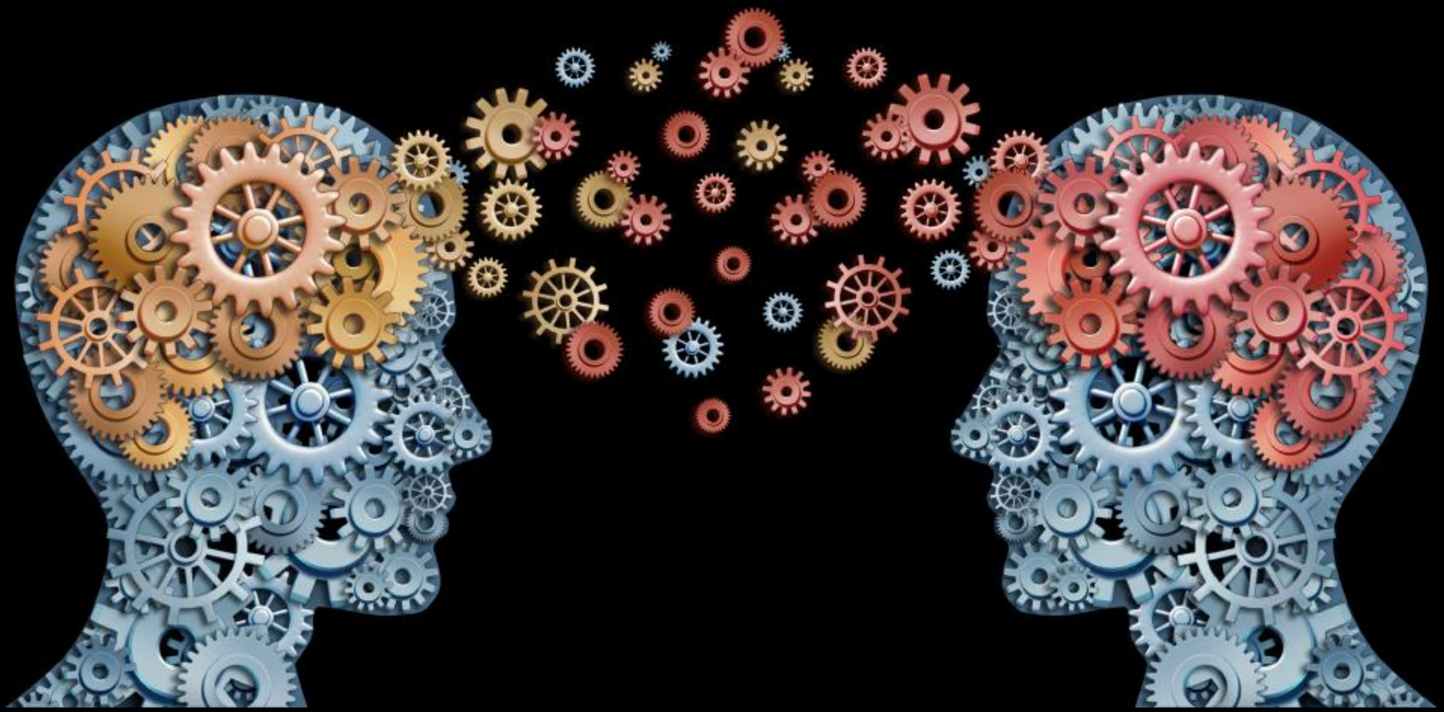
MAI Deep Learning

Transfer learning



Armand Vilalta

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Transfer Learning

The idea

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Your first driving lesson

Your first driving lesson

- Training a deep neural network typically requires **thousands** if not **millions** of training **examples**.
- **Each example** is passed through the network, an **output** is obtained, and, if the prediction is **wrong**, the weights are **adjusted.**

Your first driving lesson

- Training a deep neural network typically requires **thousands** if not **millions** of training **examples**.
- **Each example** is passed through the network, an **output** is obtained, and, if the prediction is **wrong**, the weights are **adjusted.**
- Imagine **learning to drive** a car this way…

Your first driving lesson

Learning like a Deep Neural Network

Your first driving lesson

- Training a deep neural network typically requires **thousands** if not **millions** of training examples.
- **Each example** is passed through the network, an **output** is obtained, and, if the prediction is **wrong**, the weights are **adjusted.**
- Imagine learning to drive a car this way…

We naturally reuse what we previously learnt to be able to solve a new task.

Why do we need **transfer** learning?

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Image classification

- **1998 LeNet-5**

  Gradient-based learning applied to document recognition.

  Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner

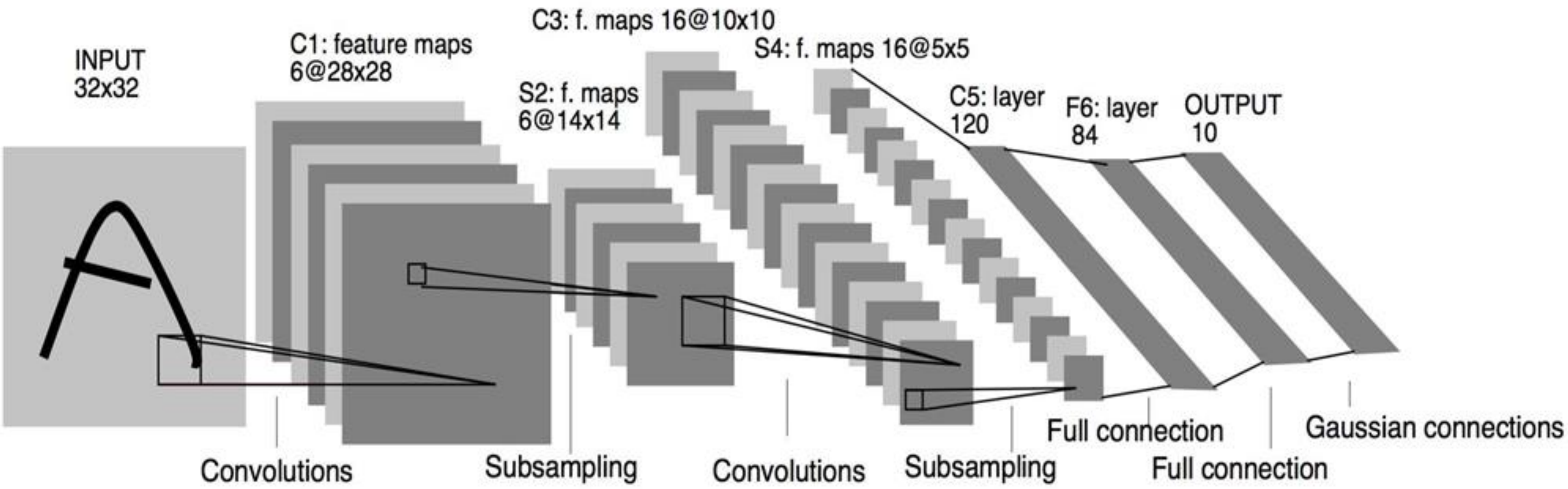
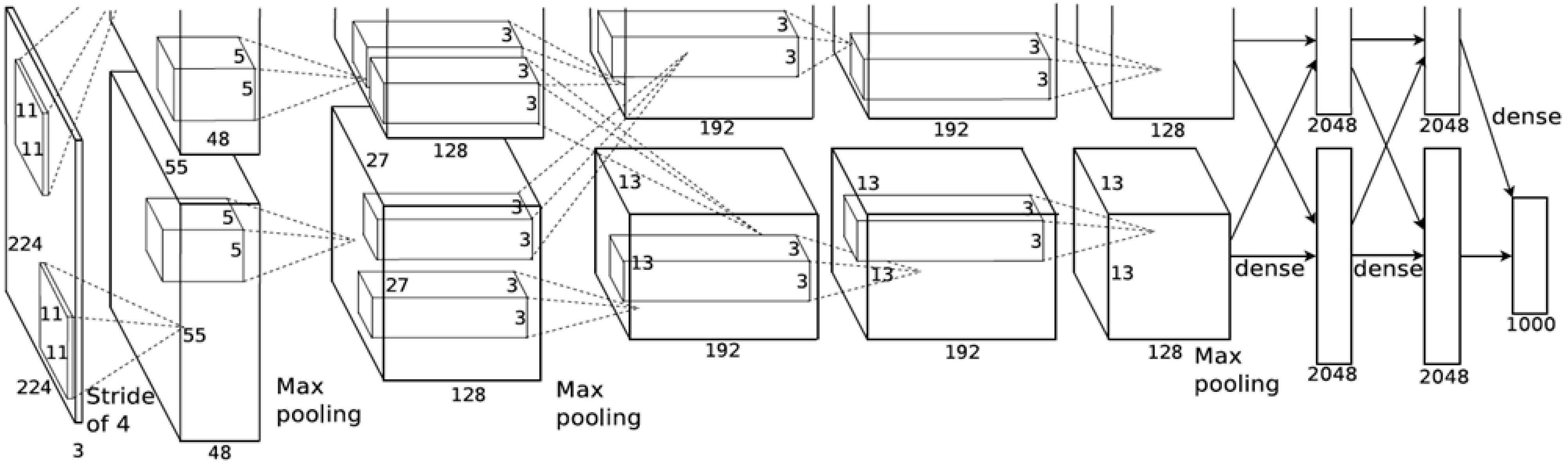


Image classification

- **2012 AlexNet**

ImageNet Classification with Deep Convolutional Neural Networks
Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton



1998 LeNet-5

2012 AlexNet

2014 VGG19

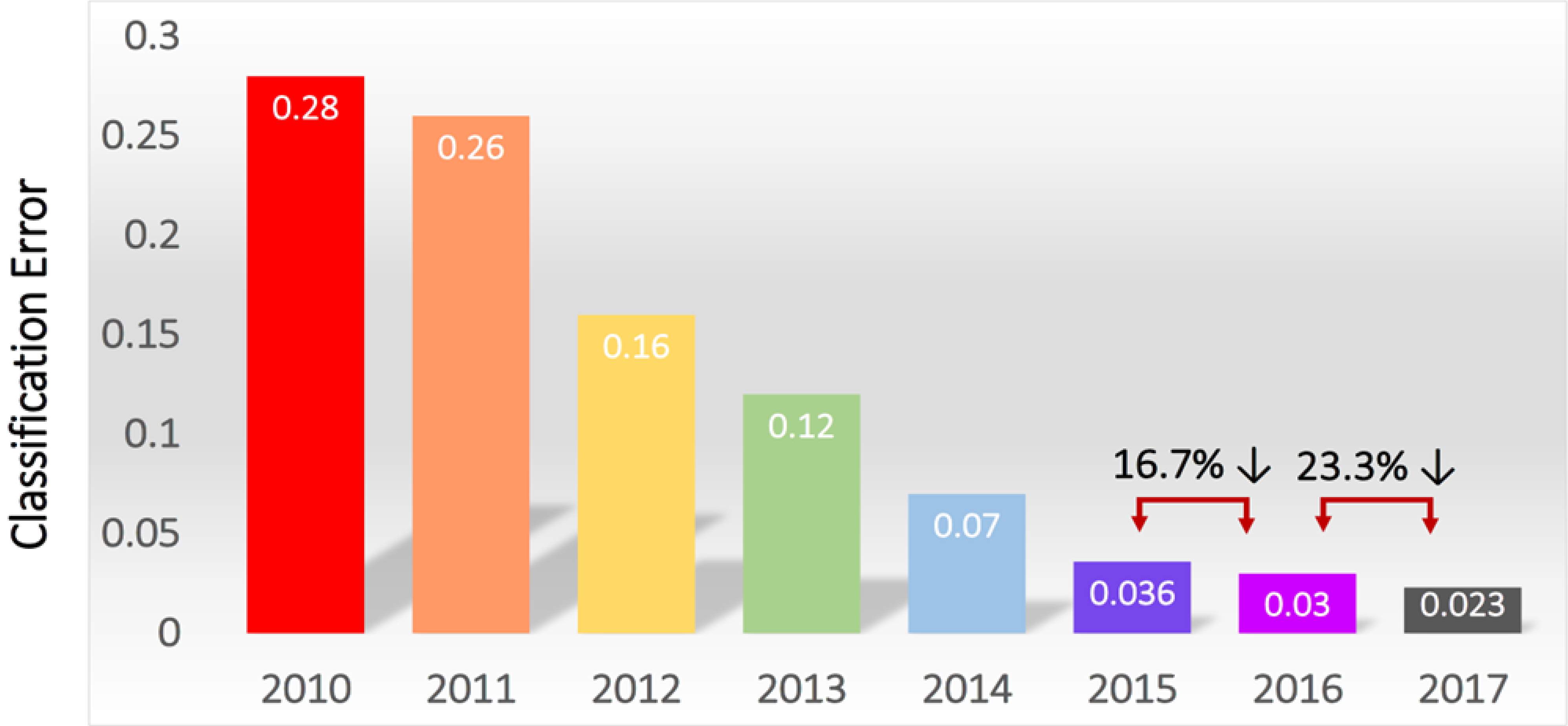
2014 GoogLeNet

2015 Inception-V3

2015 ResNet-56

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ImageNet classification results



From image-net.org

At what price?

- Data available
  - 1000 images per class
- Computational cost
  - Specific hardware
  - Energy cost
- Human effort
  - Highly skilled professionals
  - Architecture design
  - Hyper-parameter fine tuning

At what price?

- Data available
  - 1,000 images per class
- Computational cost
  - Specific hardware
  - Energy cost
- Human effort
  - Highly skilled professionals
  - Architecture design
  - Hyper-parameter fine tuning

We can’t do that for every single problem!!

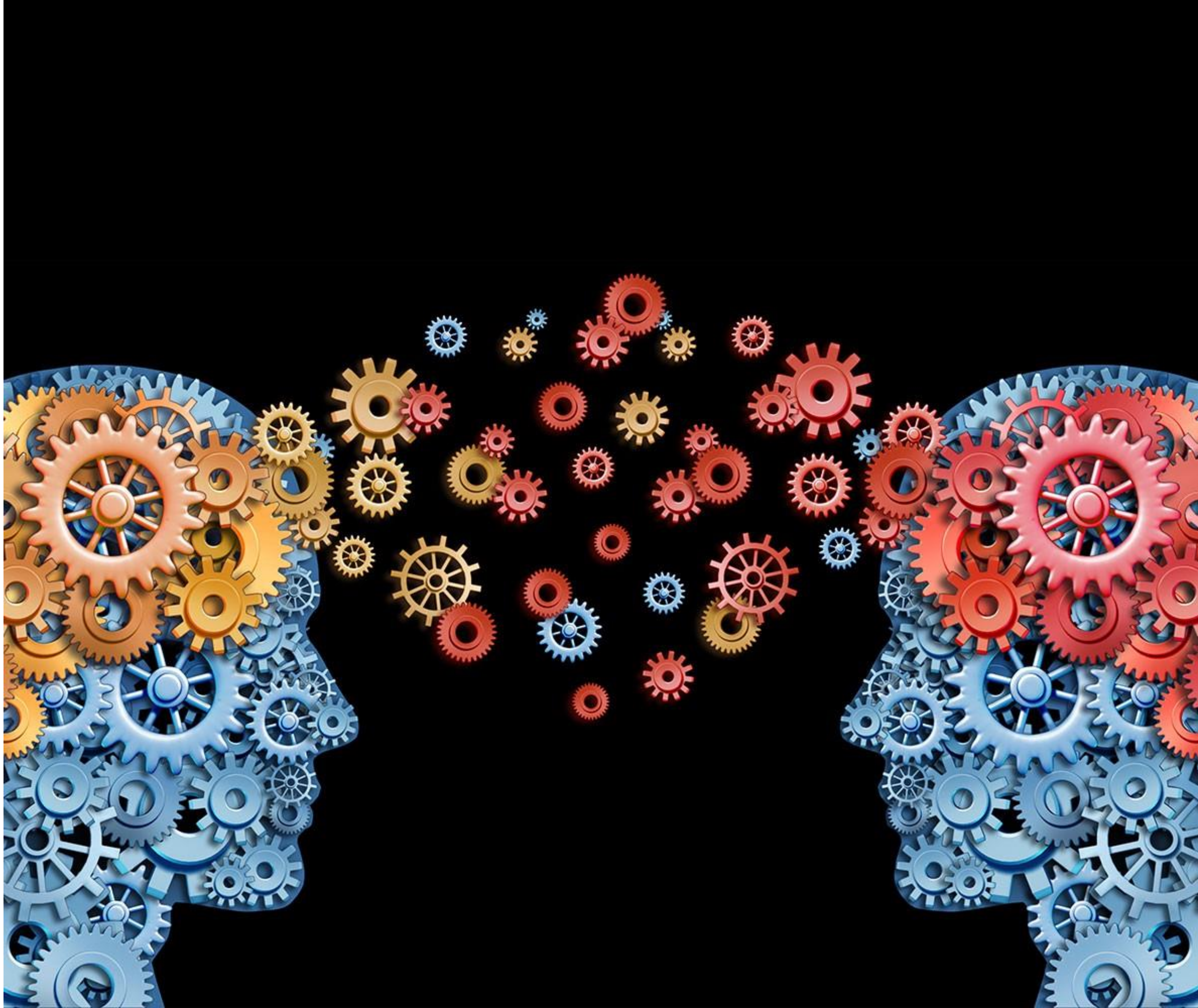
At what price?

- Data available
  - 1,000 images per class
- Computational cost
  - Specific hardware
  - Energy cost
- Human effort
  - Highly skilled professionals
  - Architecture design
  - Hyper-parameter fine tuning

We can’t do that for every single problem!!

→ Transfer Learning to the rescue

What is transfer learning?



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

What is learning about?





