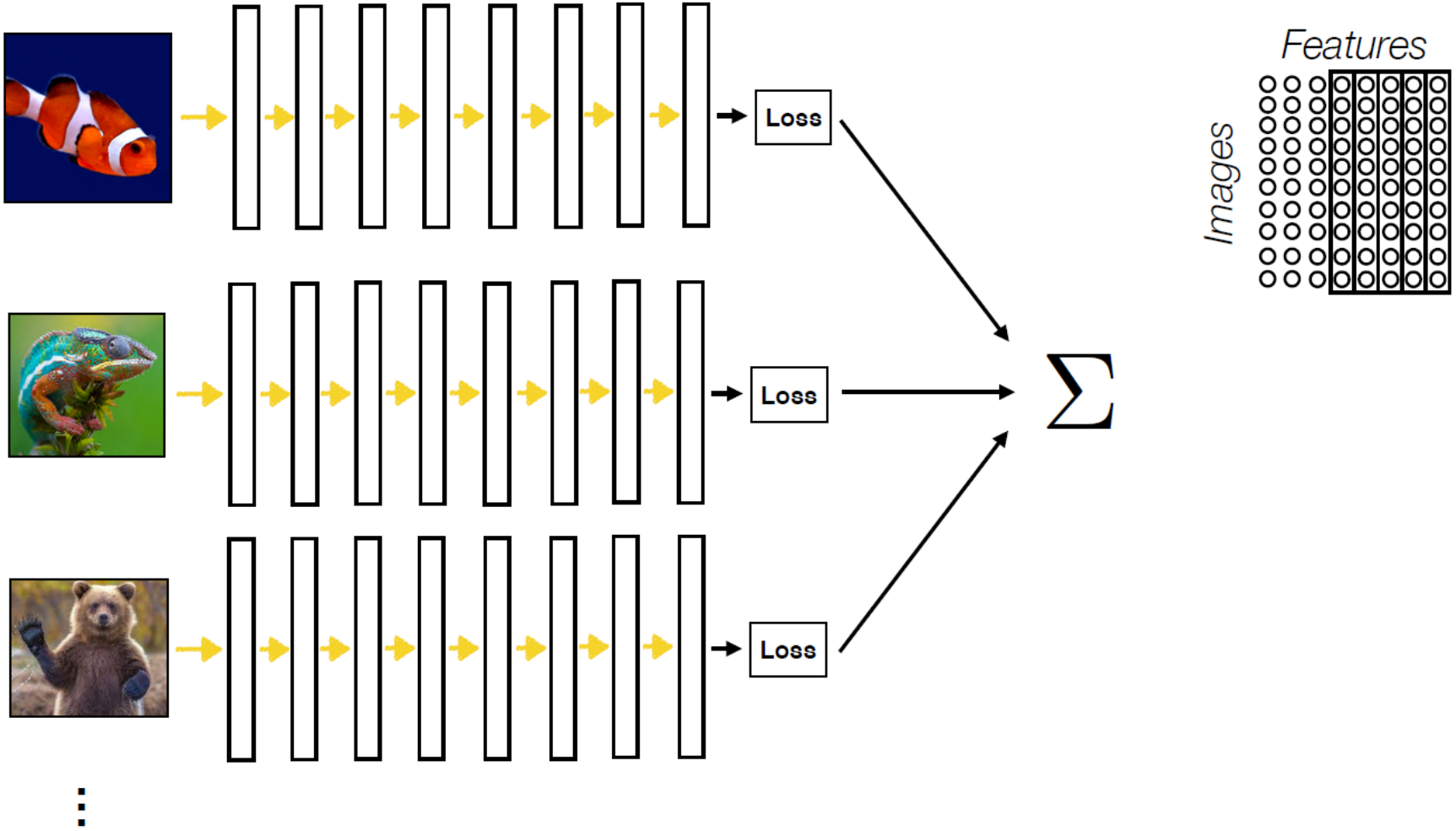
$\mathbf{y}_i$
"chameleon"

$\mathbf{x}_i$

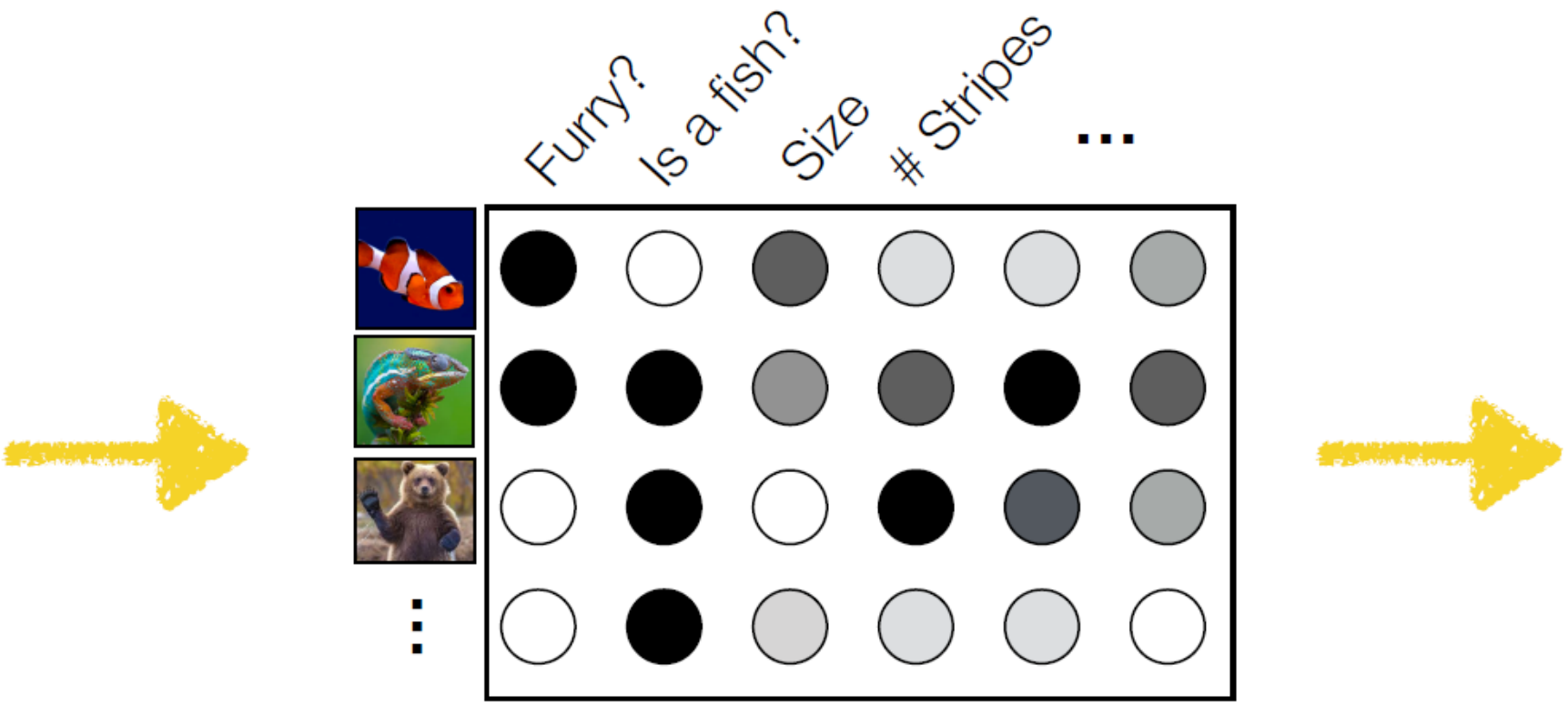


$$\theta^* = \arg\min_\theta \sum_{i=1}^{N} \mathcal{L}(f_\theta(\mathbf{x}_i), \mathbf{y}_i)$$

Batch (parallel) processing



Tensors (multi-dimensional arrays)

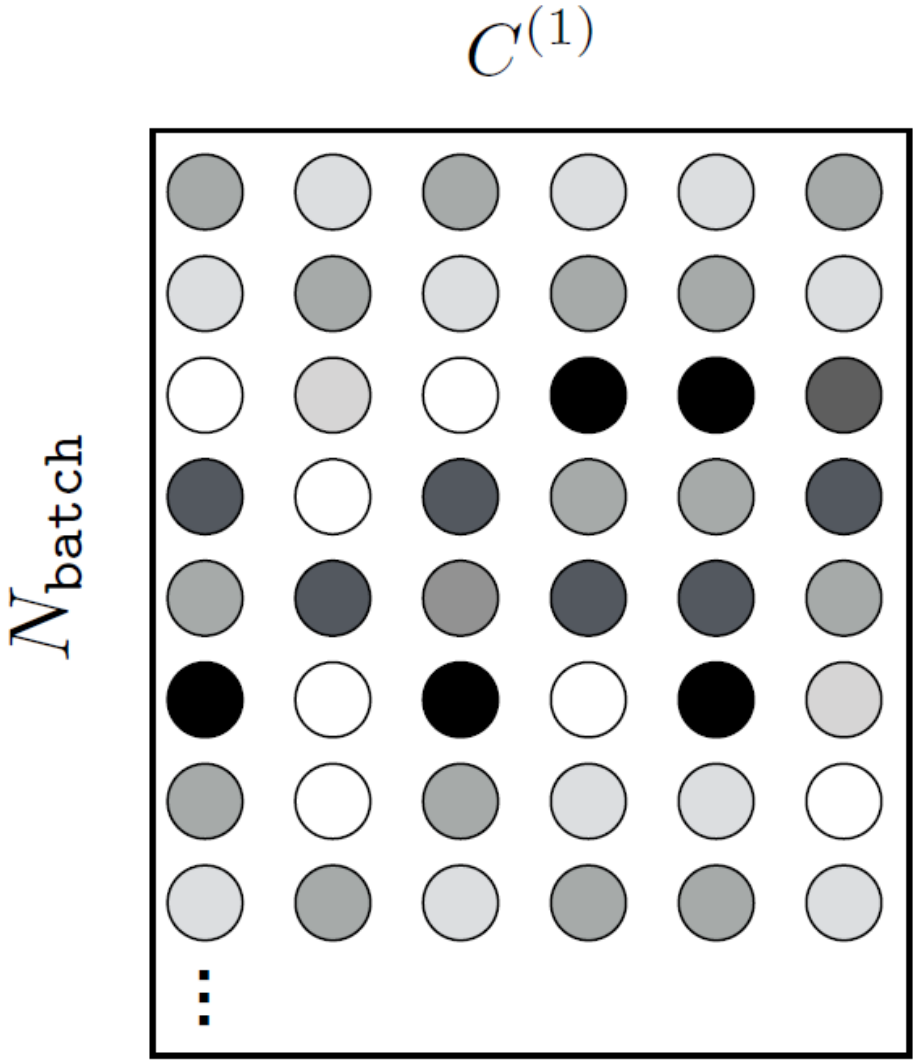


Each layer is a representation of the data

Tensors (multi-dimensional arrays)

$$\mathbf{h}^{(1)} \in \mathbb{R}^{N_{\texttt{batch}} \times C^{(1)}}$$

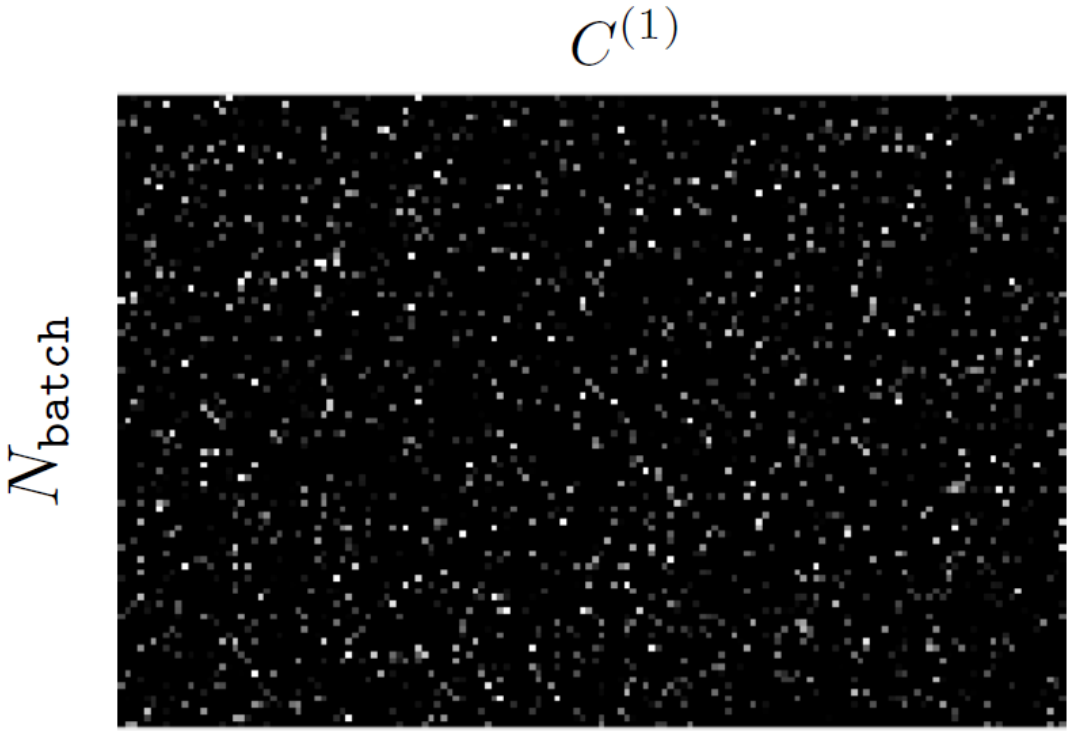
neurons
features
units
"channels"



Tensors (multi-dimensional arrays)

$$\mathbf{h}^{(1)} \in \mathbb{R}^{N_{\texttt{batch}} \times C^{(1)}}$$

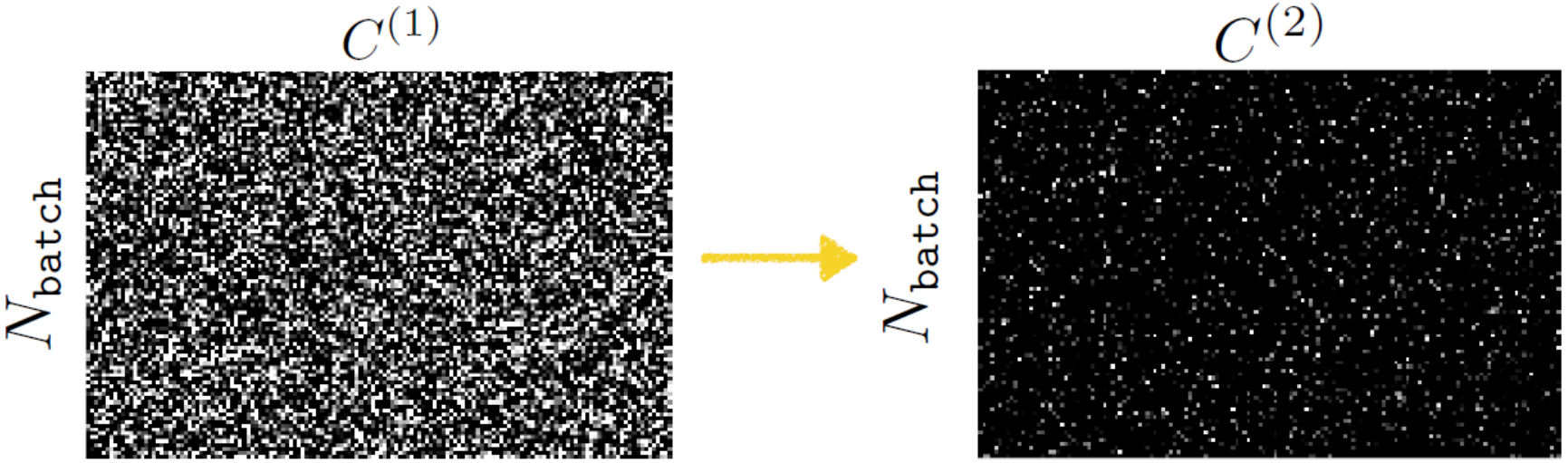
neurons
features
units
"channels"

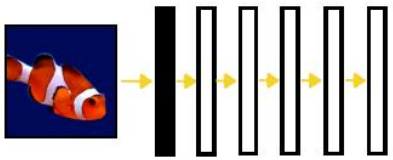


“Tensor flow”

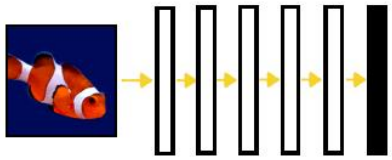
$$\mathbf{h}^{(1)} \in \mathbb{R}^{N_{\text{batch}} \times C^{(1)}}$$