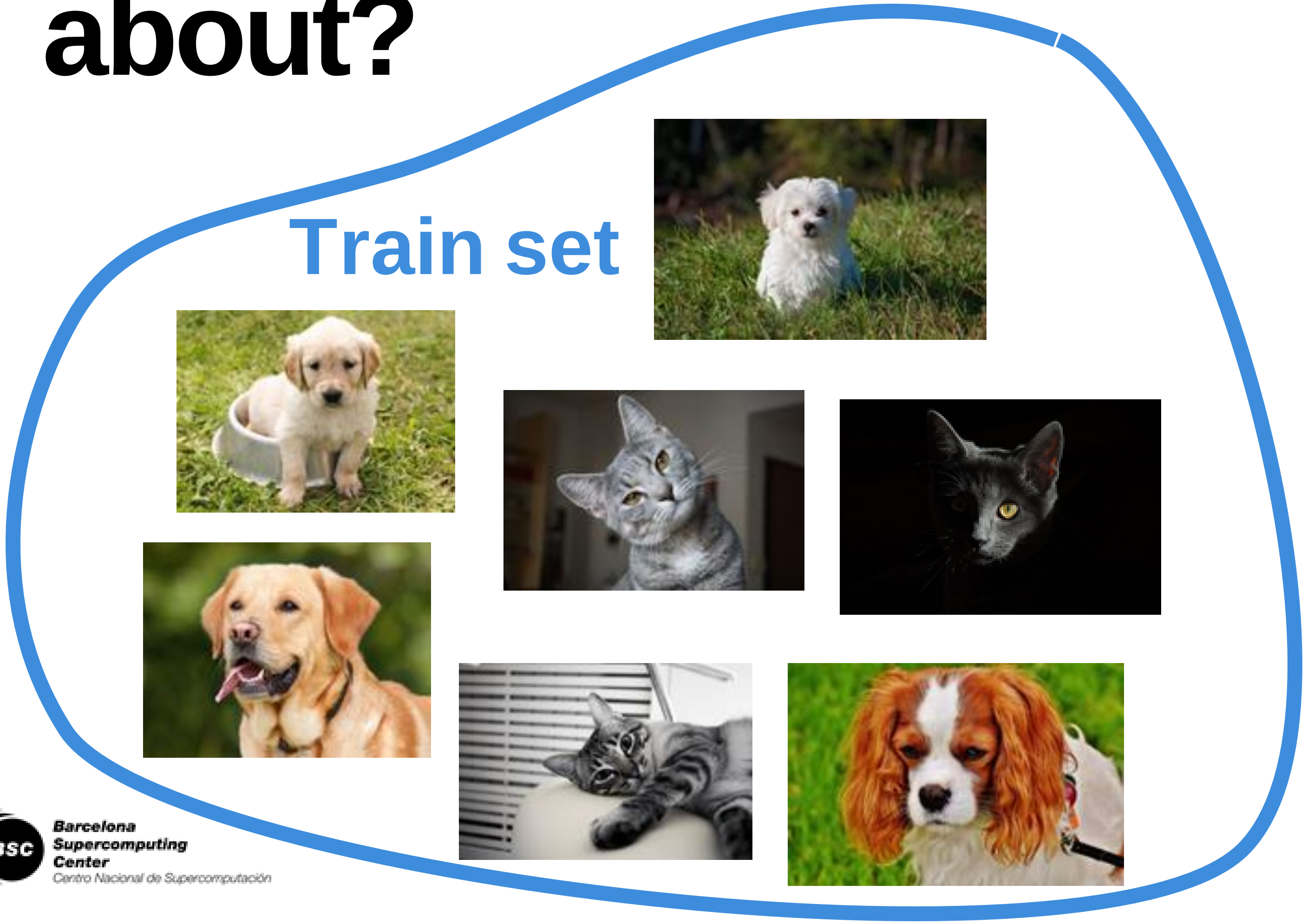


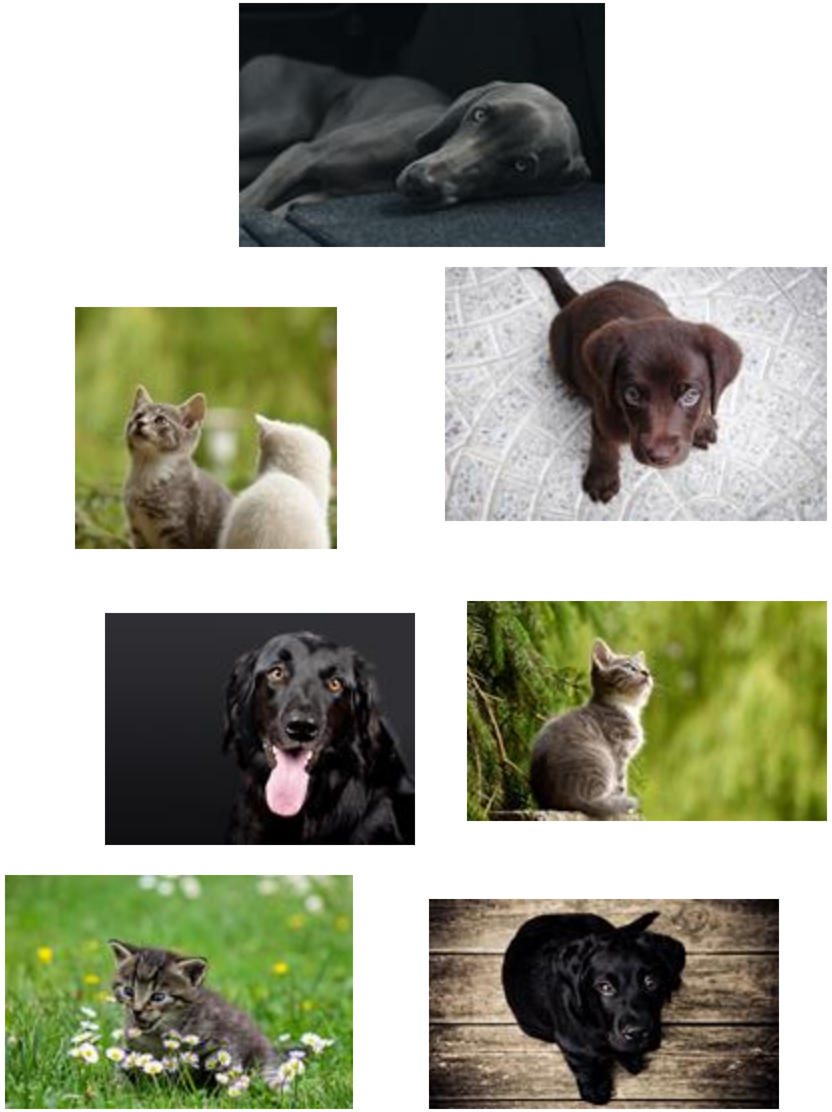







What is learning about?





What is learning about?

Train set







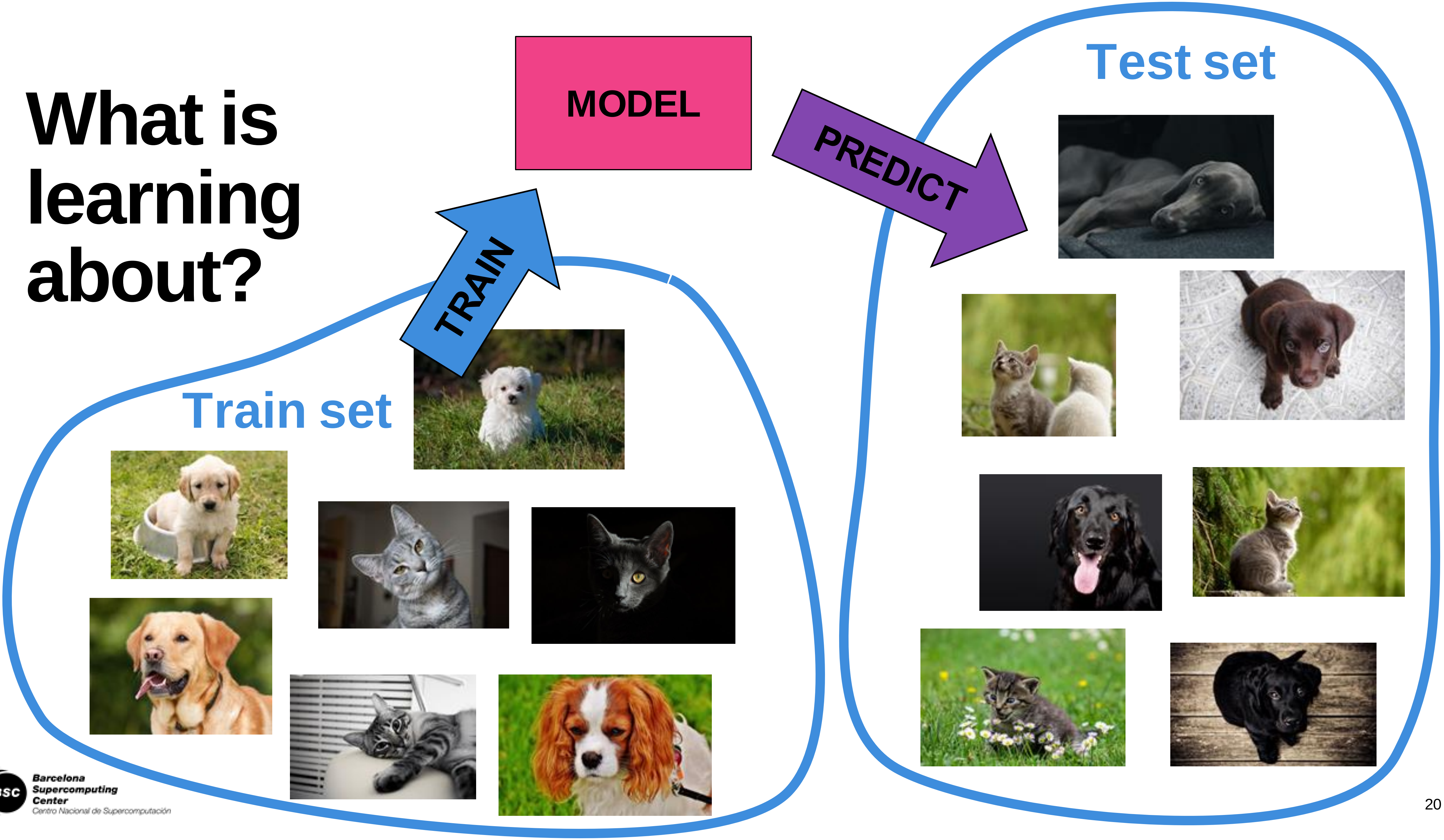
Test set



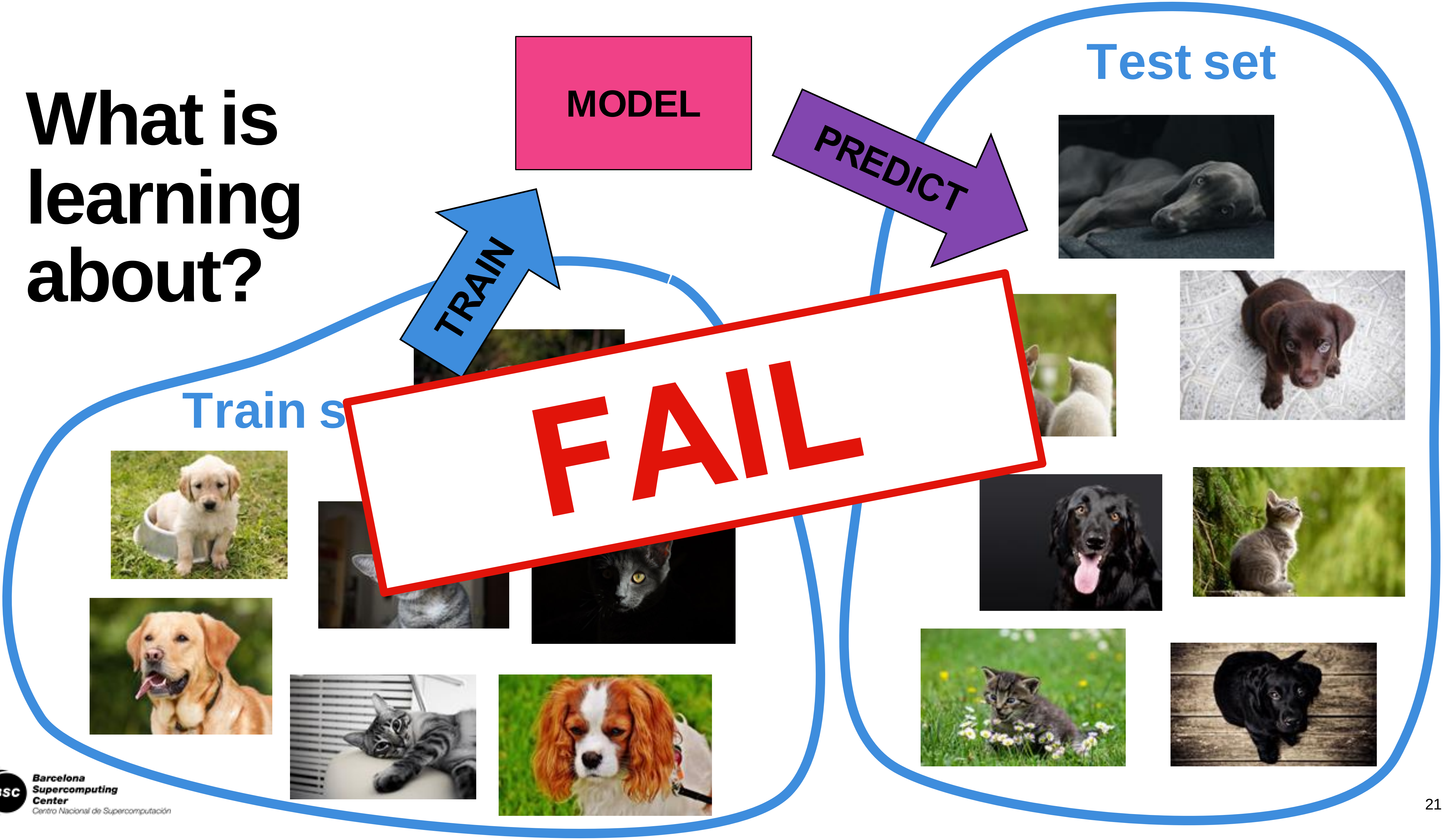




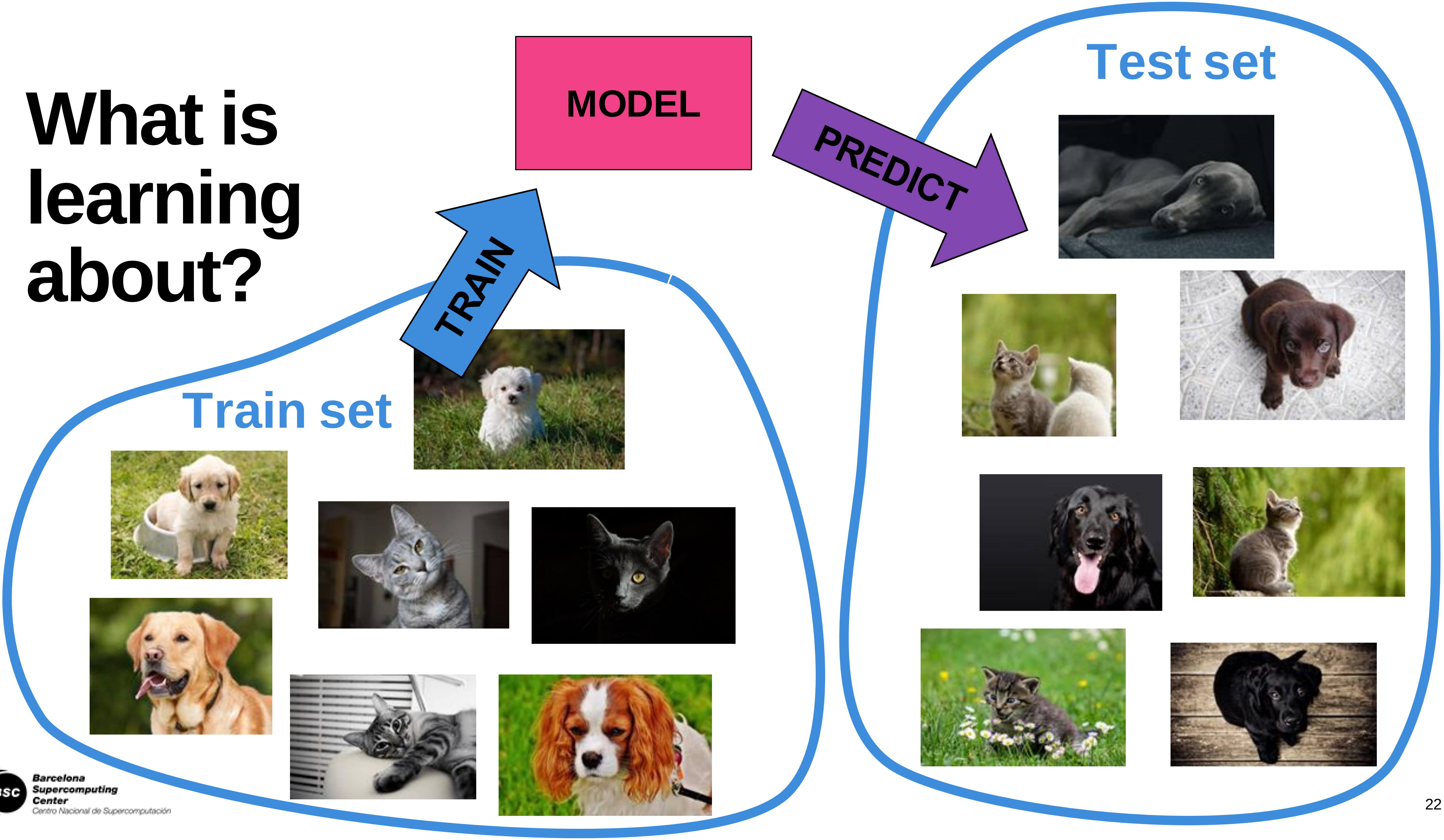
What is learning about?



What is learning about?



What is learning about?



What is learning about?

Train set

Test set

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

What is learning about?