$$\mathbf{h}^{(2)} \in \mathbb{R}^{N_{\text{batch}} \times C^{(2)}}$$

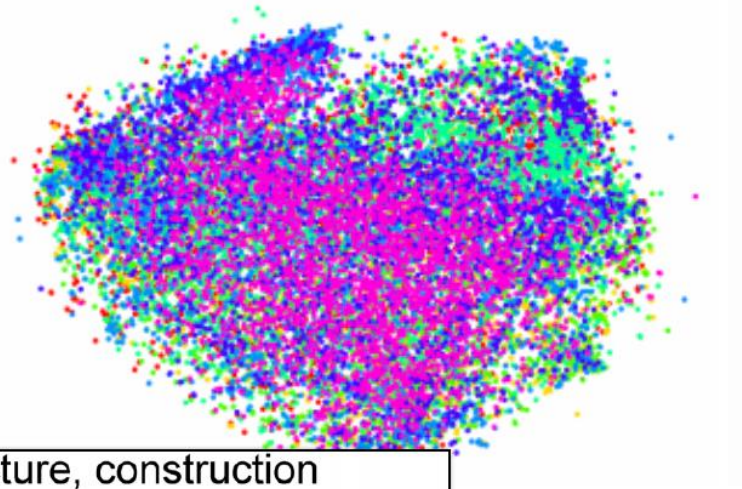


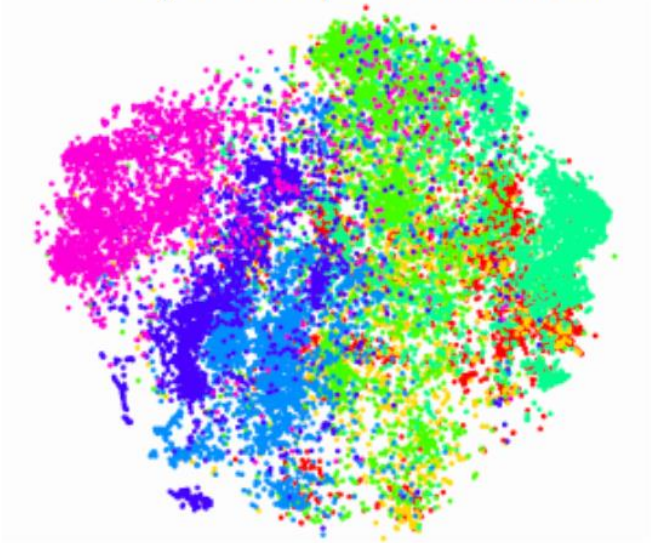


Layer 1 representation



Layer 6 representation





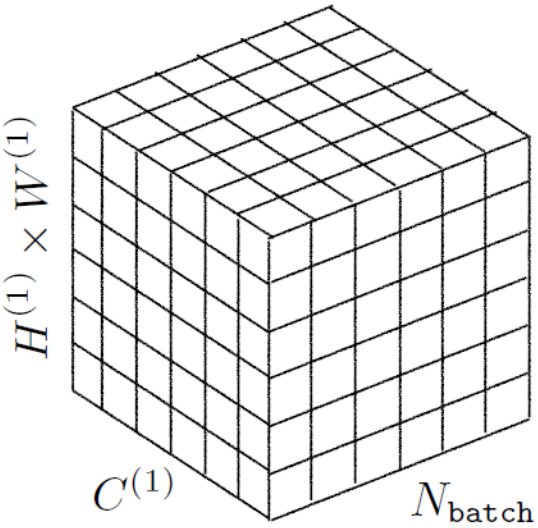
- structure, construction
- covering
- commodity, trade good, good
- conveyance, transport
- invertebrate
- bird
- hunting dog

[DeCAF, Donahue, Jia, et al. 2013]

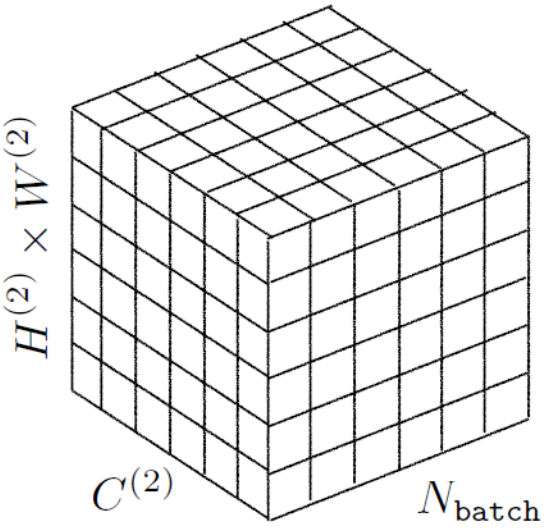
[Visualization technique : t-sne, van der Maaten & Hinton, 2008]

“Tensor flow”

$$\mathbf{h}^{(1)} \in \mathbb{R}^{N_{\text{batch}} \times H^{(1)} \times W^{(1)} \times C^{(1)}}$$



$$\mathbf{h}^{(2)} \in \mathbb{R}^{N_{\text{batch}} \times H^{(2)} \times W^{(2)} \times C^{(2)}}$$



Regularization

Regularizing deep nets

Deep nets have millions of parameters!

On many datasets, it is easy to overfit — we may have more free parameters than data points to constrain them.

How can we regularize to prevent the network from overfitting?

1. Fewer neurons, fewer layers
2. Weight decay
3. Dropout
4. Normalization layers
5. …

Regularized least squares

$$f_\theta(x) = \sum_{k=0}^{K} \theta_k x^k$$

$$R(\theta) = \lambda \left\|\theta\right\|_2^2$$ ⟵ Only use polynomial terms if you really need them! Most terms should be zero

ridge regression, a.k.a., **Tikhonov regularization**

Probabilistic interpretation: R is a Gaussian **prior** over values of the parameters.

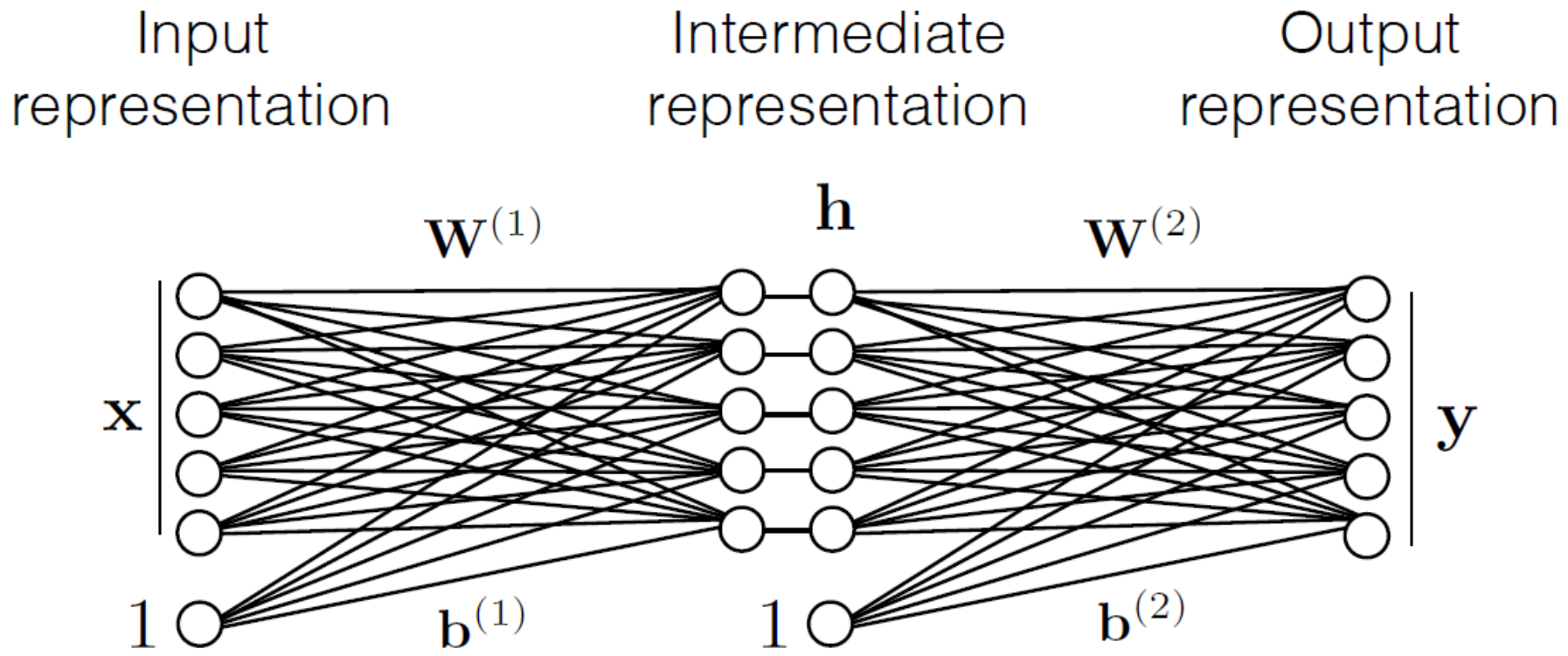
Regularizing the weights in a neural net

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(f_\theta(\mathbf{x}_i), \mathbf{y}_i) + R(\theta)$$

$$R(\mathbf{W}) = \lambda \|\mathbf{W}\|_2^2$$ ← **weight decay**

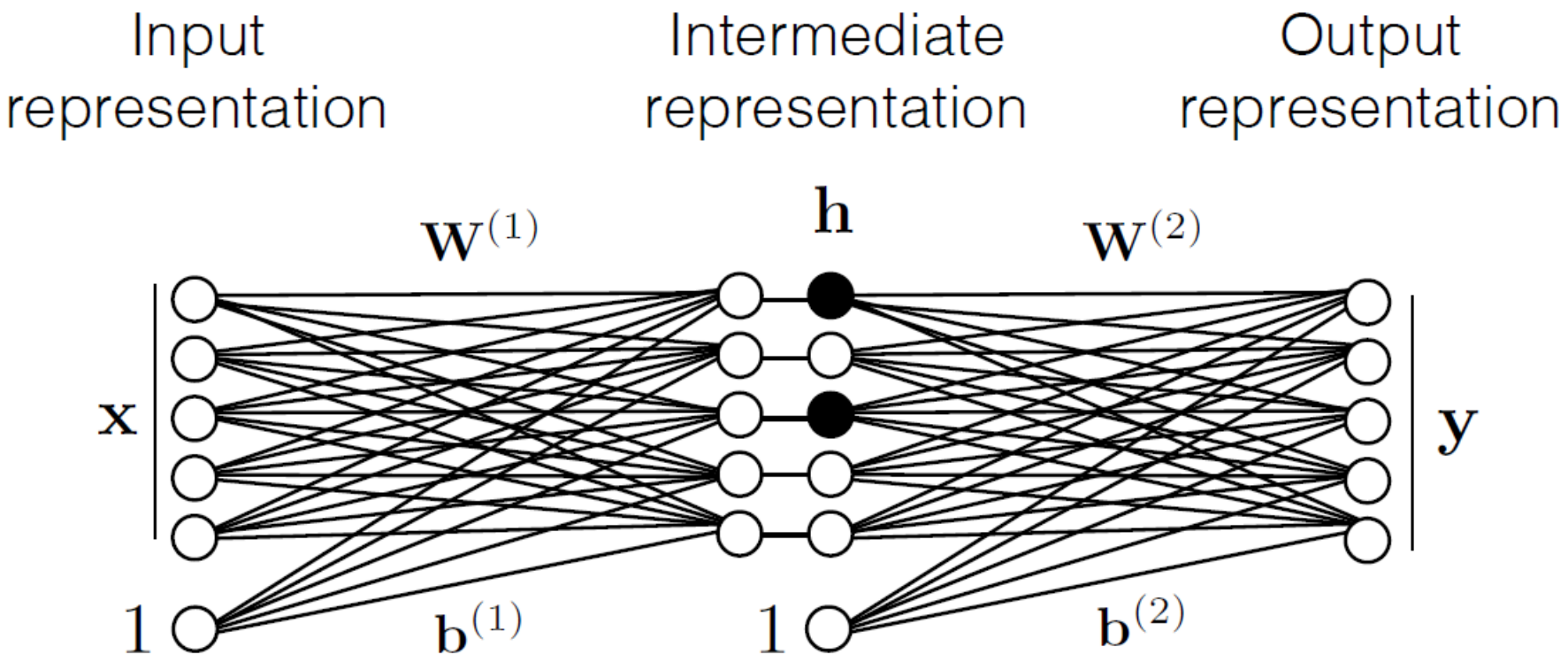
"We prefer to keep weights small."

Dropout



$$\theta = \{\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)}, \mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(L)}\}$$

Dropout



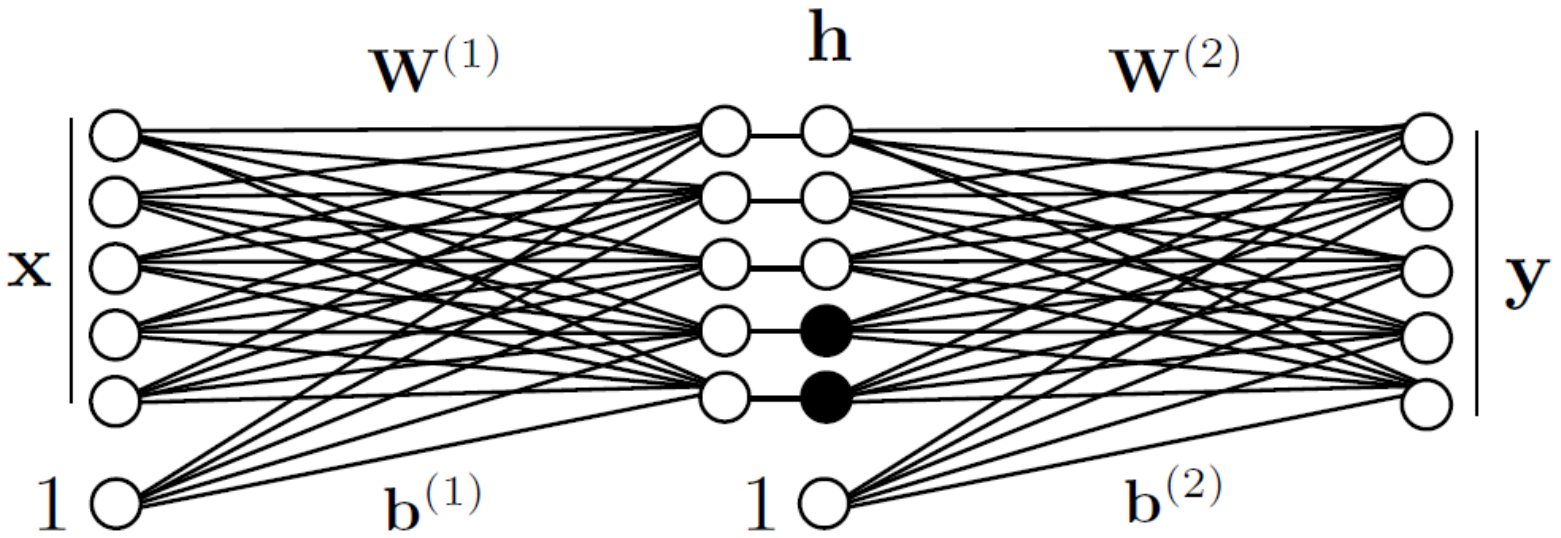
$$\theta = \{\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)}, \mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(L)}\}$$

Dropout

Input representation

Intermediate representation

Output representation



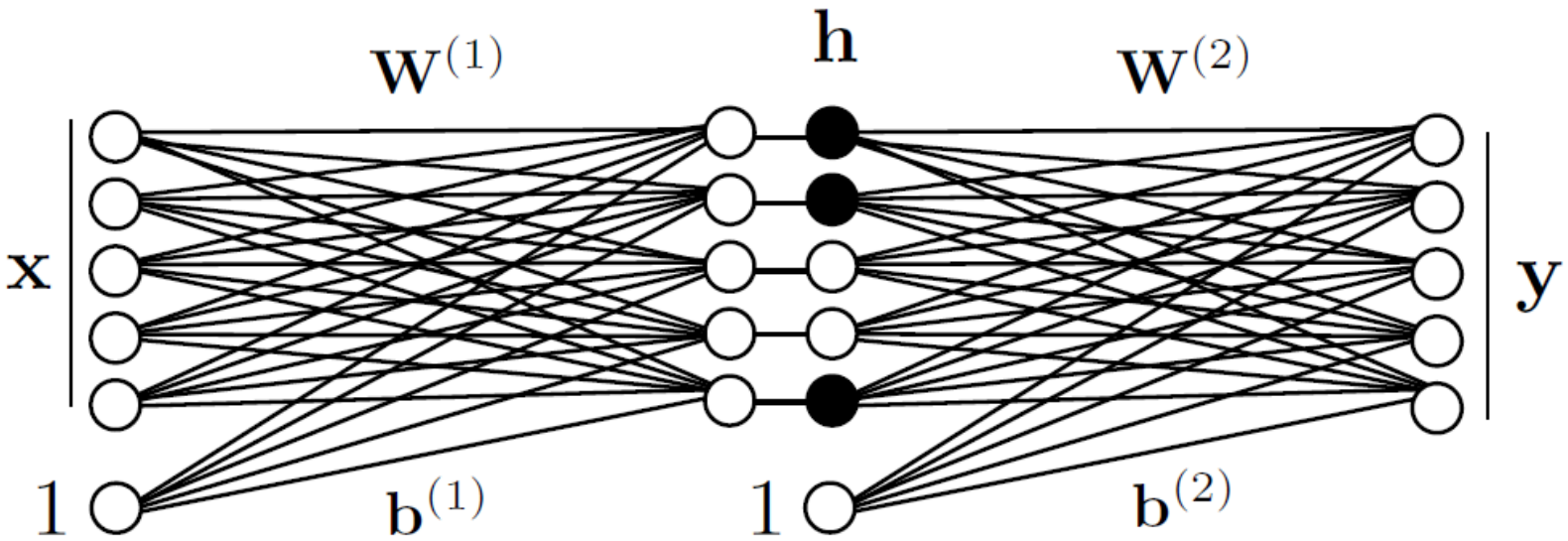
$$\theta = \{\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)}, \mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(L)}\}$$

Dropout

Input representation

Intermediate representation

Output representation



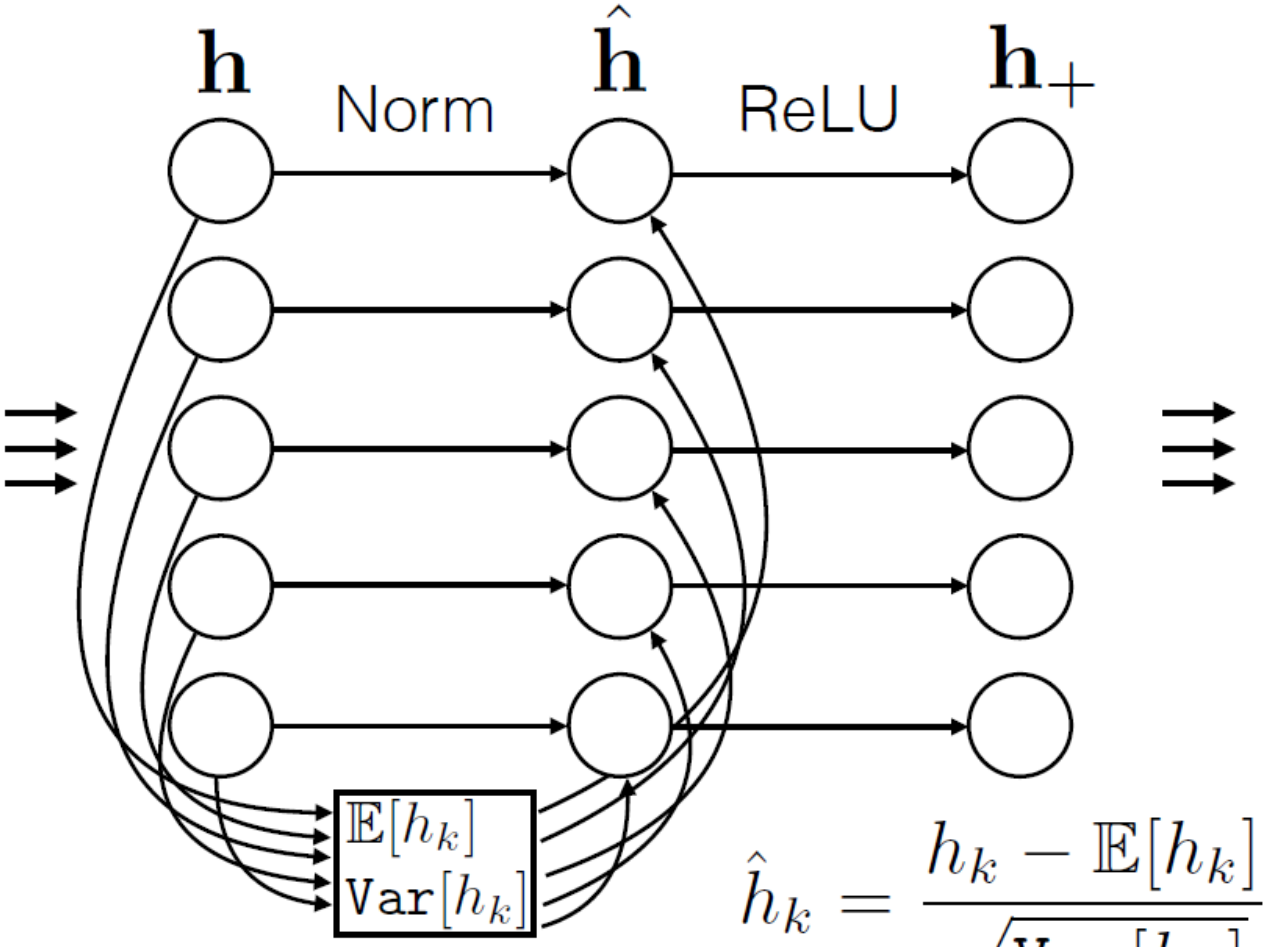
$$\theta = \{\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)}, \mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(L)}\}$$