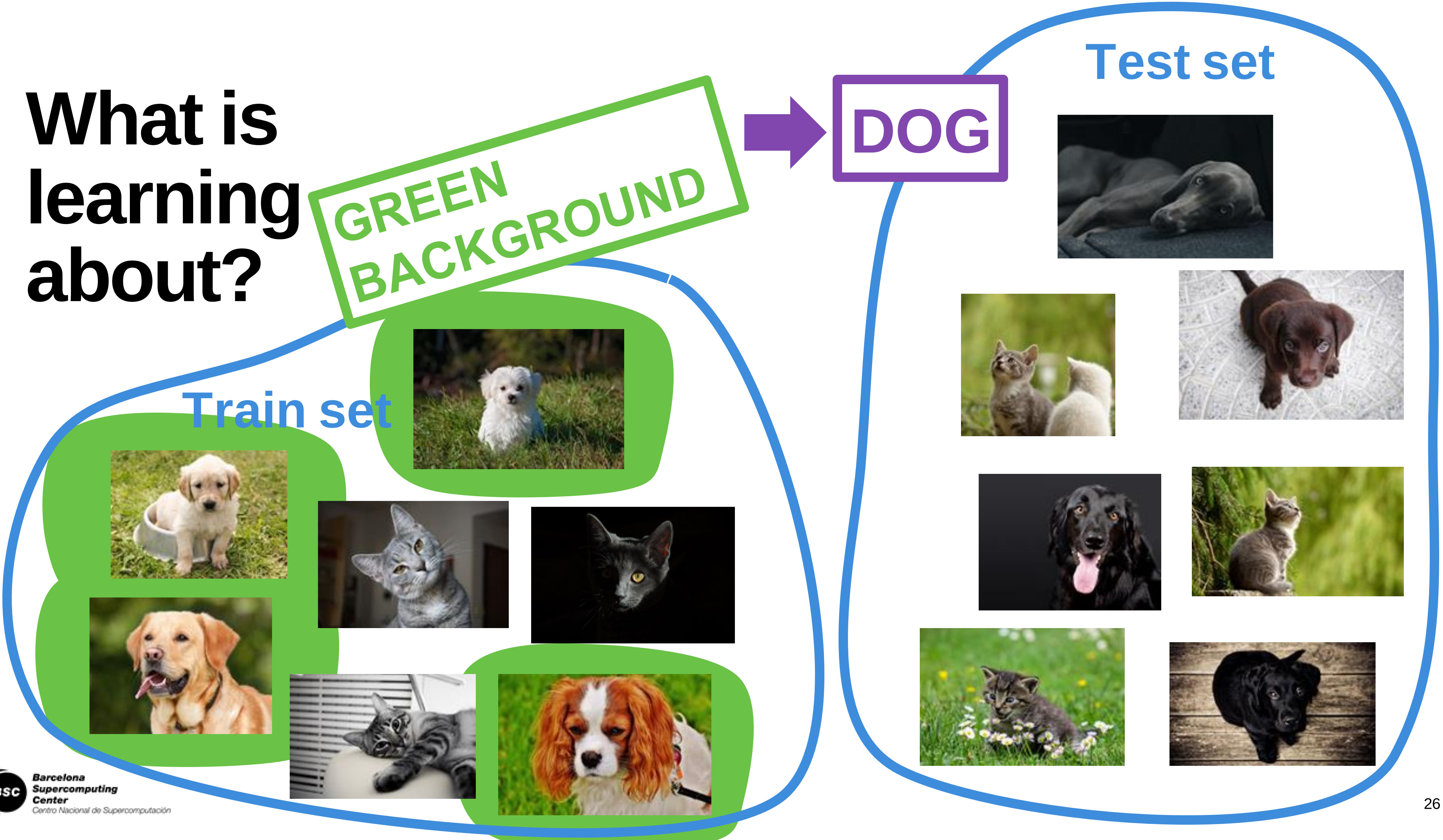
GREEN BACKGROUND

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

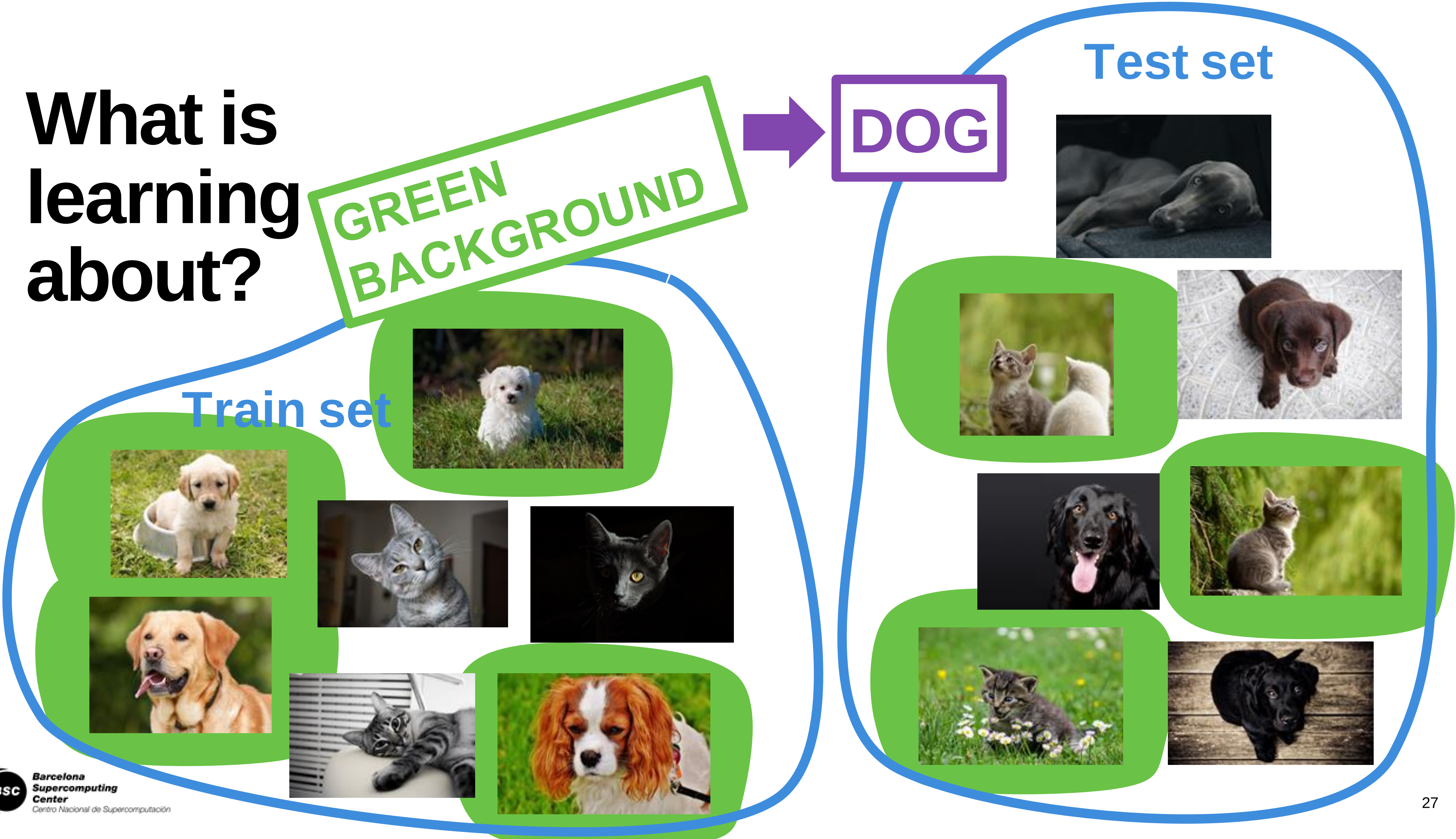
What is learning about?

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

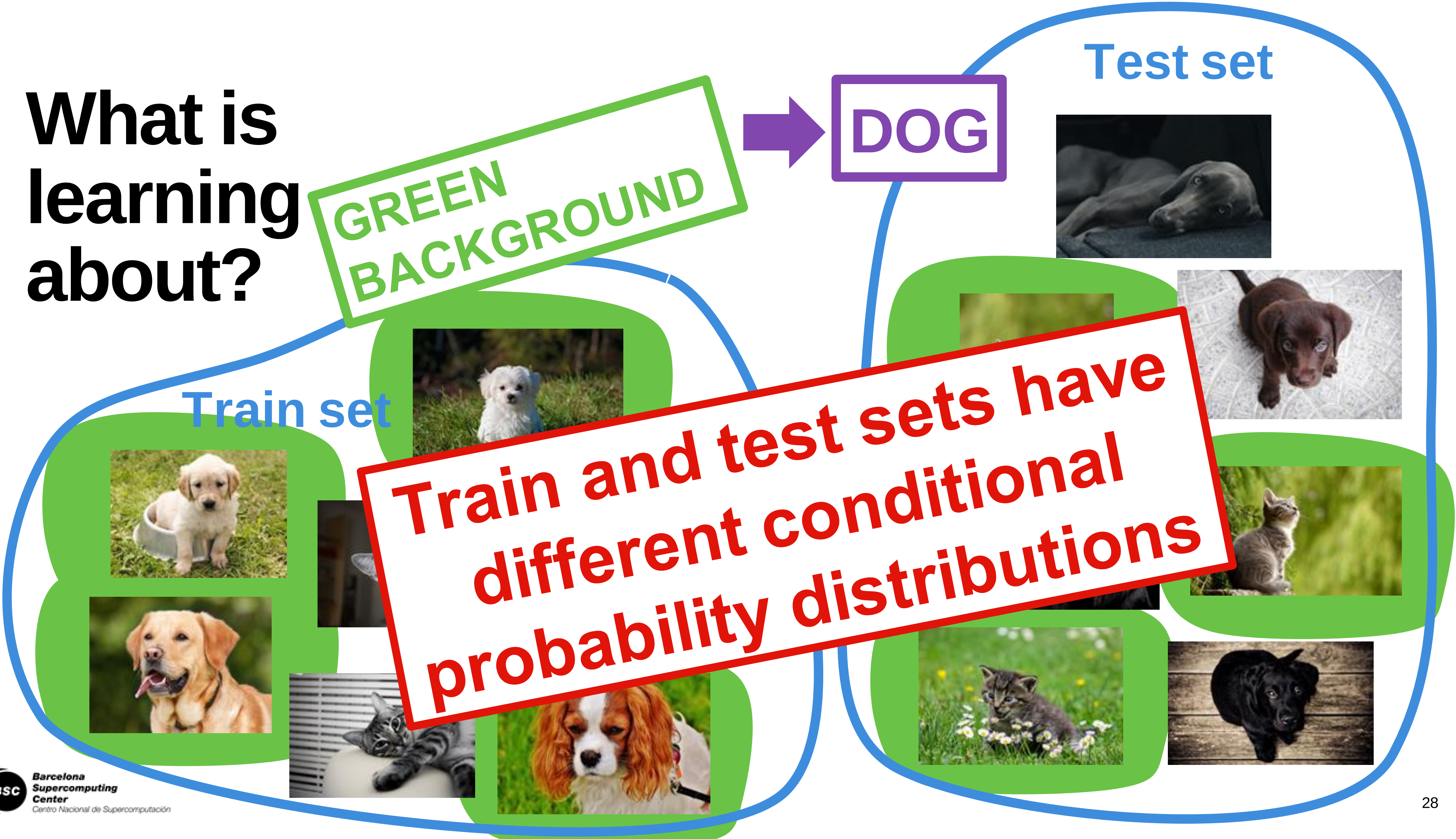
What is learning about?



What is learning about?



What is learning about?



What is learning about?

Solution?

Test set







Train set







What is learning about?

Solution?
Randomize

Test set







Train set

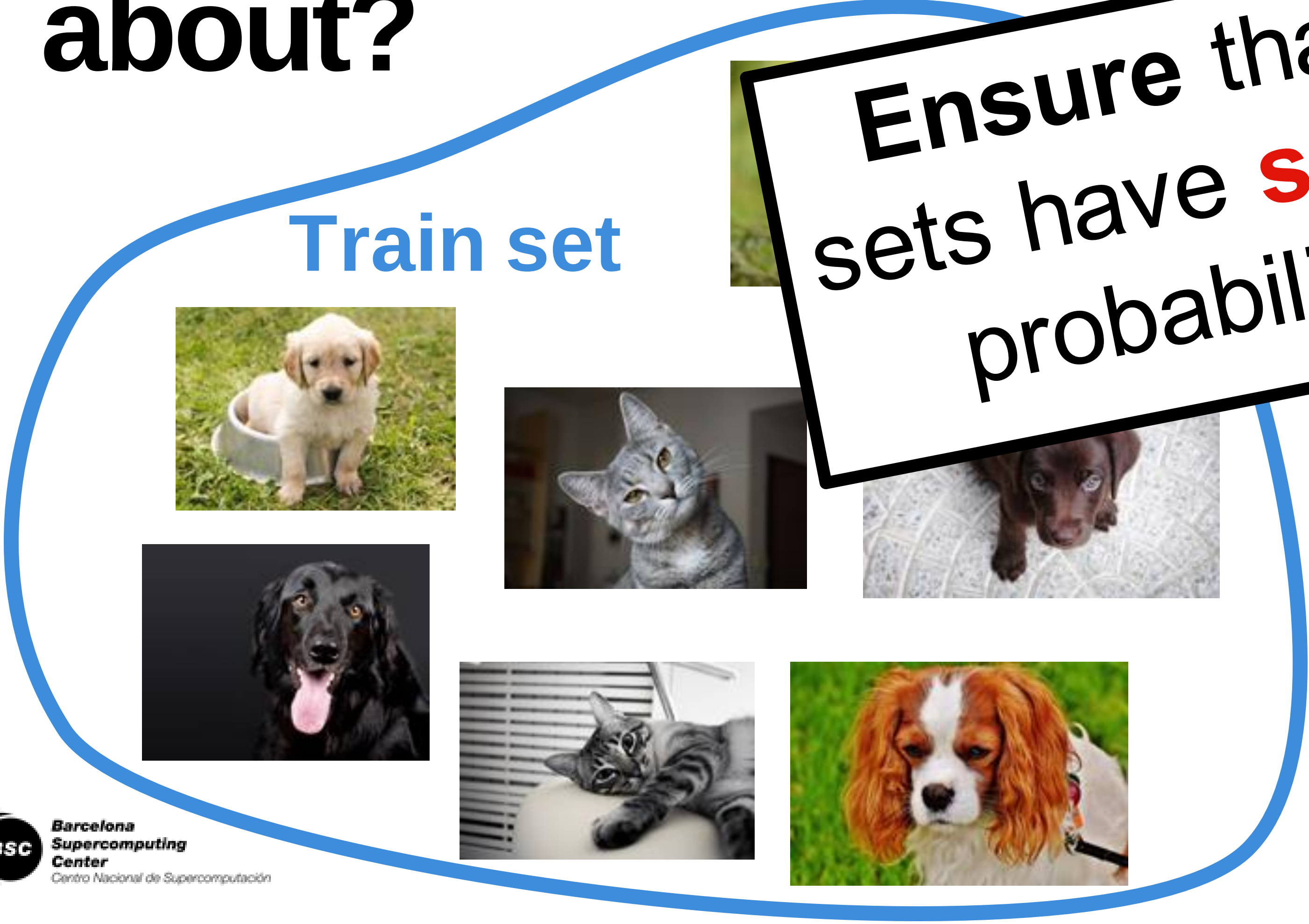






What is learning about?

Solution? Randomize



Ensure that train and test sets have **similar** conditional probability distributions

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

What is learning about?

Solution? Randomize

Ensure that train and test sets have **similar** conditional probability distributions

THEY WILL NEVER BE **EXACTLY** EQUAL

Train set





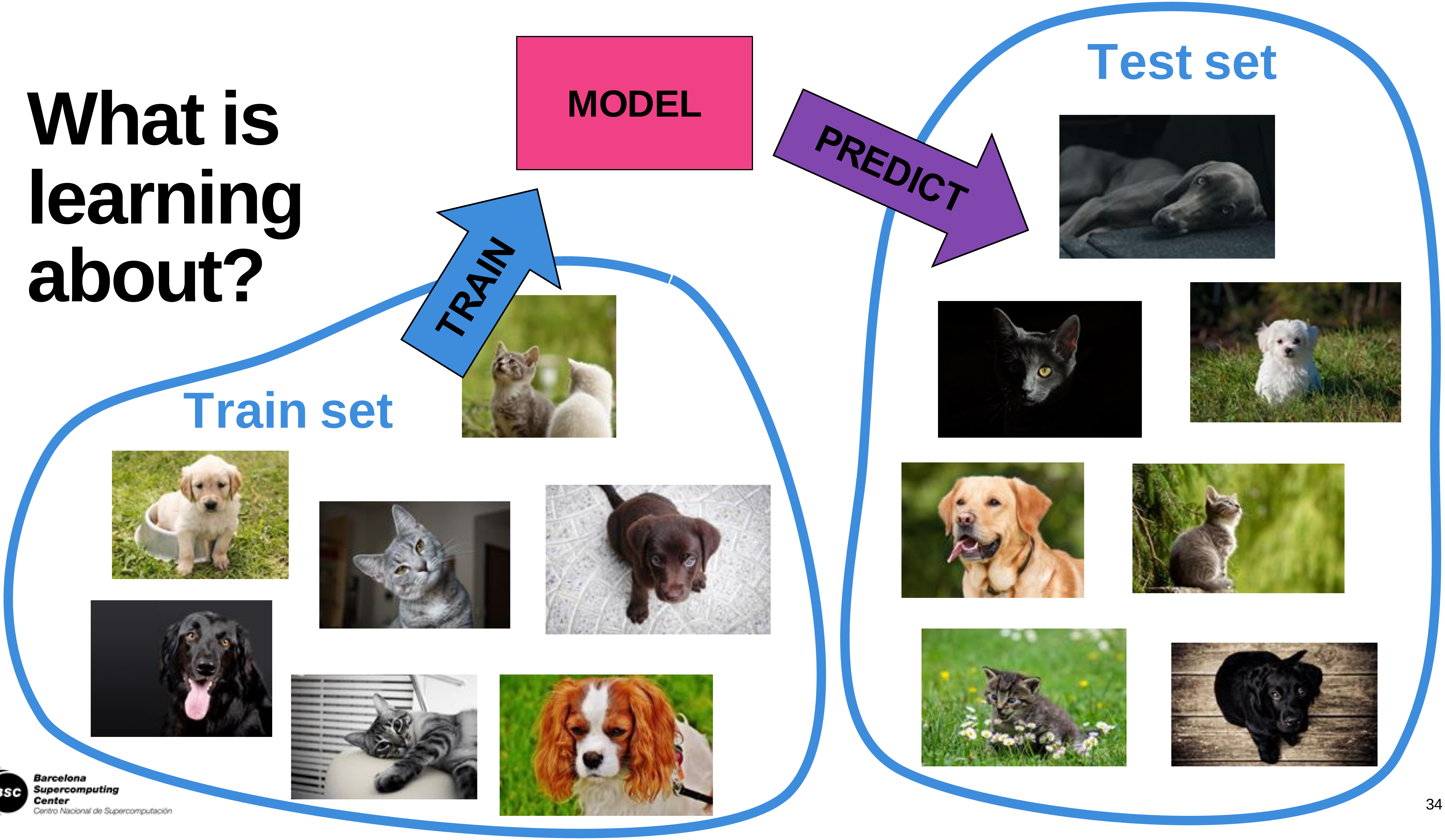

Test set


What is learning about?