Dropout

- Randomly zero out hidden units
- Prevents network from relying too much on spurious correlations between different hidden units.
- Can be understood as averaging over an exponential ensemble of subnetworks. This averaging smooths the function, thereby reducing the effective capacity of the network.

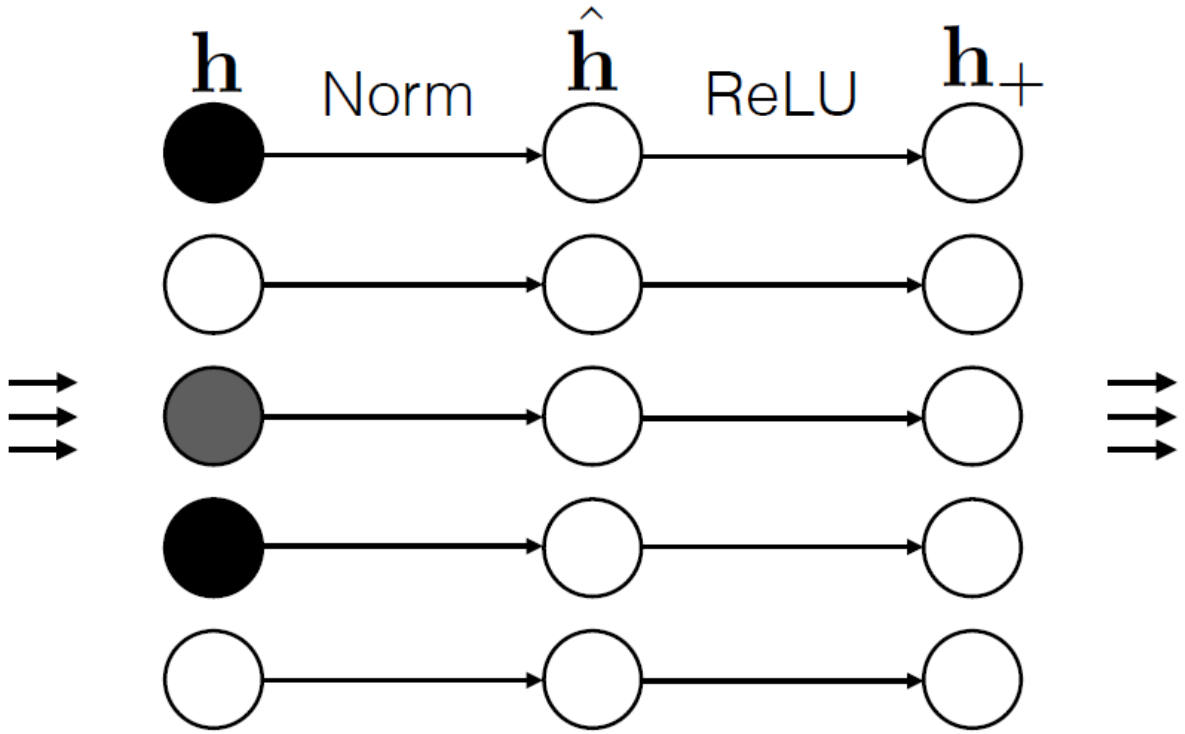
Normalization

Normalization layers



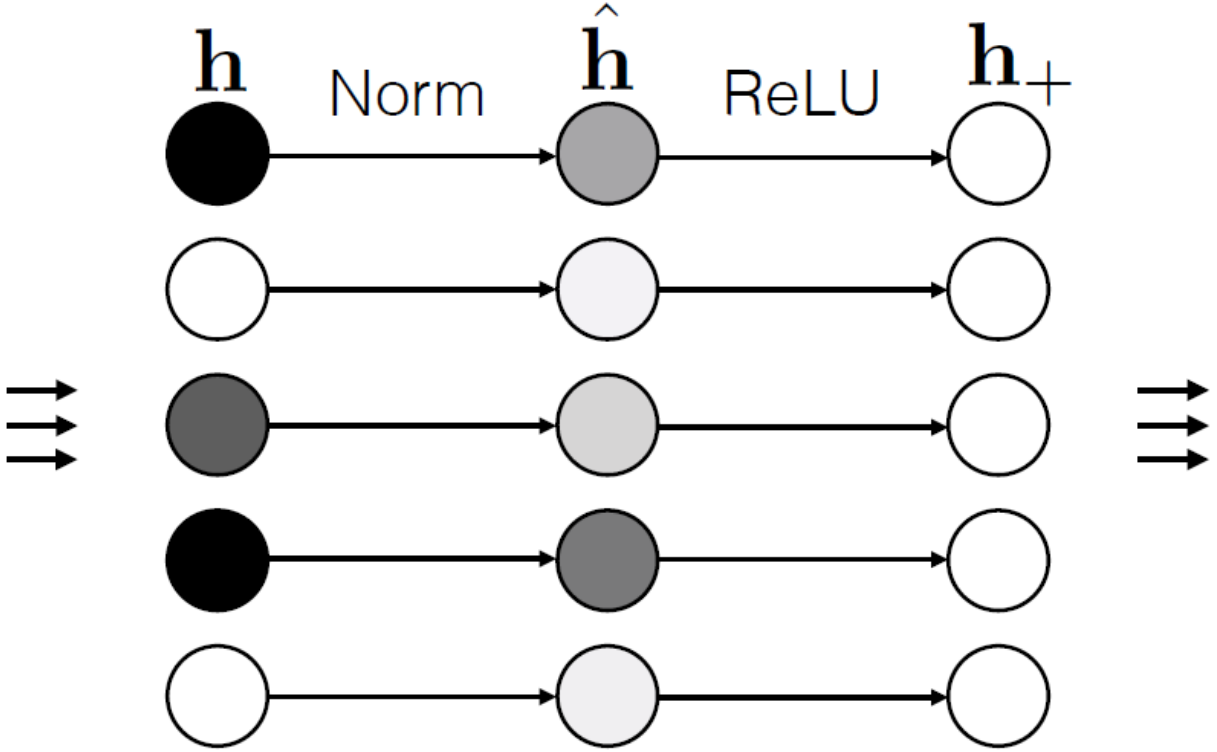
$$\hat{h}_k = \frac{h_k - \mathbb{E}[h_k]}{\sqrt{\mathtt{Var}[h_k]}}$$

Normalization layers



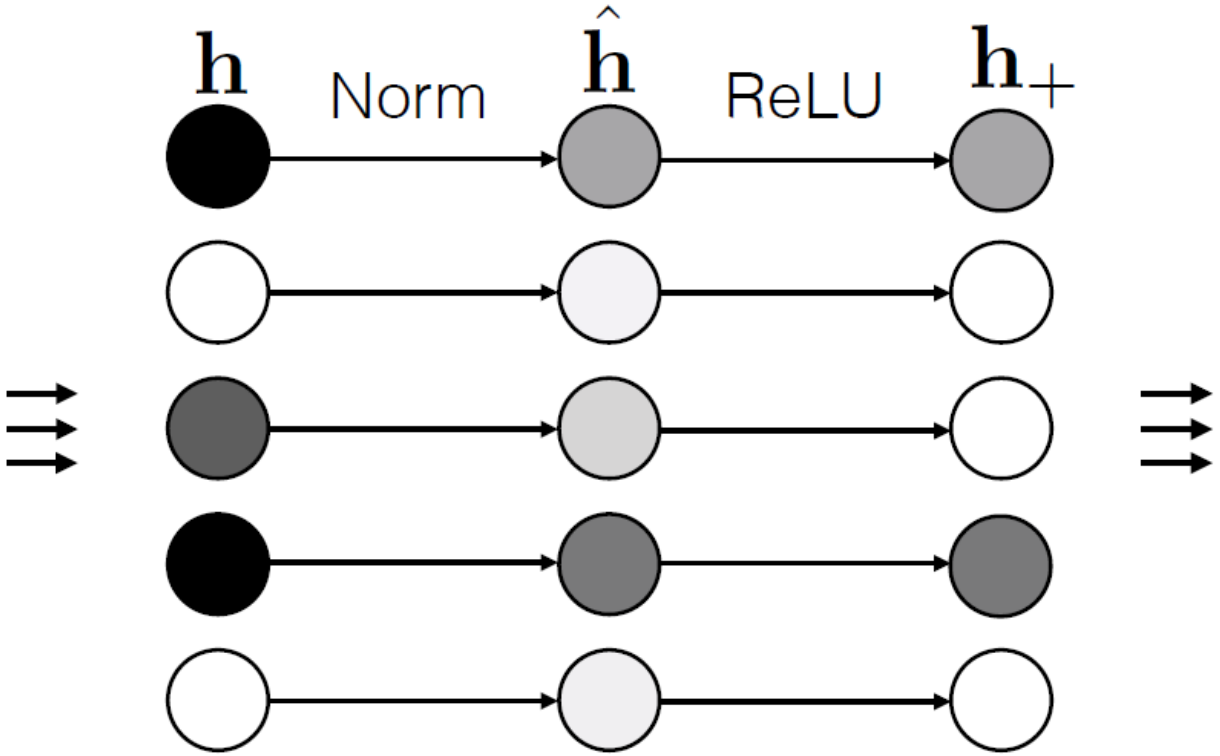
$$\hat{h}_k = \frac{h_k - \mathbb{E}[h_k]}{\sqrt{\mathtt{Var}[h_k]}}$$

Normalization layers



$$\hat{h}_k = \frac{h_k - \mathbb{E}[h_k]}{\sqrt{\mathtt{Var}[h_k]}}$$

Normalization layers

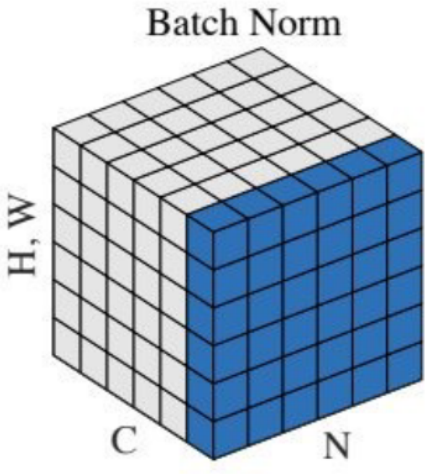


$$\hat{h}_k = \frac{h_k - \mathbb{E}[h_k]}{\sqrt{\texttt{Var}[h_k]}}$$

Normalization layers

- Keep track of mean and variance of a unit (or a population of units) over time.
- Standardize unit activations by subtracting mean and dividing by variance.
- Squashes units into a standard range, avoiding overflow.
- Also achieves invariance to mean and variance of the training signal.
- Both these properties reduce the effective capacity of the model, i.e. regularize the model.

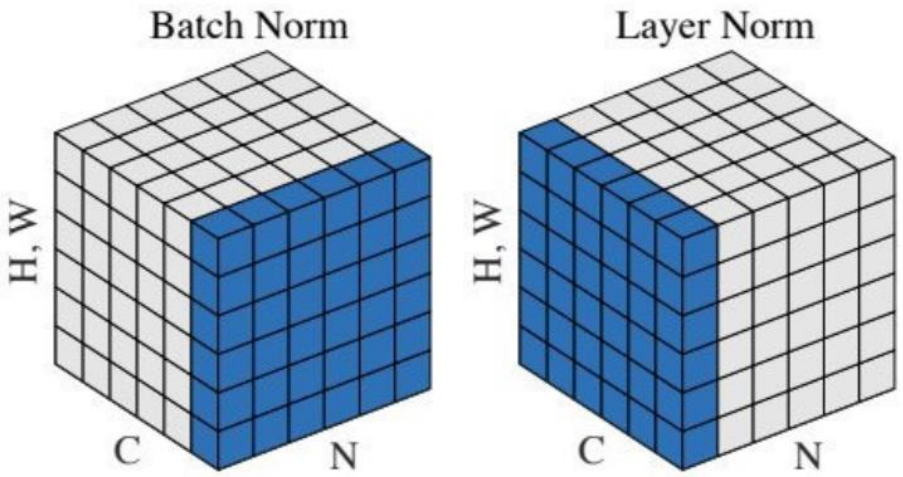
Normalization layers

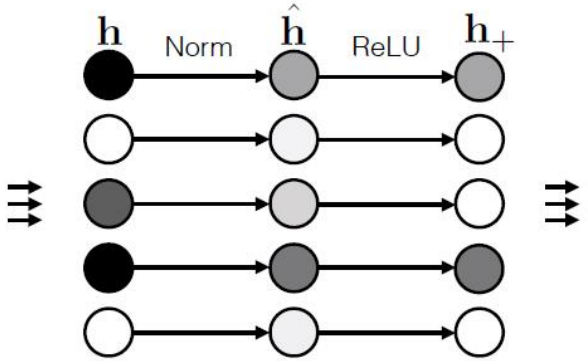


Normalize w.r.t. a single hidden unit's pattern of activation over training examples (a batch of examples).

[Figure from Wu & He, arXiv 2018]

Normalization layers

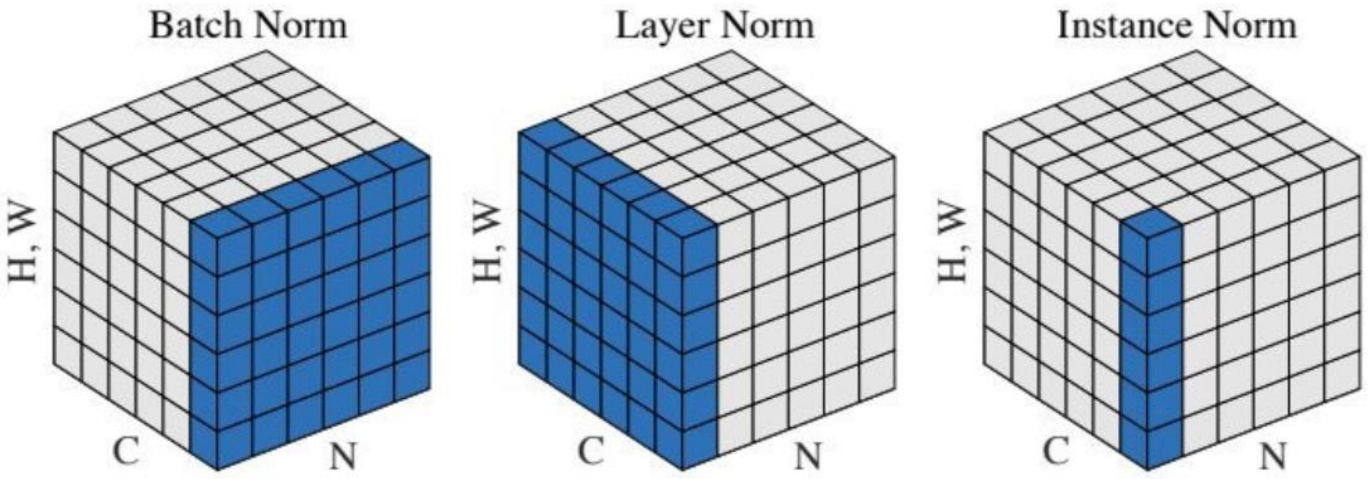




Normalize w.r.t. the mean and variance of the activations of all the hidden units (neurons) on this layer (c).

[Figure from Wu & He, arXiv 2018]

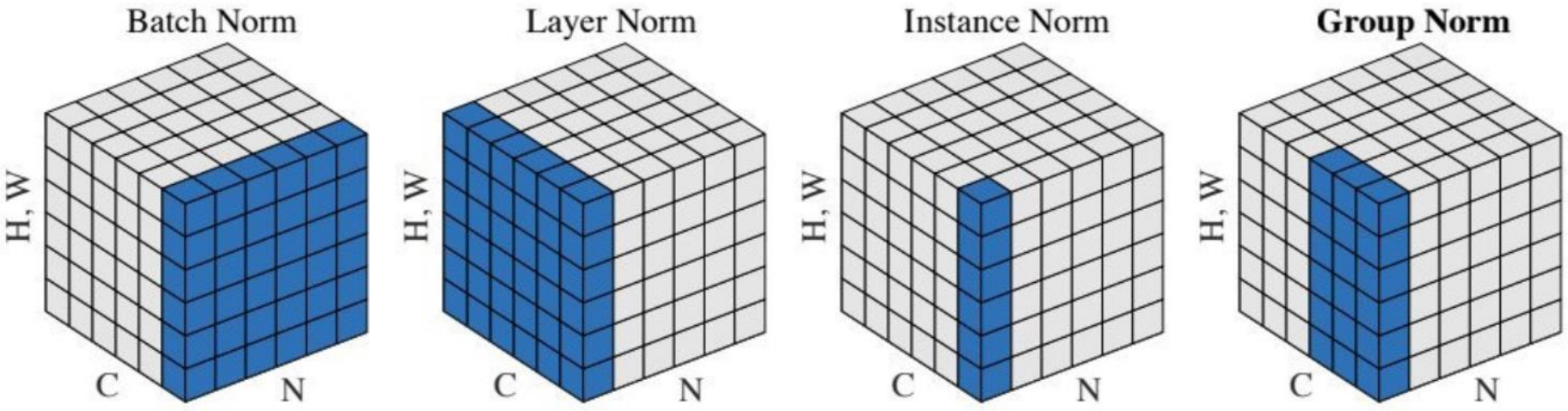
Normalization layers



Normalize w.r.t. the mean and variance of the activations of all the hidden units (neurons) on this layer (c) that process this **particular example.**