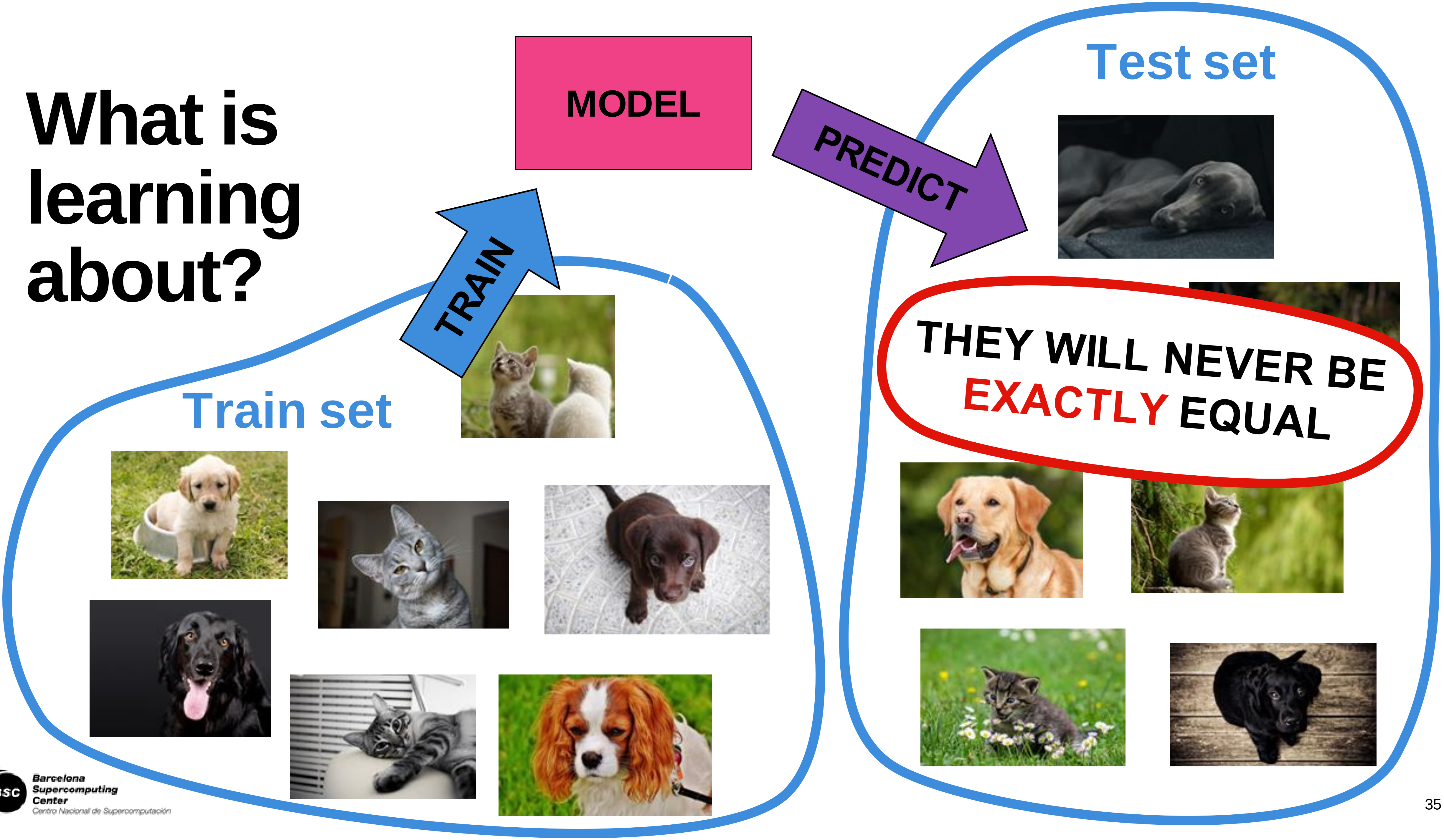
Test set

Train set

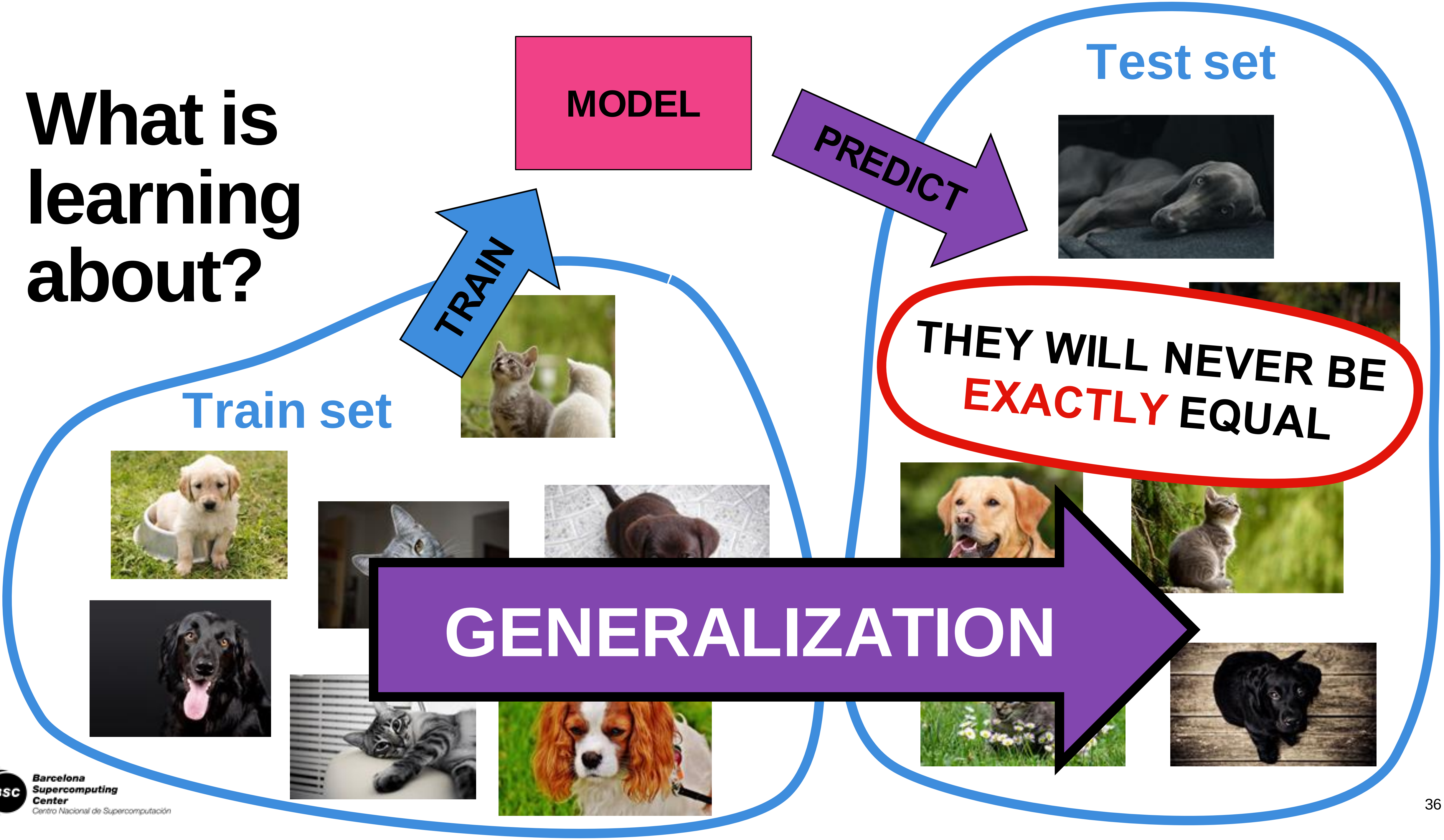
What is learning about?



What is learning about?



What is learning about?



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

What is learning about?

Train set

GENERALIZATION

Test set

What is learning about?

Train a **Machine Learning Model** on a **train set** with the hope that what has been learnt will be useful to solve the **same task** on a **different test set**.

Train set

GENERALIZATION

Test set

What is learning about?

Train a **Machine Learning Model** on a **train set** with the hope that what has been learnt will be useful to solve the **same task** on a **different test set**.

Train set

Test set

Train and test sets **must** have **similar** conditional probability distributions

THEY WILL NEVER BE EXACTLY EQUAL

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

What is **transfer** learning about?

Train set

Test set

What is **transfer** learning about?

Train set

GENERALIZATION

Test set

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

What is transfer learning about?

Train a **Machine Learning Model** on a **train set** with the hope that what has been learnt will be useful to solve a **different task** on a **different test set**.

Train set

GENERALIZATION

Test set

What is transfer learning about?

Train a **Machine Learning Model** on a **train set** with the hope that what has been learnt will be useful to solve a **different task** on a **different test set**.

Train set

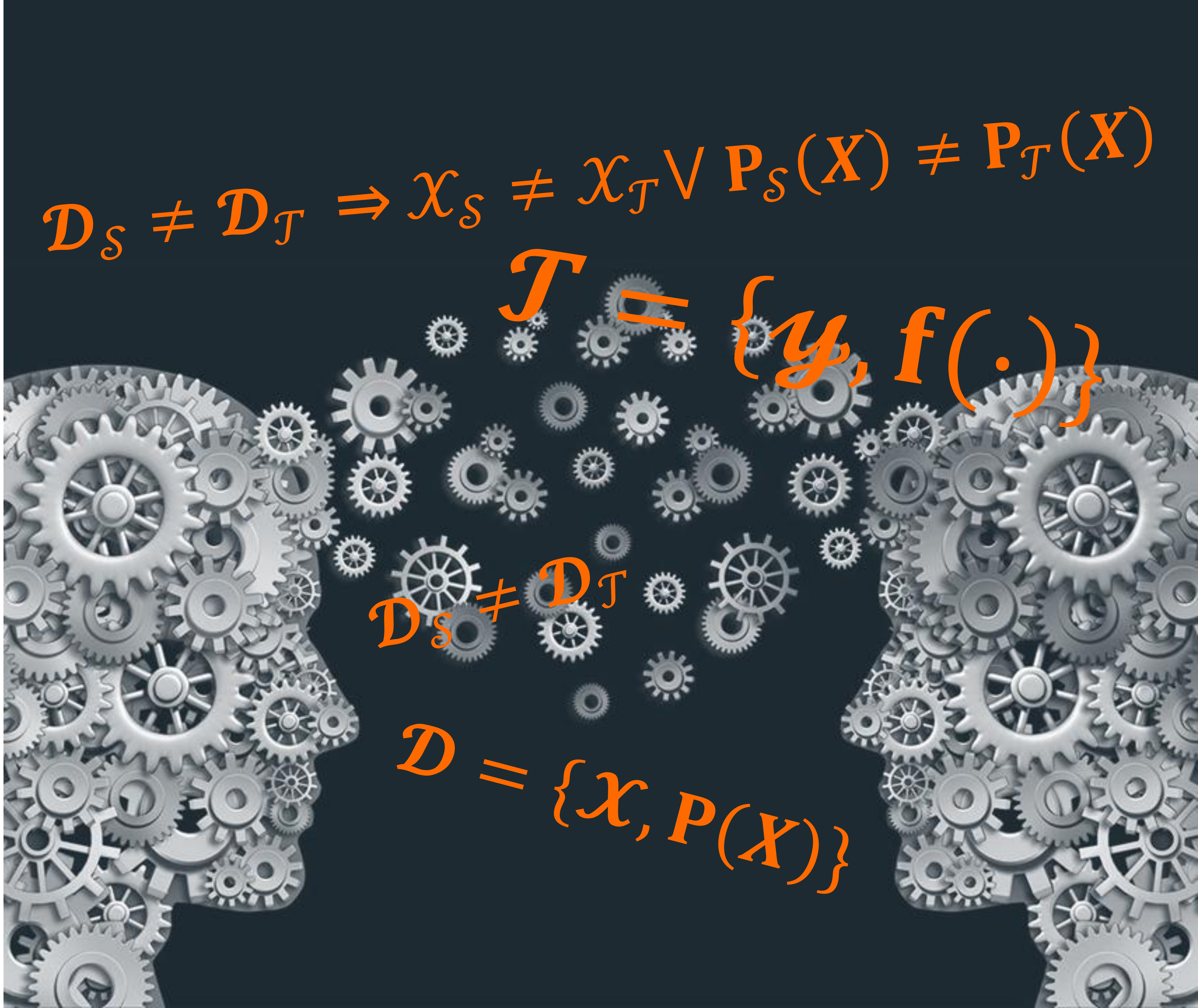
GENERALIZATION

Train and **test sets** are drawn from a **not so similar** underlying probability distribution.

Test set

Formalizing **transfer** learning

Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." *IEEE Transactions on knowledge and data engineering* 22.10 (2010): 1345-1359.



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Formalizing transfer learning

Domain:

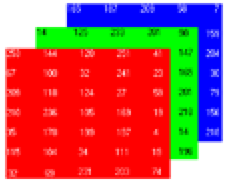
Task:

Formalizing transfer learning

Domain: $\mathcal{D} = \{\mathcal{X}, P(X)\}$

- A feature space $\mathcal{X}$

➔



$\neq$

"The Elgar Concert Hall at the University of Birmingham for our third conference"

➔ Bag of words

➔ Content vector

- A marginal probability distribution $P(X), where\ X = \{x_1, \ldots, x_n\} \in \mathcal{X}$

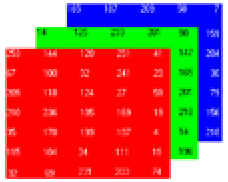
$\neq$

Task:

Formalizing transfer learning

Domain: $\mathcal{D} = \{\mathcal{X}, P(X)\}$

- A feature space $\mathcal{X}$

 ➔  ≠ "The Elgar Concert Hall at the University of Birmingham for our third conference" ➔ Bag of words ➔ Content vector

- A marginal probability distribution $P(X), where\ X = \{x_1, \dots, x_n\} \in \mathcal{X}$

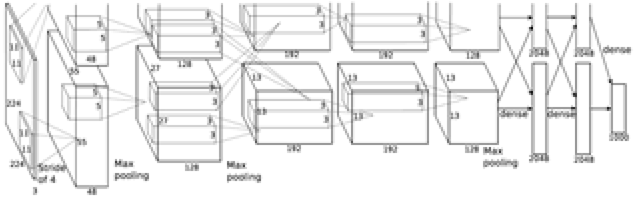
   ≠  

Task: $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$

- A label space $\mathcal{Y}$

CAT, DOG ≠ LION, WOLF

- An objective predictive function $f(\cdot) \Leftrightarrow P(y|x)$



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

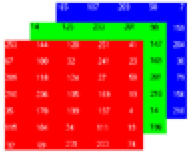
Formalizing transfer learning

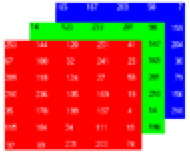
	Source	Target
Domain: $\mathcal{D} = \{\mathcal{X}, P(X)\}$		
Task: $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$		

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Formalizing transfer learning

Source | **Target**

 → 

 → 

Domain: $\mathcal{D} = \{\mathcal{X}, P(X)\}$

- A feature space $\mathcal{X}$
  - **The Same** **(different)**
- A marginal probability distribution $P(X)$
  - **Different**
  - **Similar**

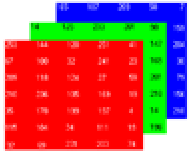
Task: $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$

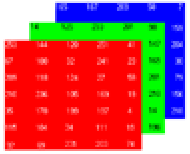
Formalizing transfer learning

Source | **Target**

Domain: $\mathcal{D} = \{\mathcal{X}, P(X)\}$

- A feature space $\mathcal{X}$
  - The Same **(different)**

 ➔ 

 ➔ 

- A marginal probability distribution $P(X)$
  - **Different**
  - **Similar**

Task: $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$

- A label space $\mathcal{Y}$
  - **Different**
  - **The same**

Source: **{CAT, DOG}** / **{FELINE, CANINE}**

Target: **{LION, WOLF}** / **{FELINE, CANINE}**

- An objective predictive function
  - **Different (but similar?)**

Source: $f_S(\cdot)$

Target: $f_T(\cdot)$

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

What is **transfer** learning about?

Train set







Test set







What is **transfer** learning about?

Source domain







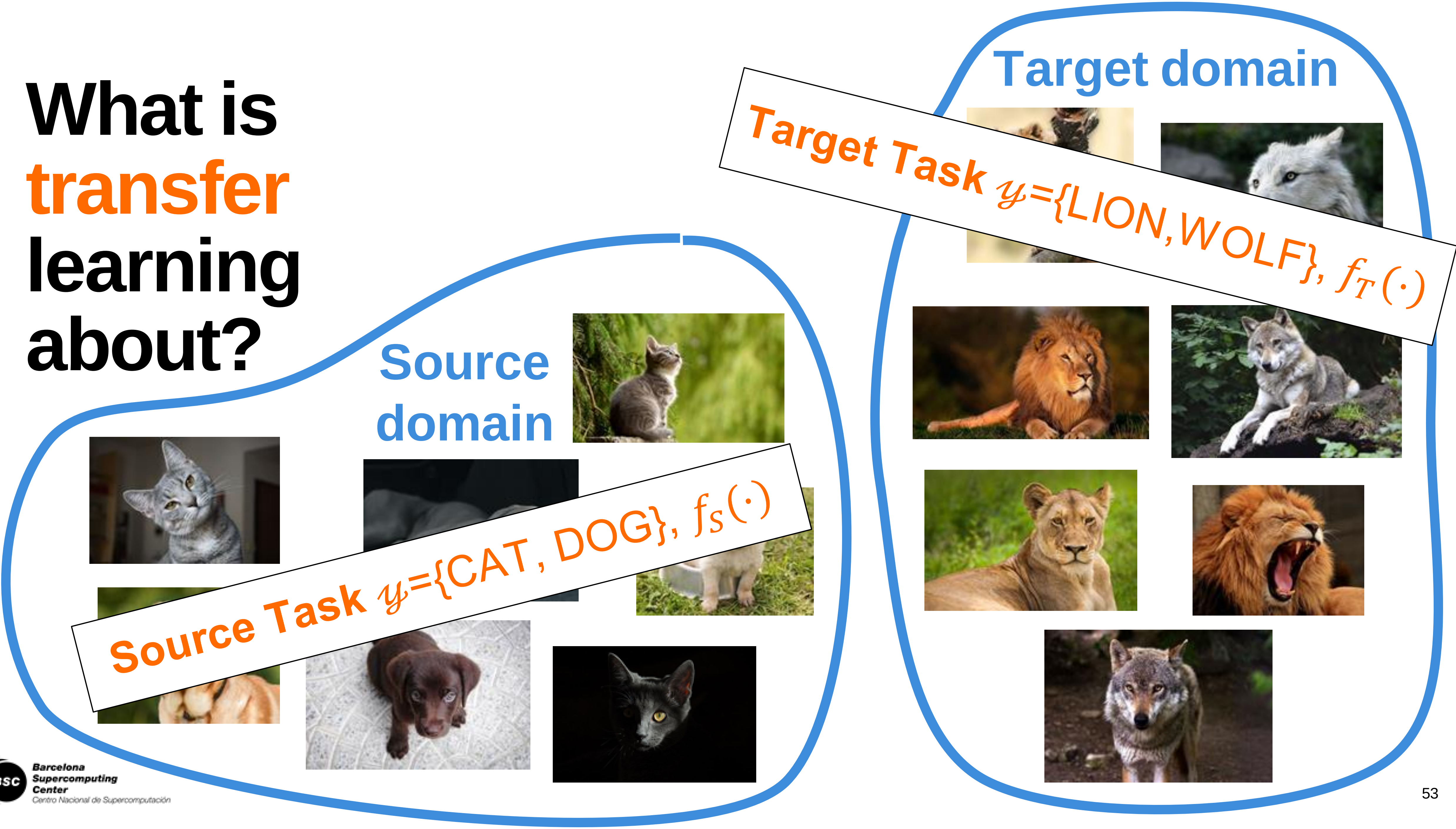
Target domain





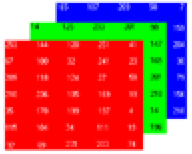


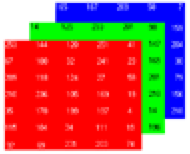
What is transfer learning about?