[Figure from Wu & He, arXiv 2018]

Normalization layers

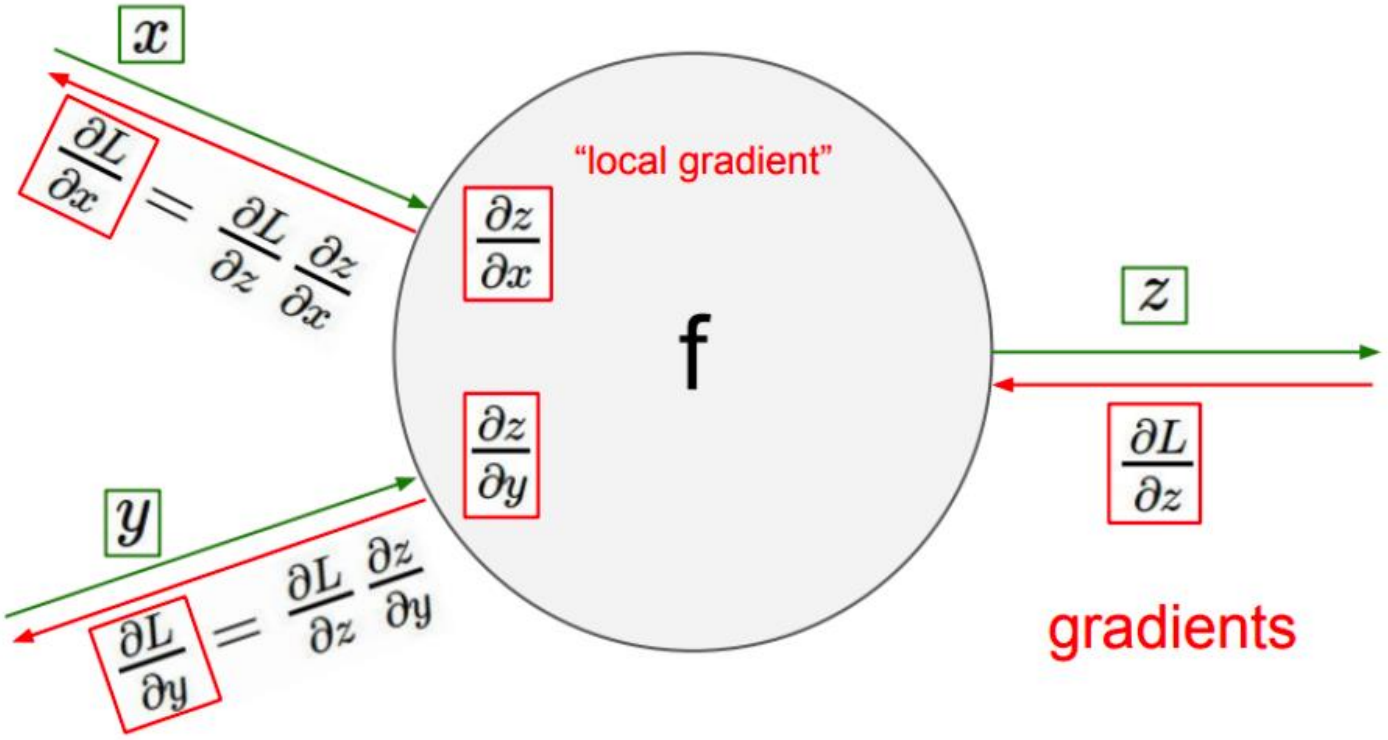


Might as well…

[Figure from Wu & He, arXiv 2018]

Next up:

Optimization of NNs (backprop)



Taxonomy of Learning Problems

Supervised learning: learning from examples

Training data

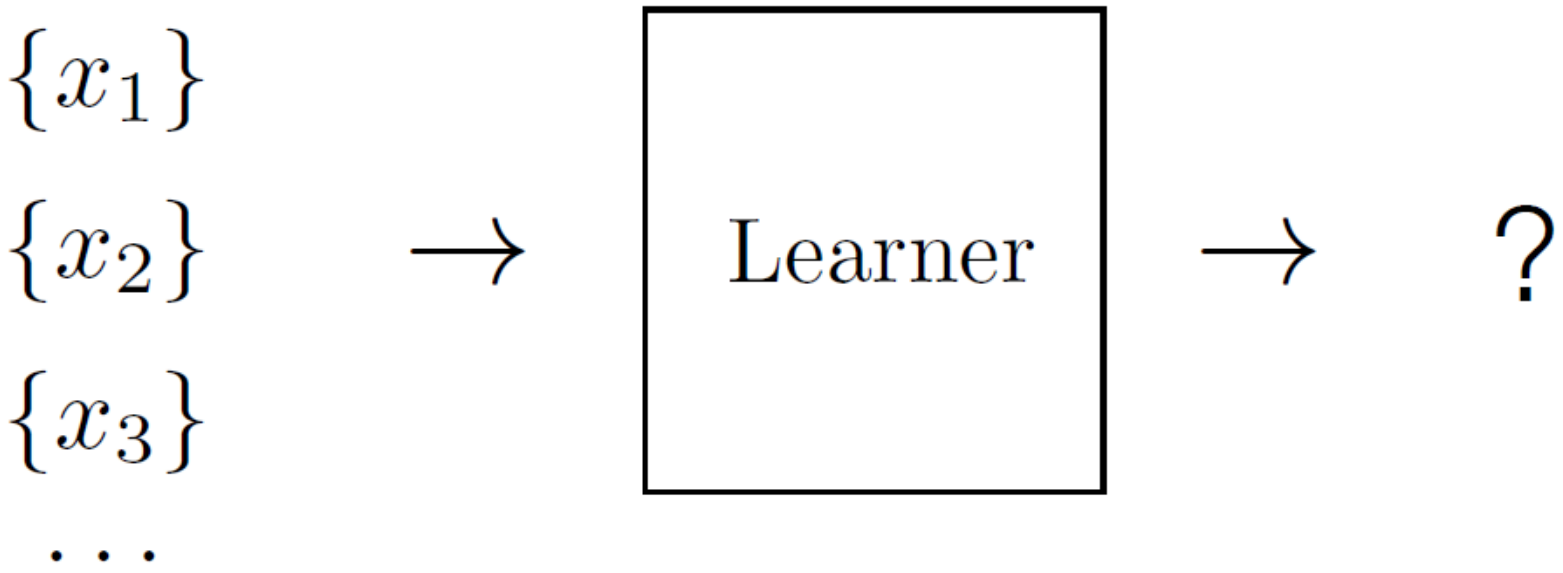
$\{x_1, y_1\}$
$\{x_2, y_2\}$
$\{x_3, y_3\}$
$\ldots$

$\rightarrow$ Learner $\rightarrow$ $f : X \rightarrow Y$

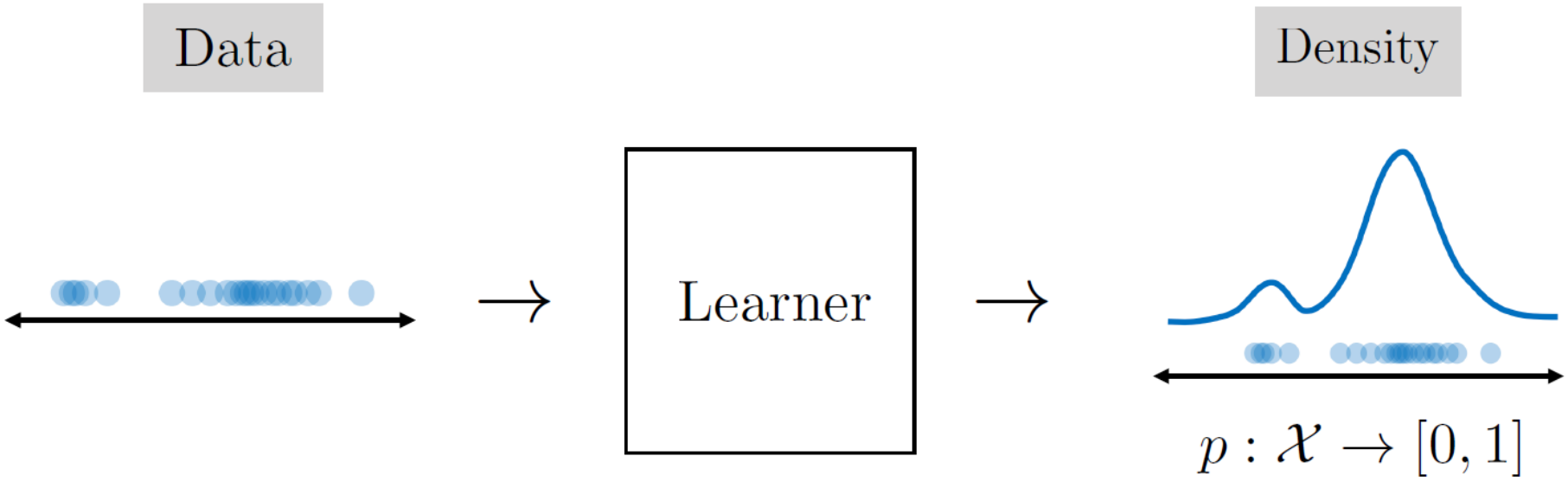
$$f^* = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{N} \mathcal{L}(f(x_i), y_i)$$