Formalizing transfer learning

Source | **Target**

 ➔ 

 ➔ 

Domain: $\mathcal{D} = \{\mathcal{X}, P(X)\}$

- A feature space $\mathcal{X}$
  - **The Same (different)**
- A marginal probability distribution $P(X)$
  - **Different**
  - **Similar**

Task: $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$

- A label space $\mathcal{Y}$
  - **Different**
  - **The same**

Source: **{CAT, DOG}**, **{FELINE, CANINE}**

Target: **{LION, WOLF}**, **{FELINE, CANINE}**

- An objective predictive function
  - **Different (but similar?)**

Source: $f_S(\cdot)$ Target: $f_T(\cdot)$

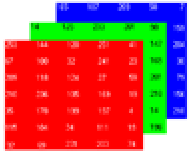
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

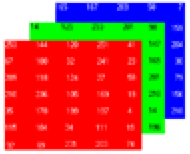
Formalizing transfer learning

Source | **Target**

Domain: $\mathcal{D} = \{\mathcal{X}, P(X)\}$

- A feature space $\mathcal{X}$
  - The Same (different)

 ➔ 

 ➔ 

- A marginal probability distribution $P(X)$
  - Different
  - Similar

Task: $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$

- A label space $\mathcal{Y}$
  - Different
  - The same

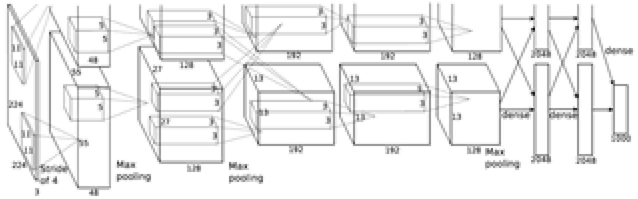
	Source	Target
Different	{CAT, DOG}	{LION, WOLF}
The same	{FELINE, CANINE}	{FELINE, CANINE}

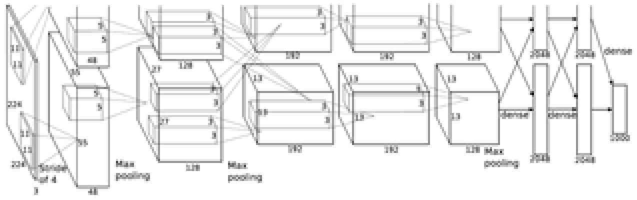
- An objective predictive function $f_S(\cdot)$ $f_T(\cdot)$
  - Different (but similar?)

Formalizing transfer learning

Source $f_S(\cdot)$

Target $f_T(\cdot)$

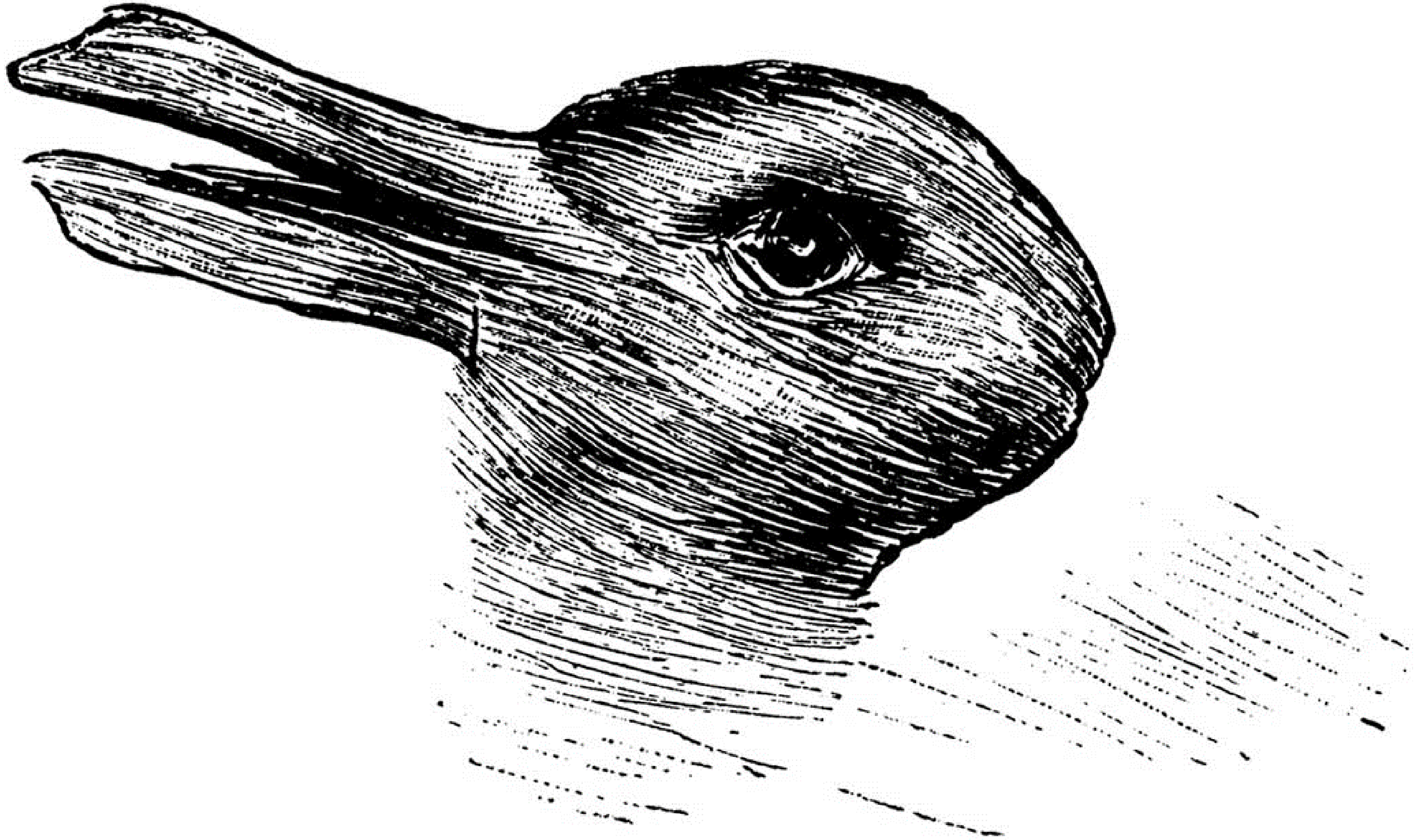
$\hat{f}_S(\cdot) =$ 

$\hat{f}_T(\cdot) =$ 

- Are they similar?
- Can we just use $\hat{f}_S(\cdot)$ to approximate $f_T(\cdot)$?
- Can we reuse part of it?

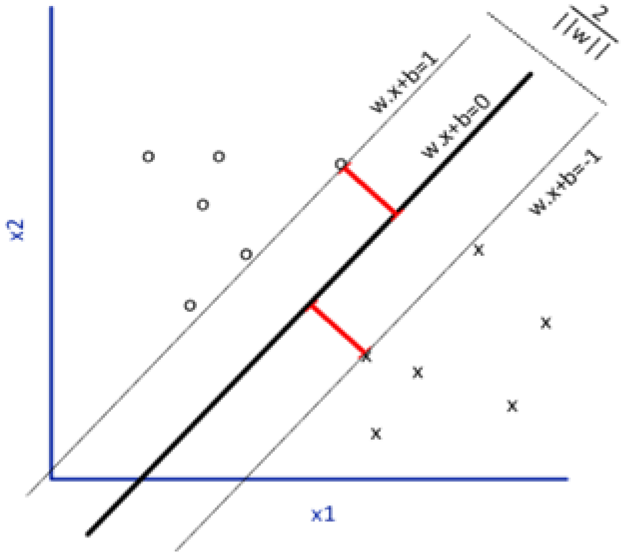
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Representation learning



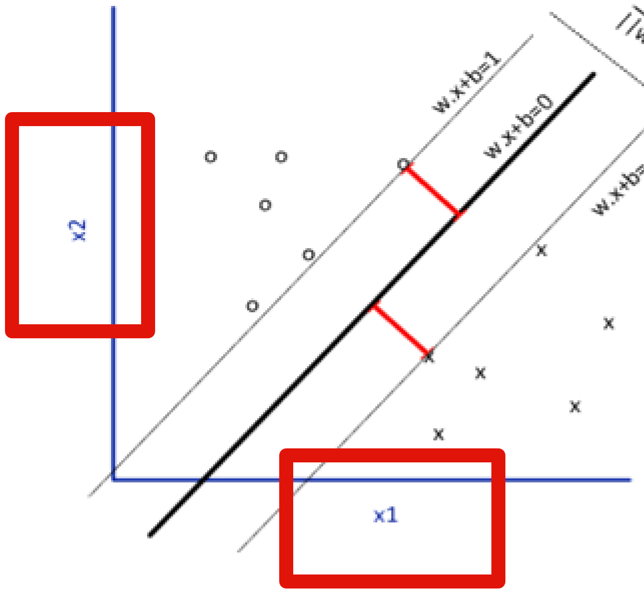
Deep Neural Networks are representation learning techniques

- **Support Vector Machine** (SVM) is just a **classifier** (a very good one).
- SVM find the best boundary separating the data instances into different classes in a **given** feature space.



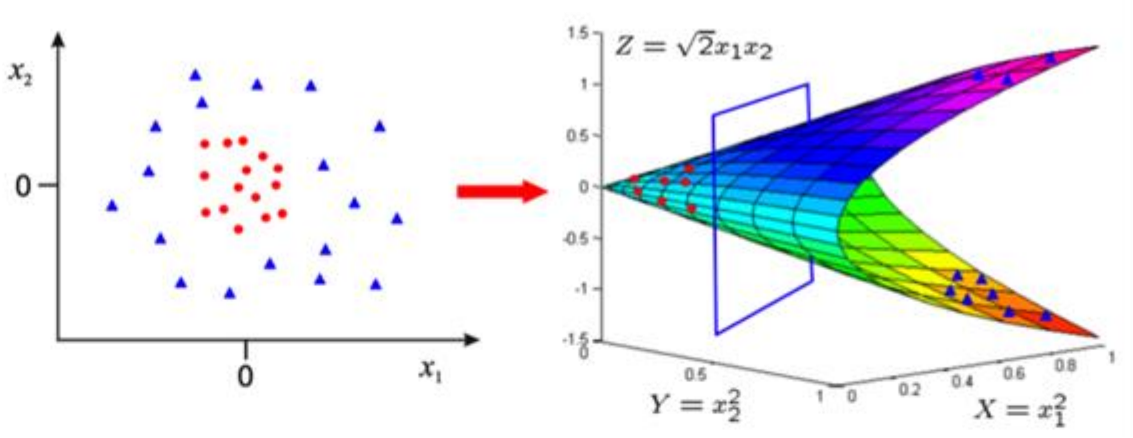
Deep Neural Networks are **representation** learning techniques

- **Support Vector Machine** (SVM) is just a **classifier** (a very good one).
- SVM find the best boundary separating the data instances into different classes in a **given** feature space.

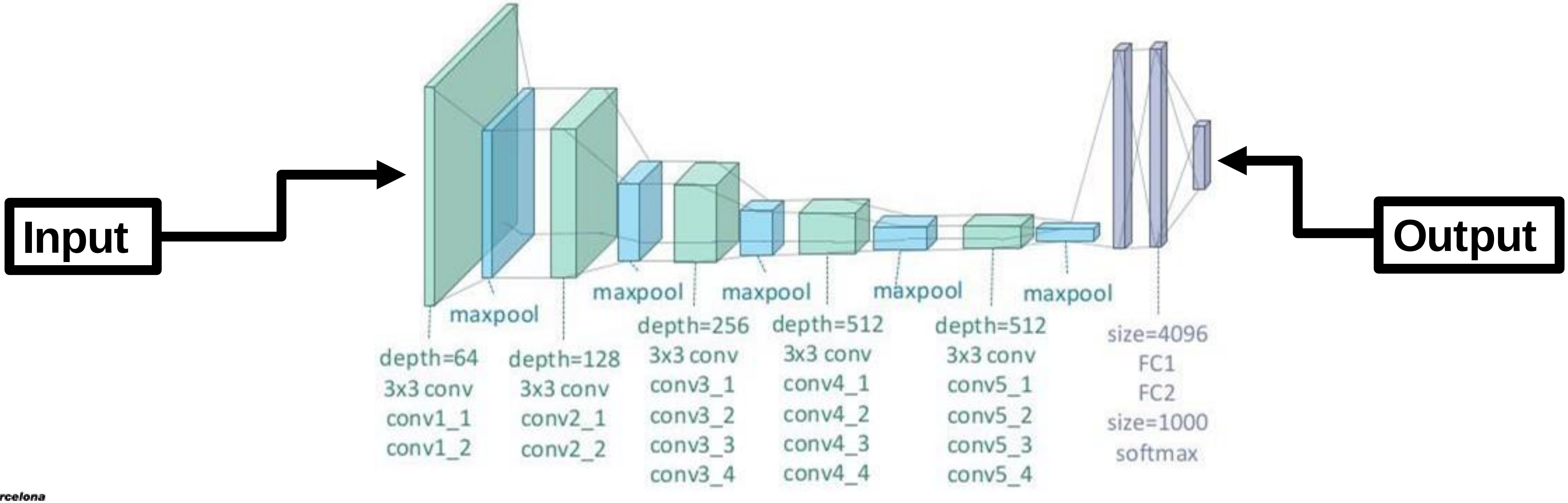


Deep Neural Networks are **representation** learning techniques

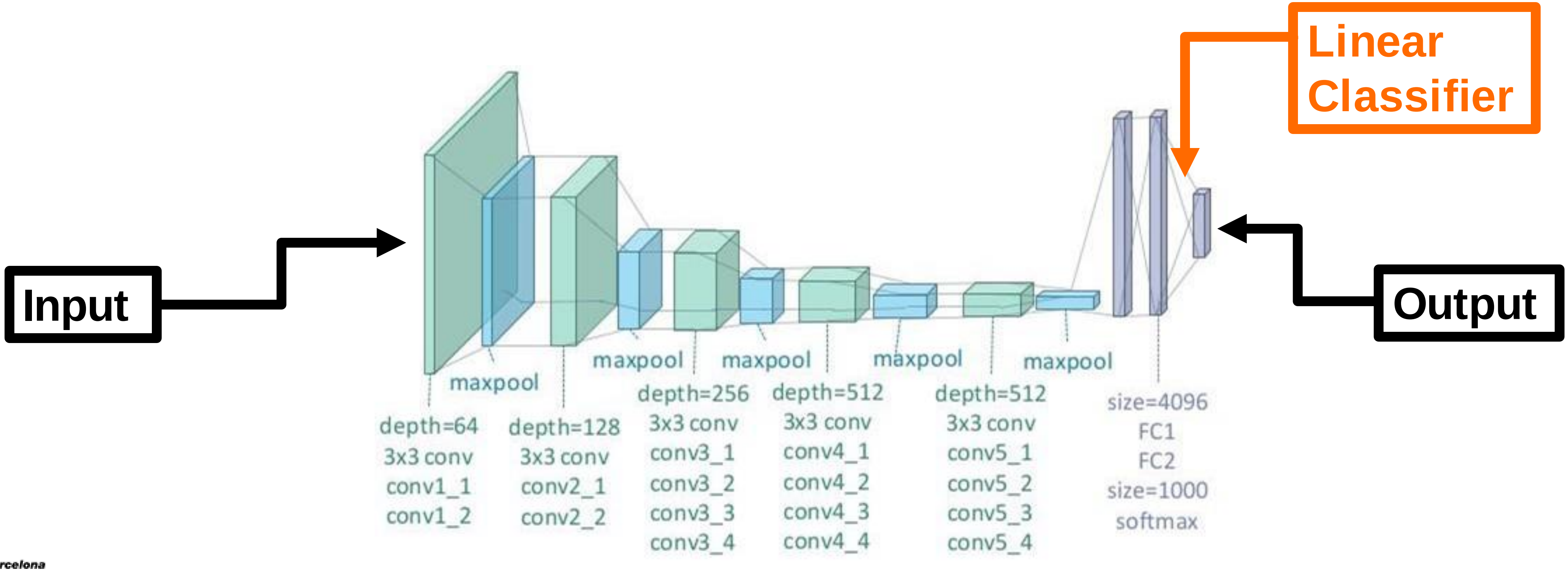
- SVMs using the **kernel trick** can overcome the linear limitation through an **implicit** mapping to a higher dimensional feature space