Unsupervised learning: Learning without examples

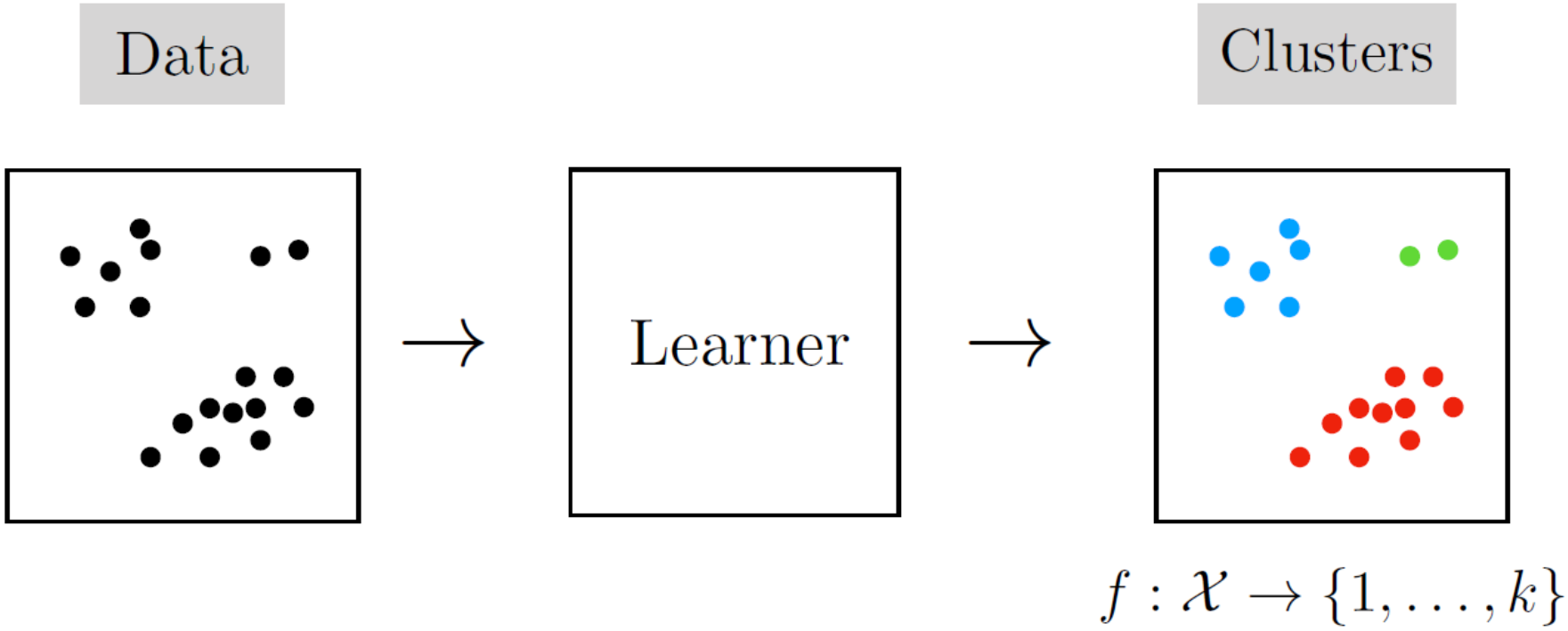
Data



Density modeling

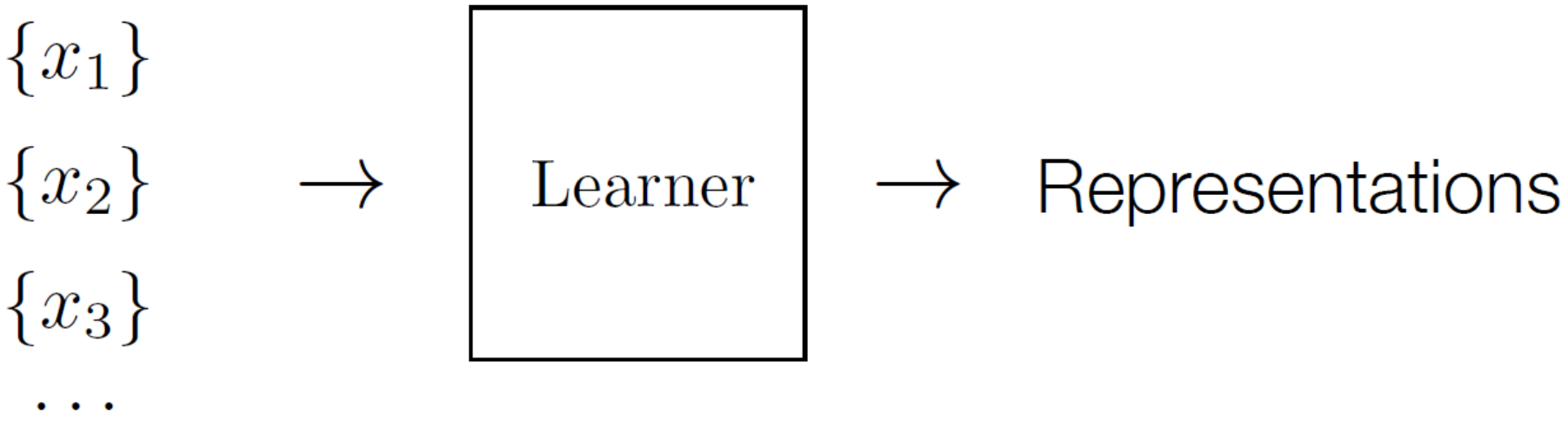


Clustering



Representation Learning

Data

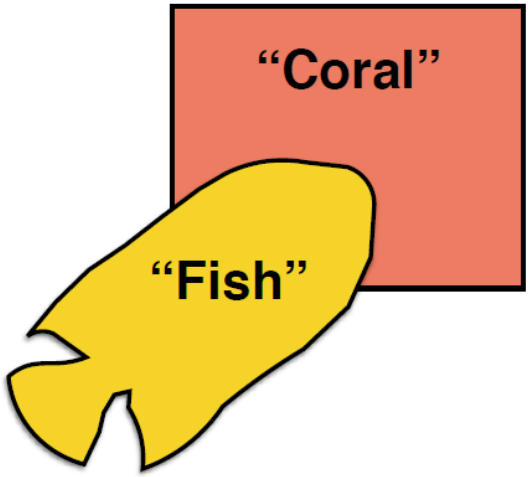


Representation Learning

$\mathbf{x}$

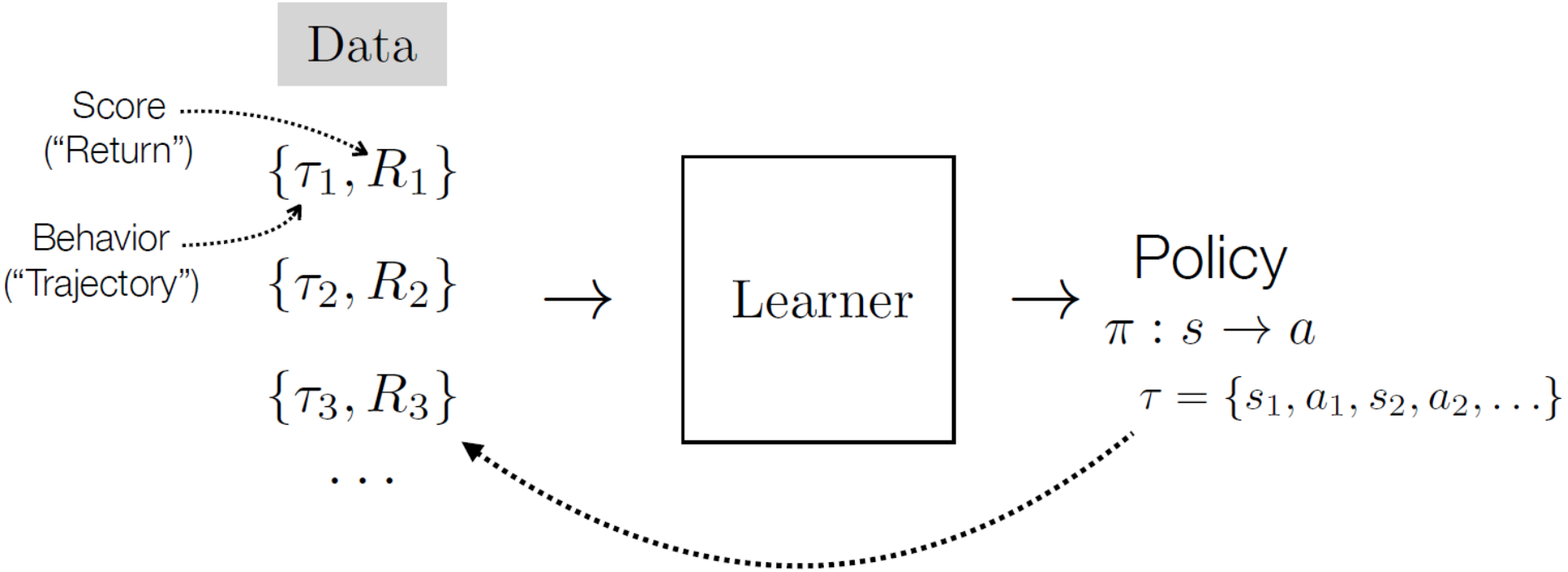
Image

$\rightarrow$ f $\rightarrow$



Compact mental representation

Reinforcement learning



Differentiable programing

Deep nets are popular for a few reasons:

1. High capacity
2. Easy to optimize (differentiable)
3. Compositional “block based programming”

An emerging term for general models with these properties is **differentiable programming**.