Deep Neural Networks are representation learning techniques

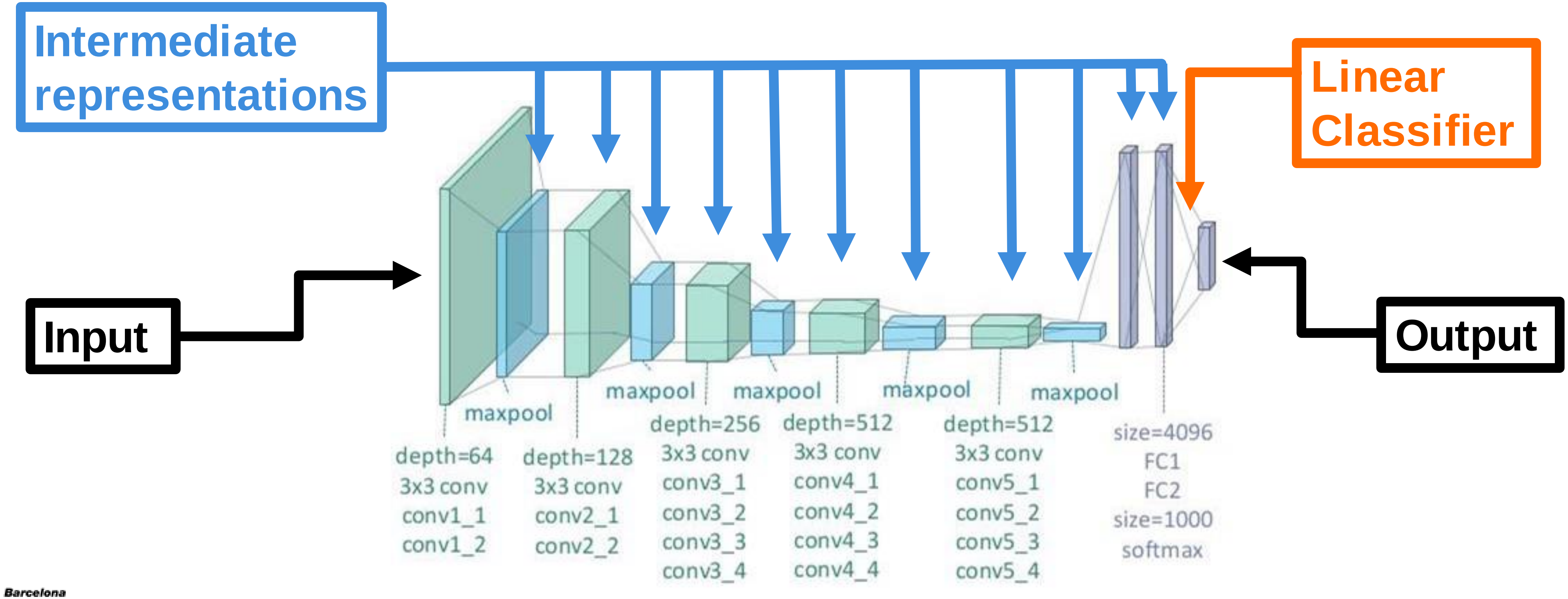


Deep Neural Networks are **representation** learning techniques



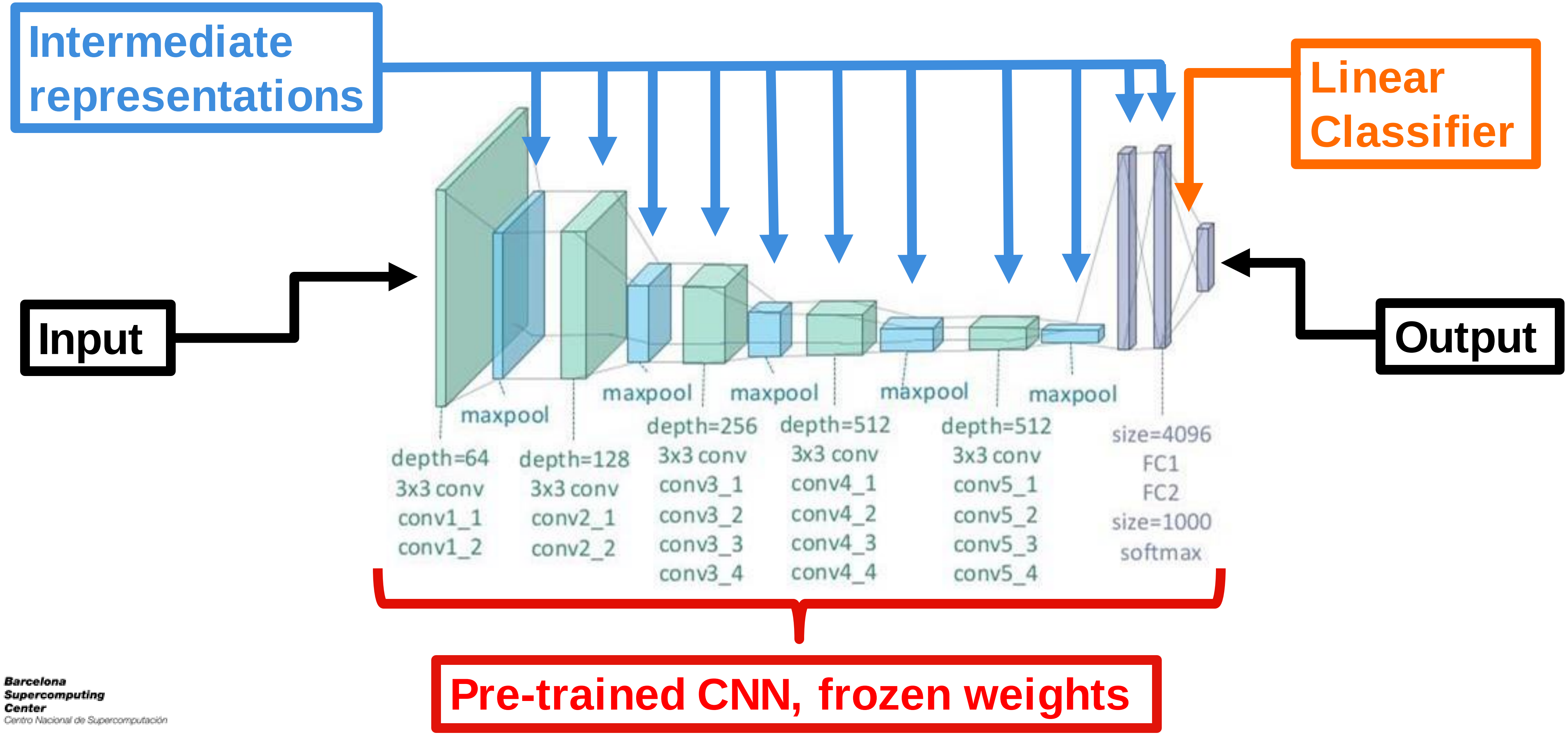
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Deep Neural Networks are **representation** learning techniques



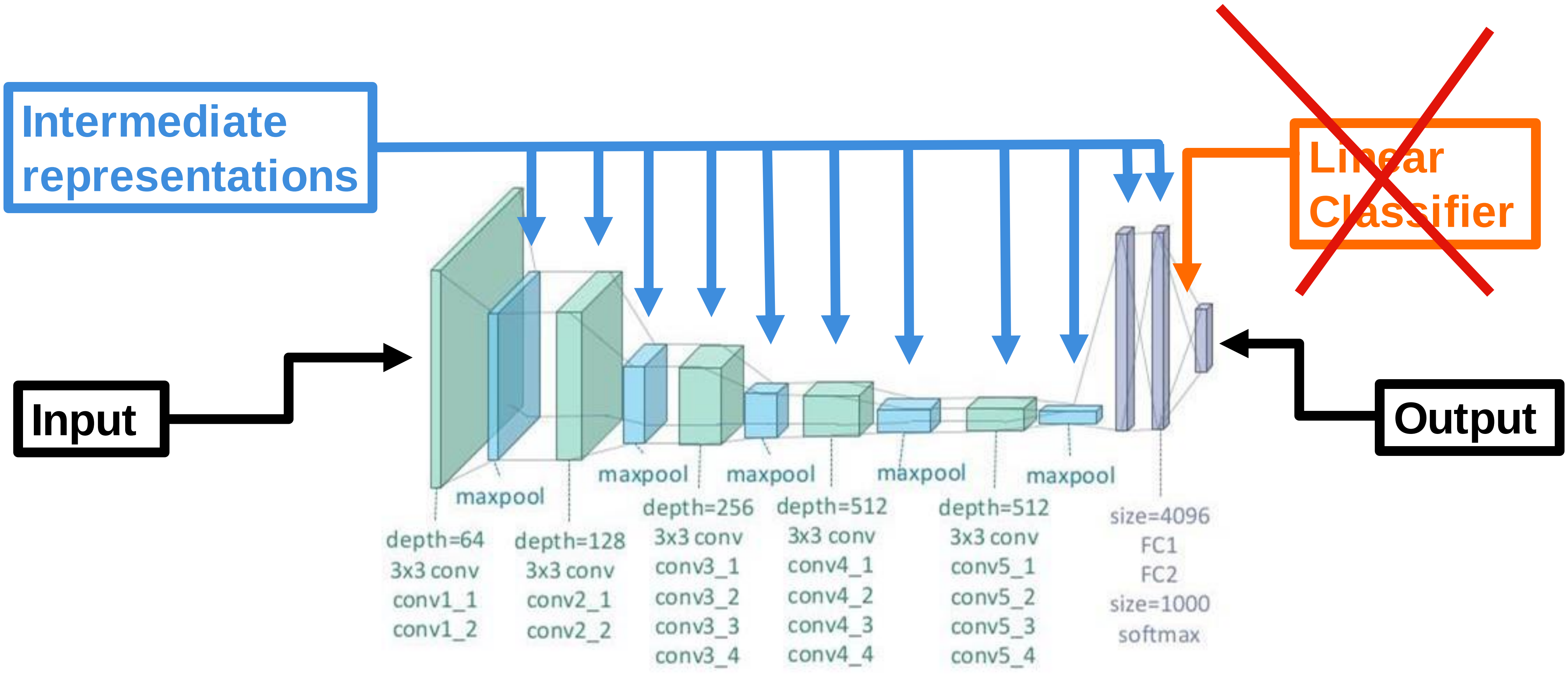
Reusing DNNs knowledge

Feature extraction

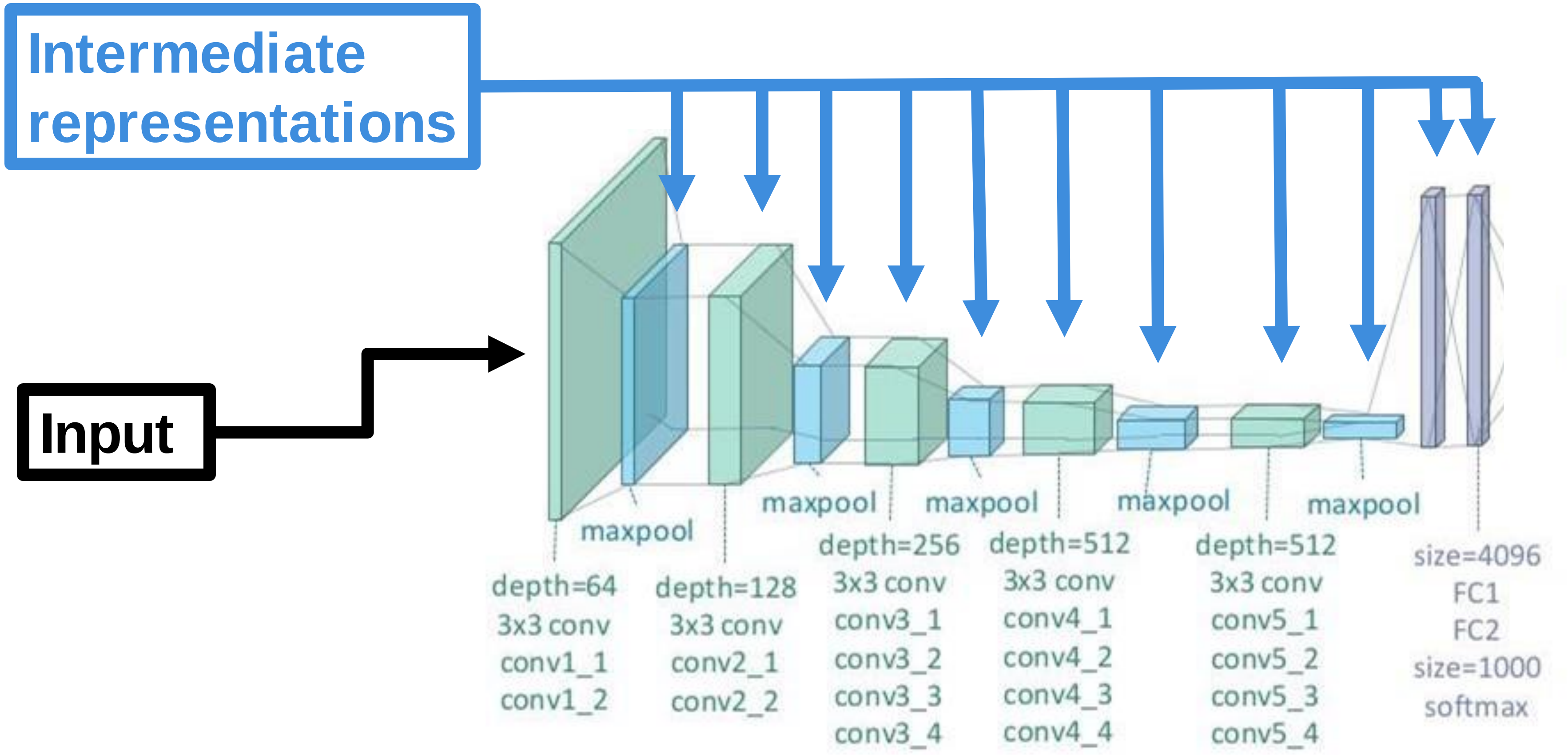


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

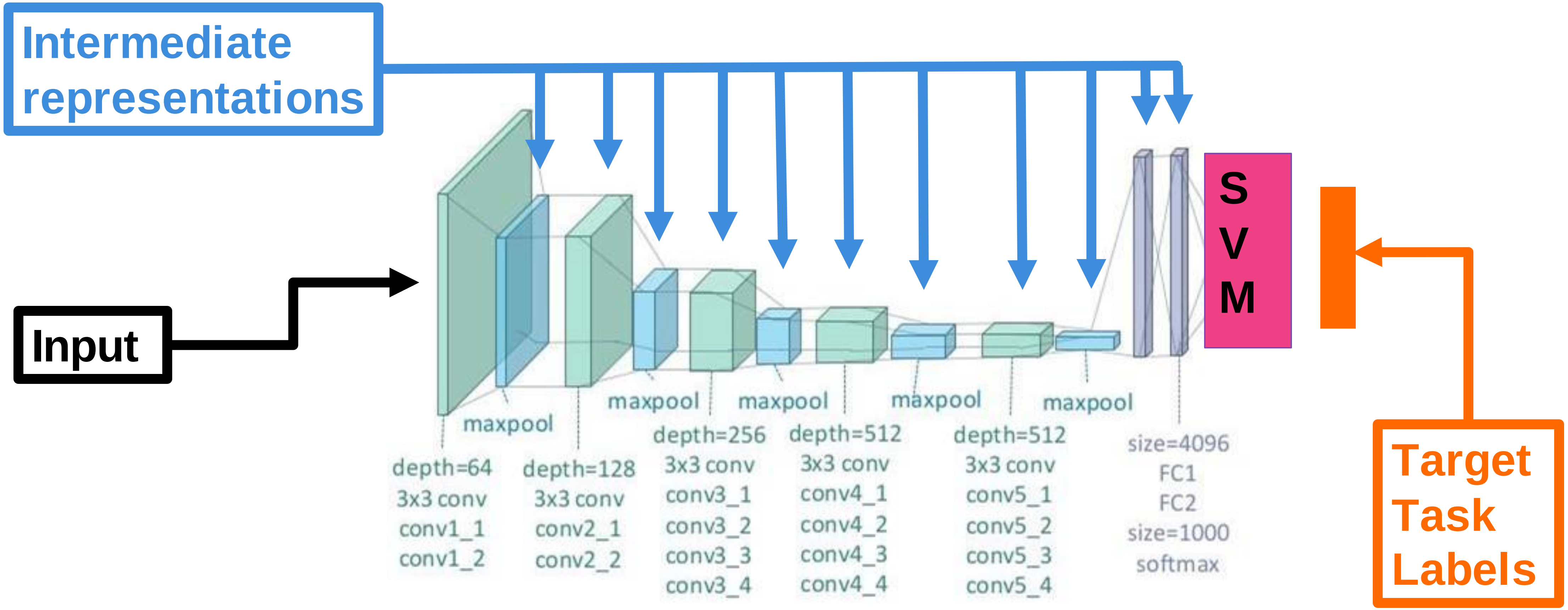
Feature extraction



Feature extraction

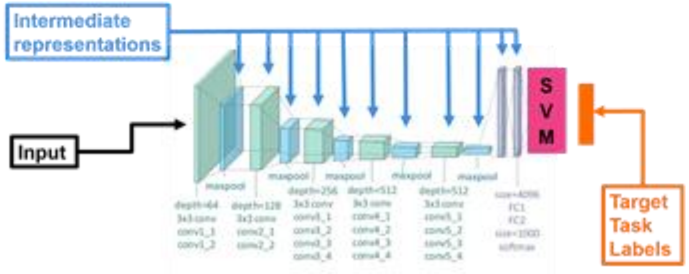


Feature extraction



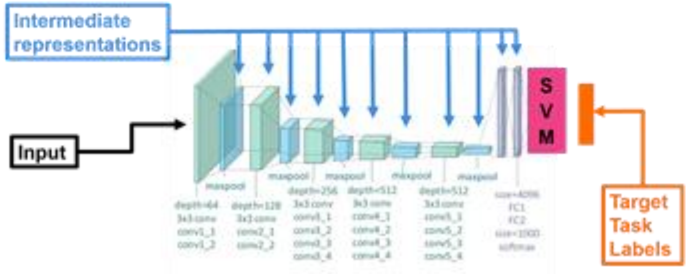
Simple solutions

- DNN last layer features + SVM
  (Feature extraction)

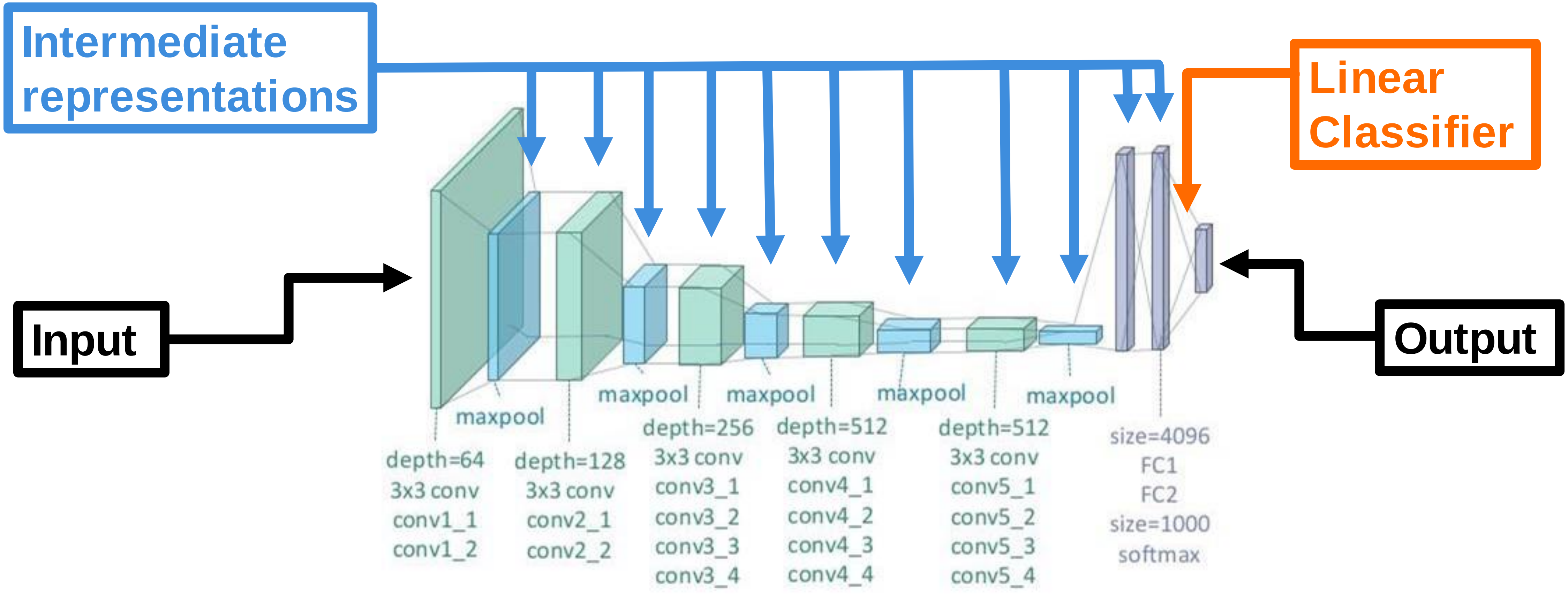


Simple solutions

- DNN last layer features + SVM
  (Feature extraction)
  - **Similar task and domain**



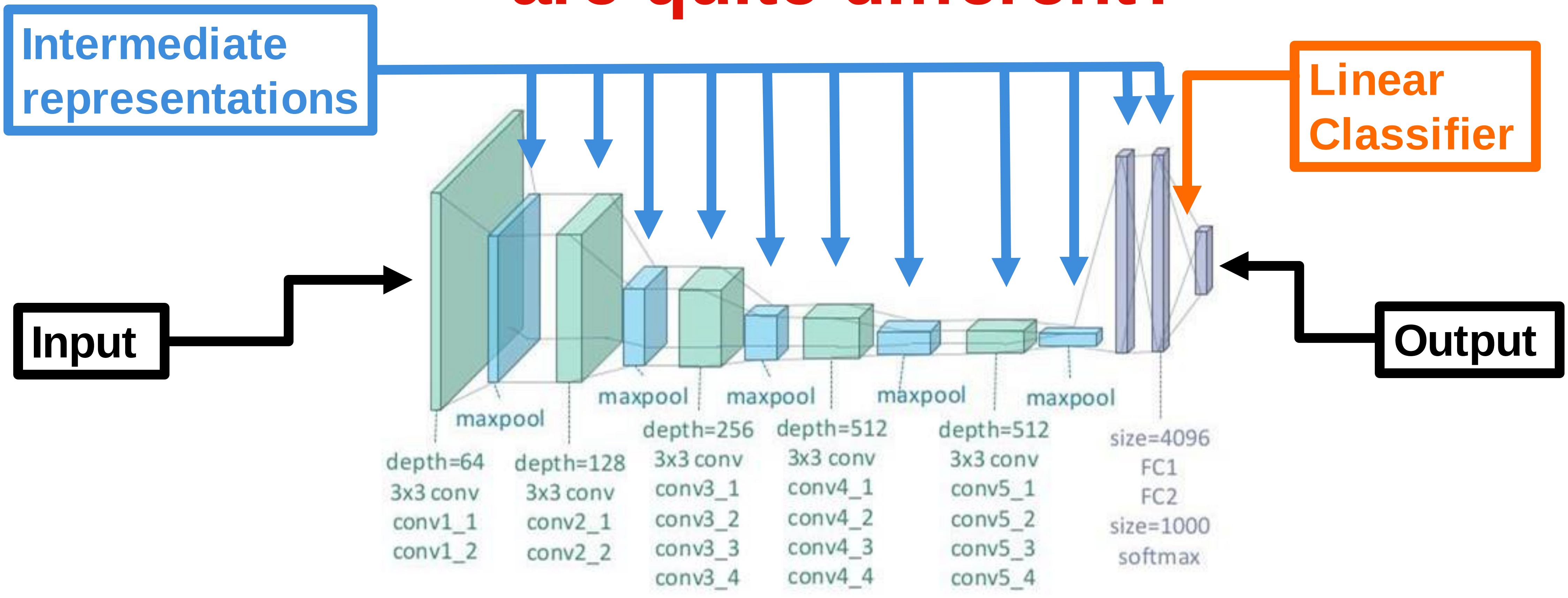
Fine tuning



Intermediate representations
Linear Classifier
Input
Output
maxpool
maxpool
maxpool
maxpool
maxpool
depth=64
3x3 conv
conv1_1
conv1_2
depth=128
3x3 conv
conv2_1
conv2_2
depth=256
3x3 conv
conv3_1
conv3_2
conv3_3
conv3_4
depth=512
3x3 conv
conv4_1
conv4_2
conv4_3
conv4_4
depth=512
3x3 conv
conv5_1
conv5_2
conv5_3
conv5_4
size=4096
FC1
FC2
size=1000
softmax

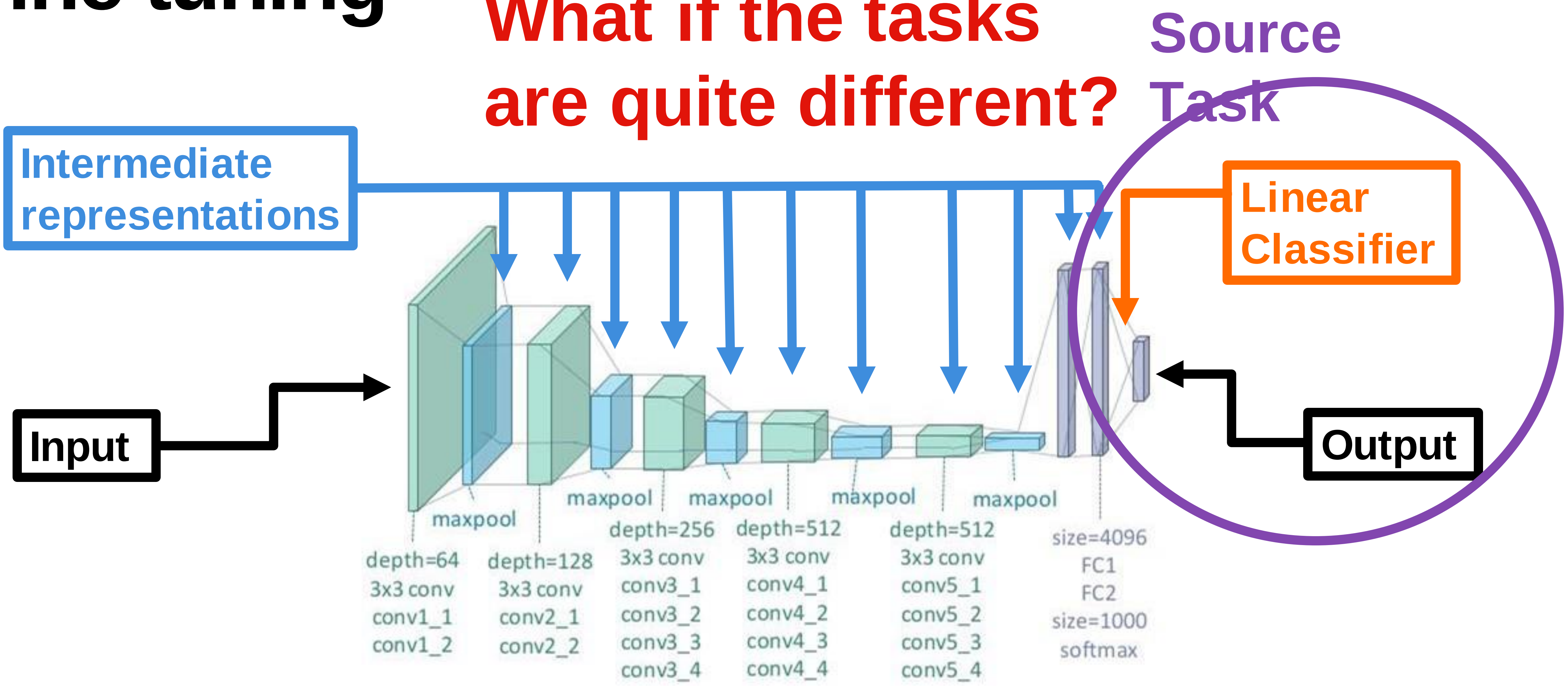
Fine tuning

What if the tasks are quite different?

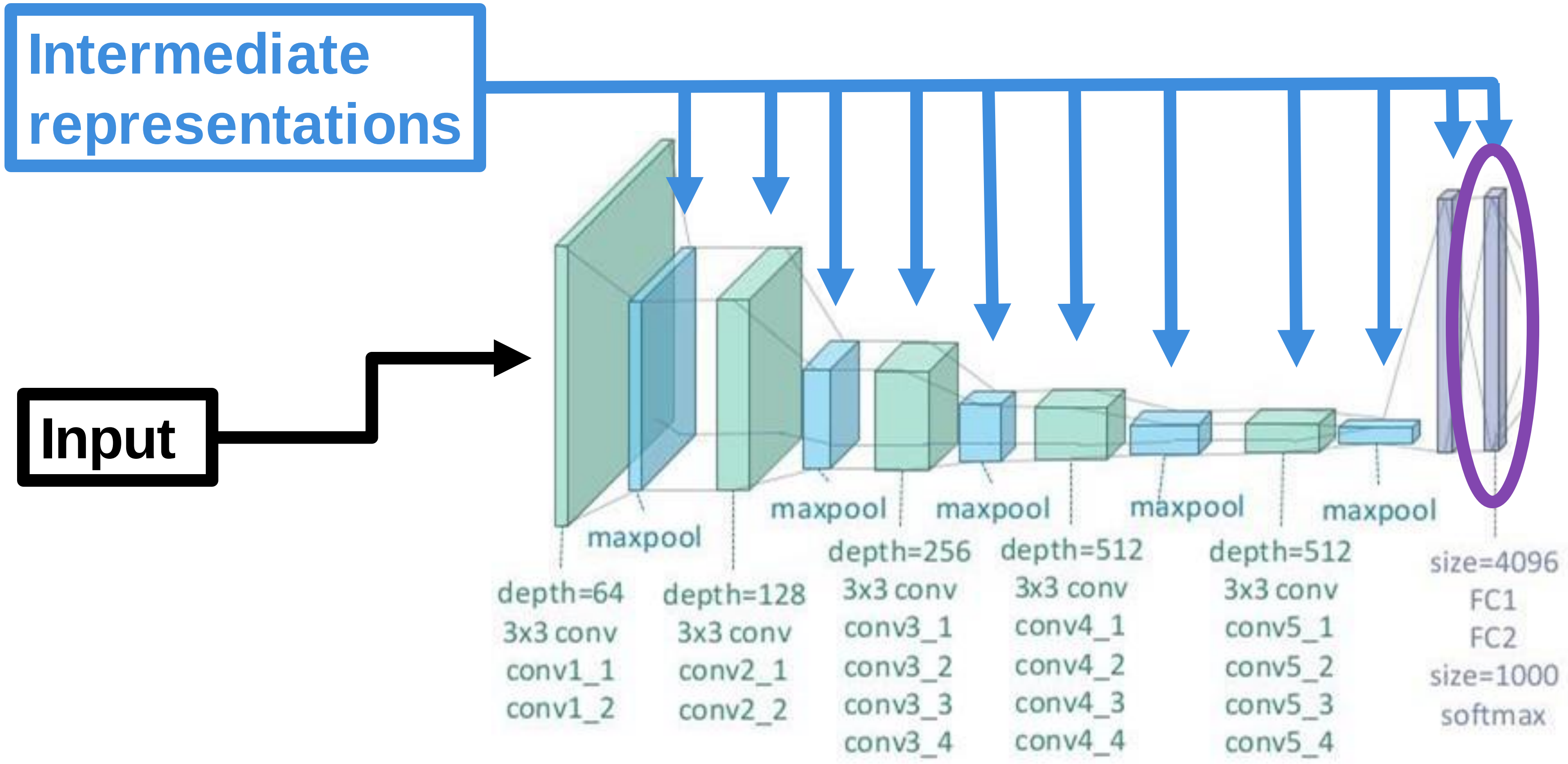


Fine tuning

What if the tasks are quite different?

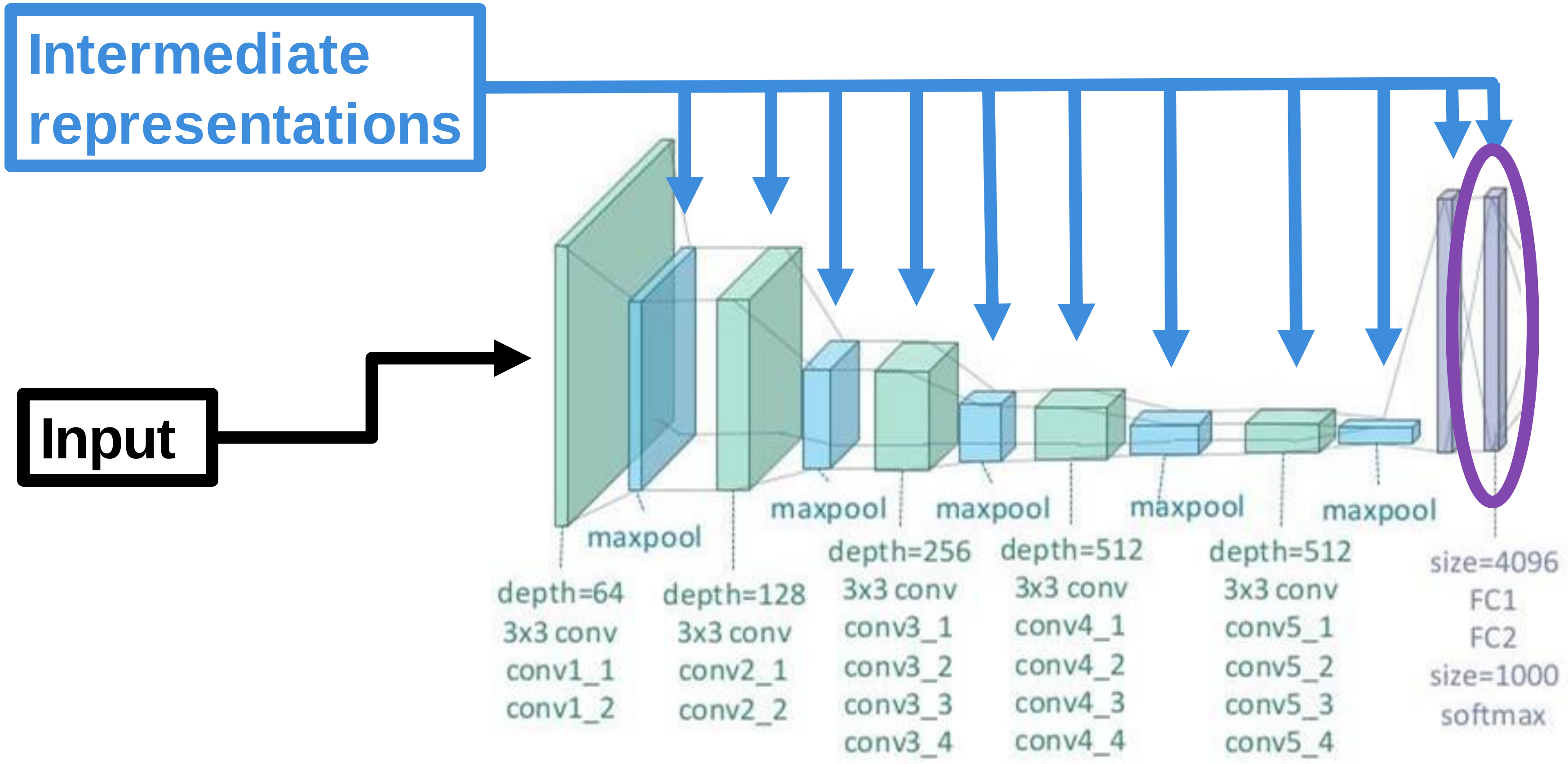


Fine tuning



Features learned for the **Source Task**

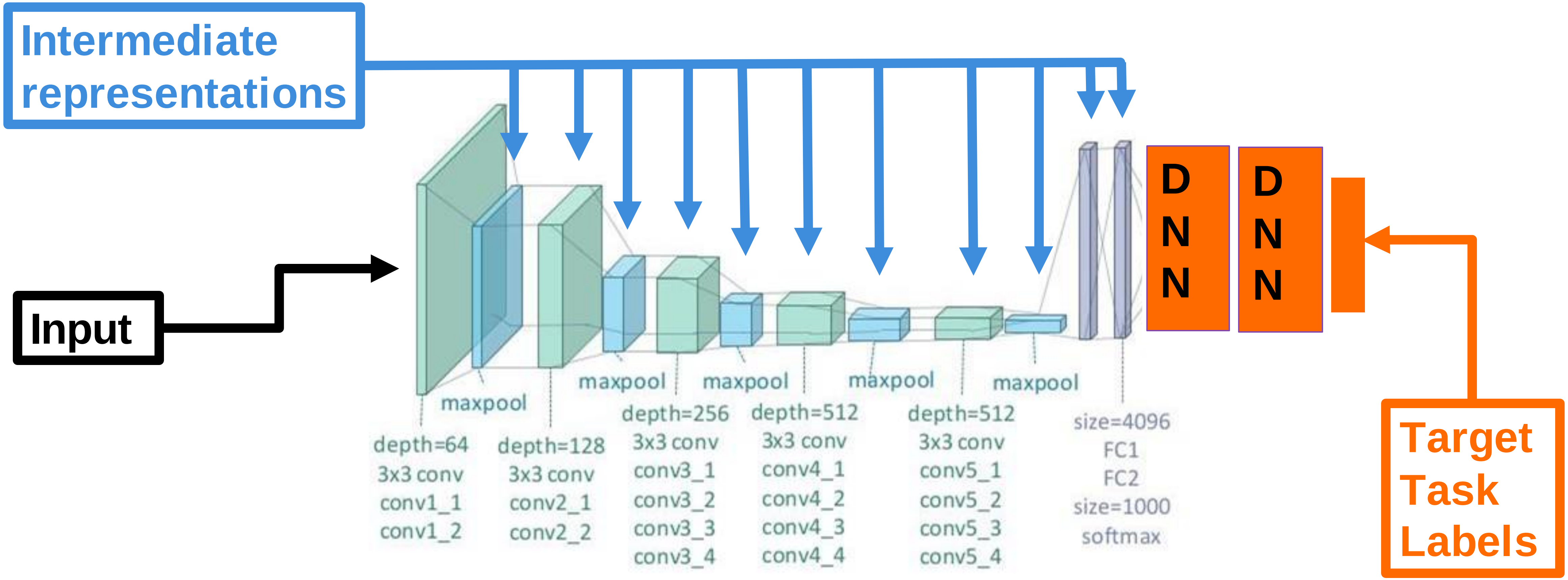
Fine tuning



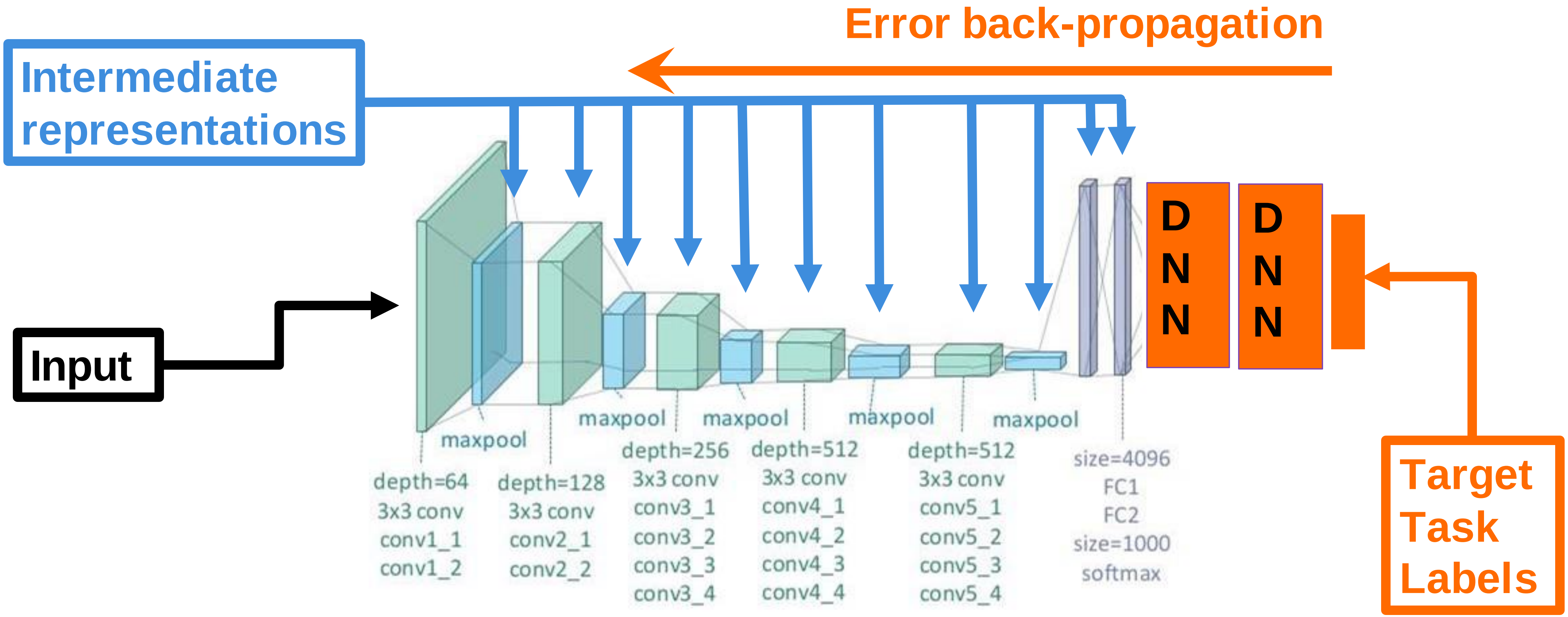
Features learned for the **Source Task**

Can we make them better?

Fine tuning

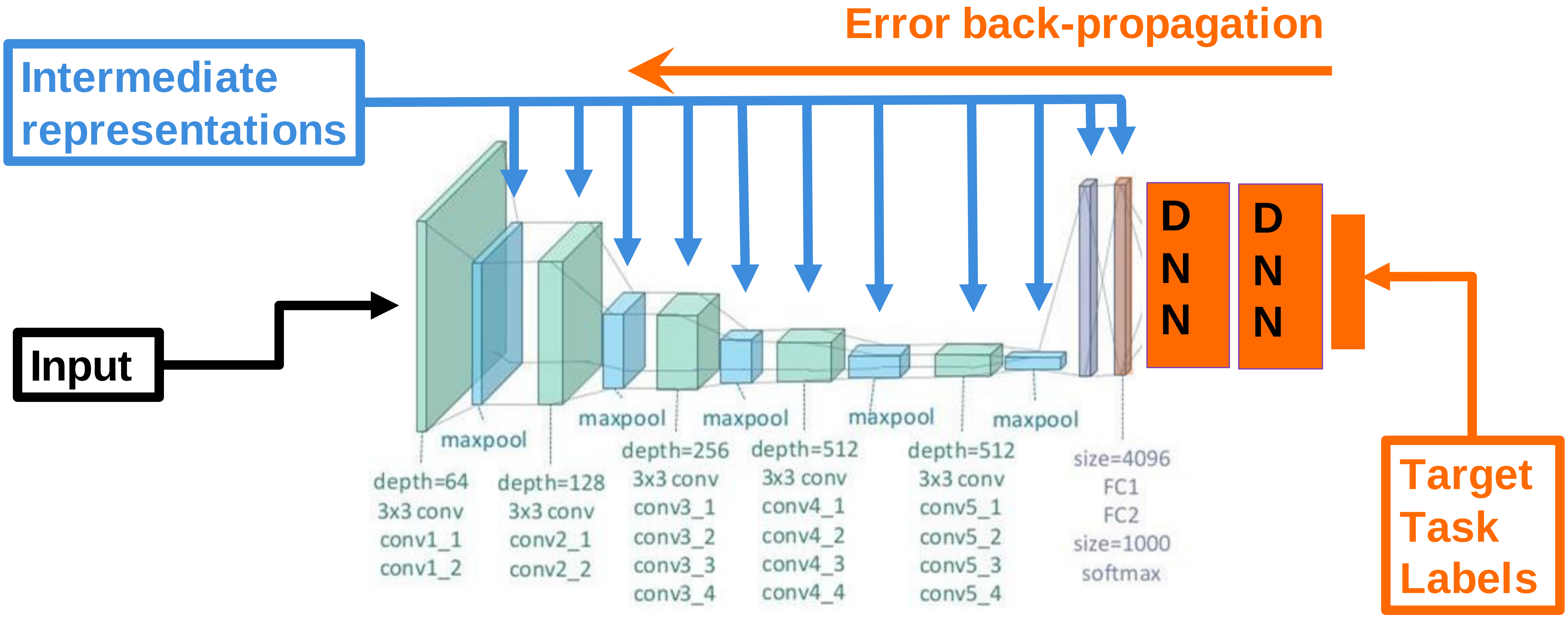

Intermediate representations
Input
maxpool
maxpool
maxpool
maxpool
maxpool
depth=64
3x3 conv
conv1_1
conv1_2
depth=128
3x3 conv
conv2_1
conv2_2
depth=256
3x3 conv
conv3_1
conv3_2
conv3_3
conv3_4
depth=512
3x3 conv
conv4_1
conv4_2
conv4_3
conv4_4
depth=512
3x3 conv
conv5_1
conv5_2
conv5_3
conv5_4
size=4096
FC1
FC2
size=1000
softmax
DNN
DNN
Target Task Labels

Fine tuning

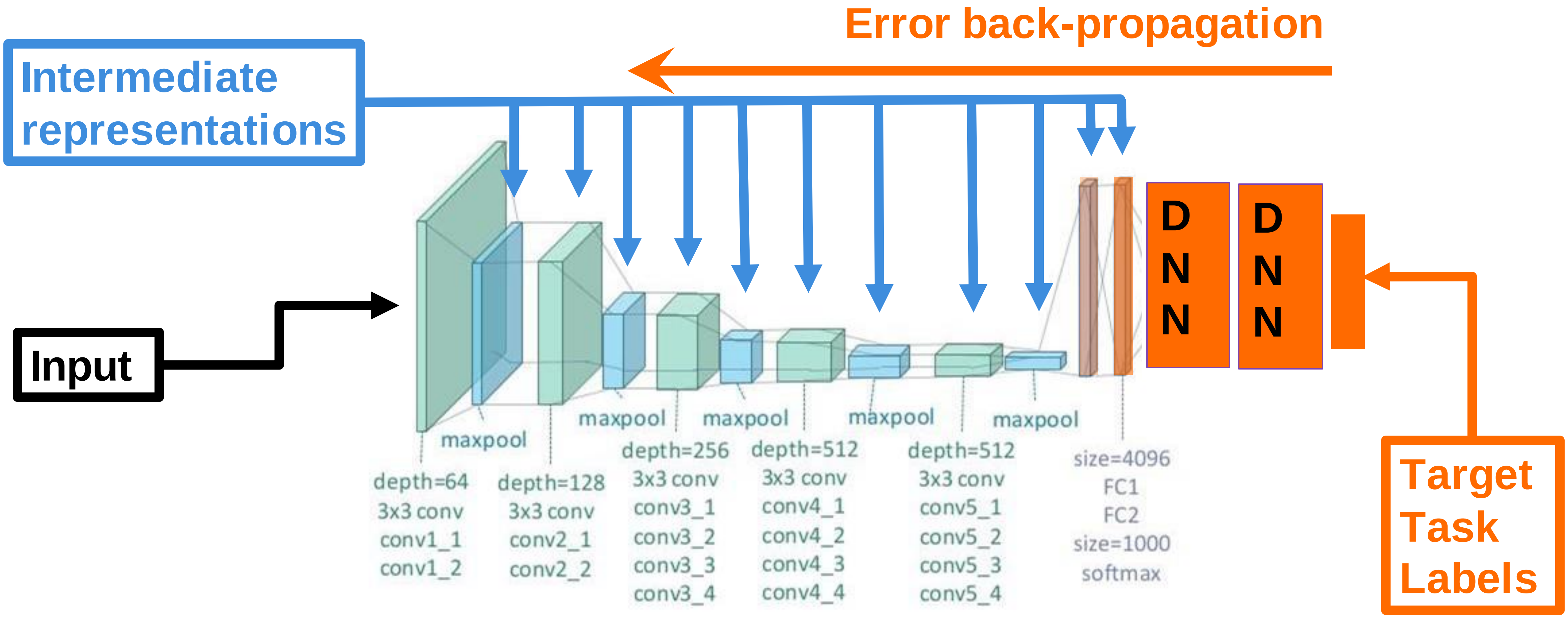


Error back-propagation
Intermediate representations
Input
maxpool
maxpool
maxpool
maxpool
maxpool
depth=64
3x3 conv
conv1_1
conv1_2
depth=128
3x3 conv
conv2_1
conv2_2
depth=256
3x3 conv
conv3_1
conv3_2
conv3_3
conv3_4
depth=512
3x3 conv
conv4_1
conv4_2
conv4_3
conv4_4
depth=512
3x3 conv
conv5_1
conv5_2
conv5_3
conv5_4
size=4096
FC1
FC2
size=1000
softmax
DNN
DNN
Target Task Labels

Fine tuning

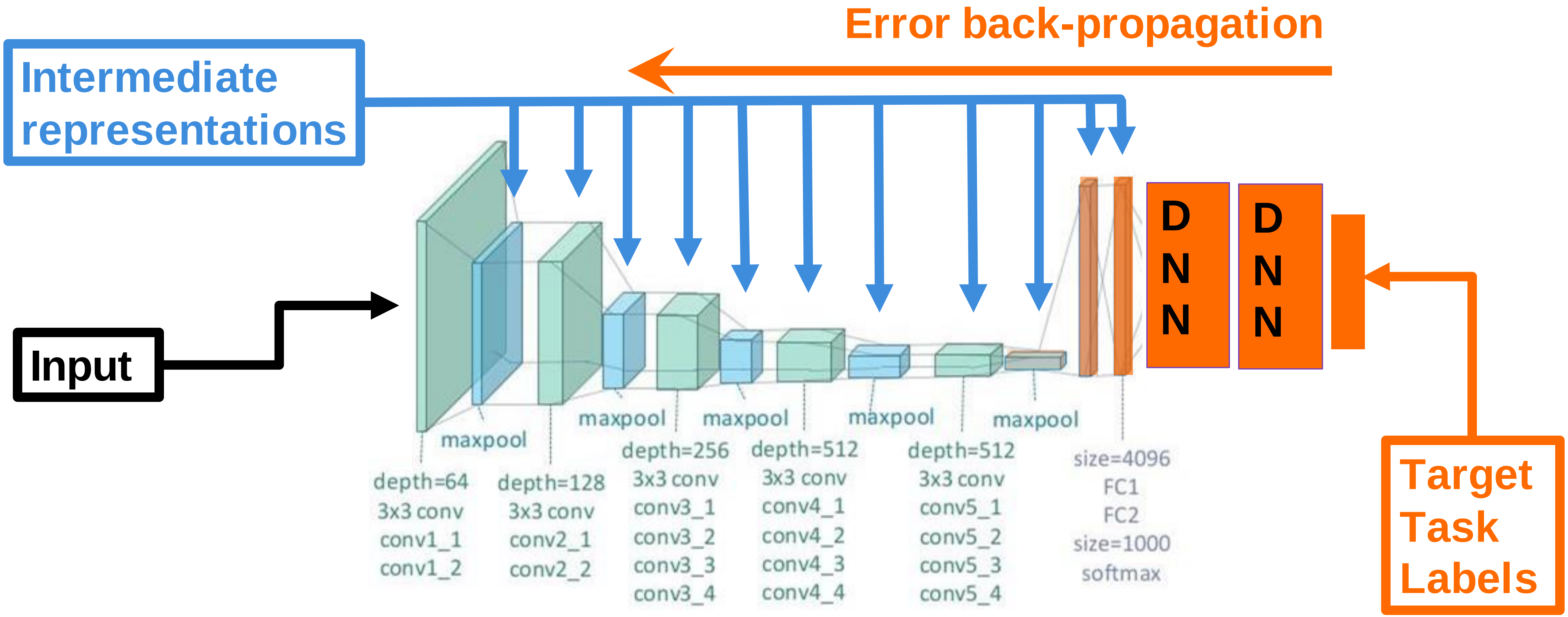

Error back-propagation
Intermediate representations
Input
maxpool
maxpool
maxpool
maxpool
maxpool
depth=64
3x3 conv
conv1_1
conv1_2
depth=128
3x3 conv
conv2_1
conv2_2
depth=256
3x3 conv
conv3_1
conv3_2
conv3_3
conv3_4
depth=512
3x3 conv
conv4_1
conv4_2
conv4_3
conv4_4
depth=512
3x3 conv
conv5_1
conv5_2
conv5_3
conv5_4
size=4096
FC1
FC2
size=1000
softmax
DNN
DNN
Target Task Labels

Fine tuning

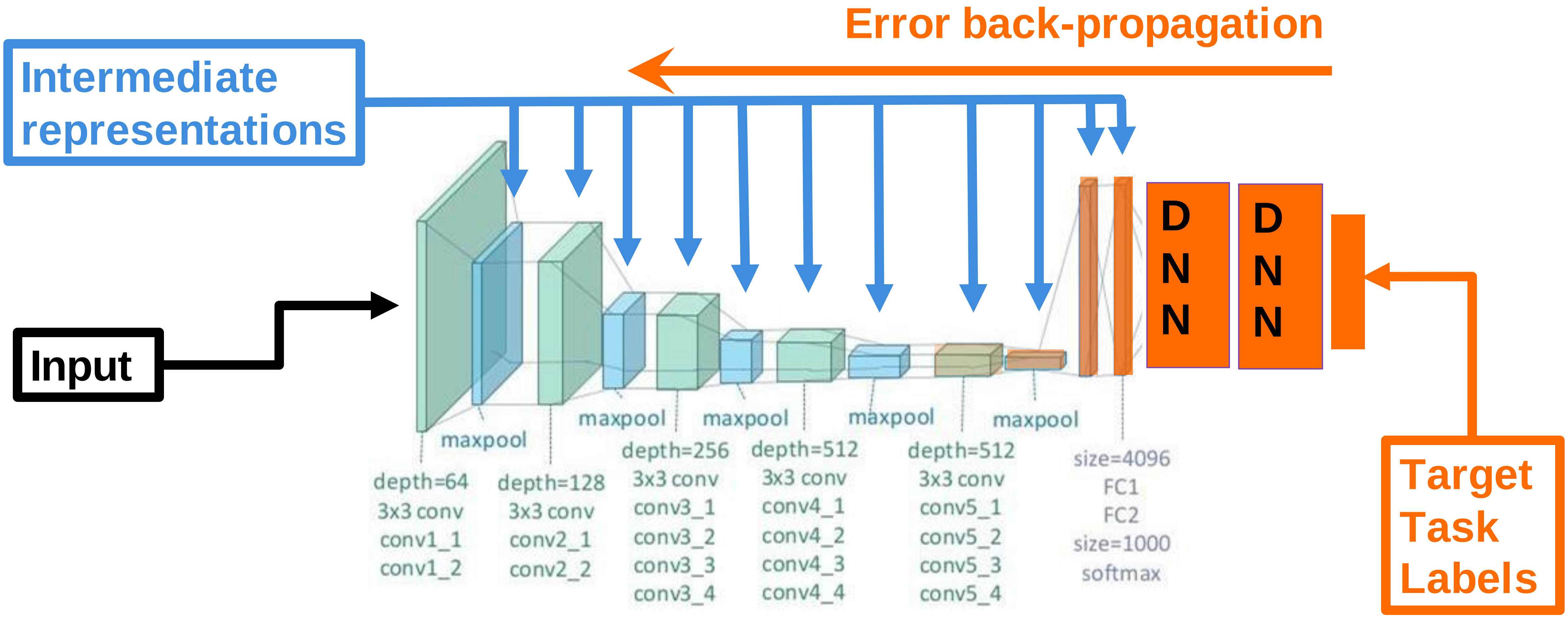


Error back-propagation
Intermediate representations
Input
maxpool
maxpool
maxpool
maxpool
maxpool
depth=64
3x3 conv
conv1_1
conv1_2
depth=128
3x3 conv
conv2_1
conv2_2
depth=256
3x3 conv
conv3_1
conv3_2
conv3_3
conv3_4
depth=512
3x3 conv
conv4_1
conv4_2
conv4_3
conv4_4
depth=512
3x3 conv
conv5_1
conv5_2
conv5_3
conv5_4
size=4096
FC1
FC2
size=1000
softmax
DNN
DNN
Target Task Labels

Fine tuning

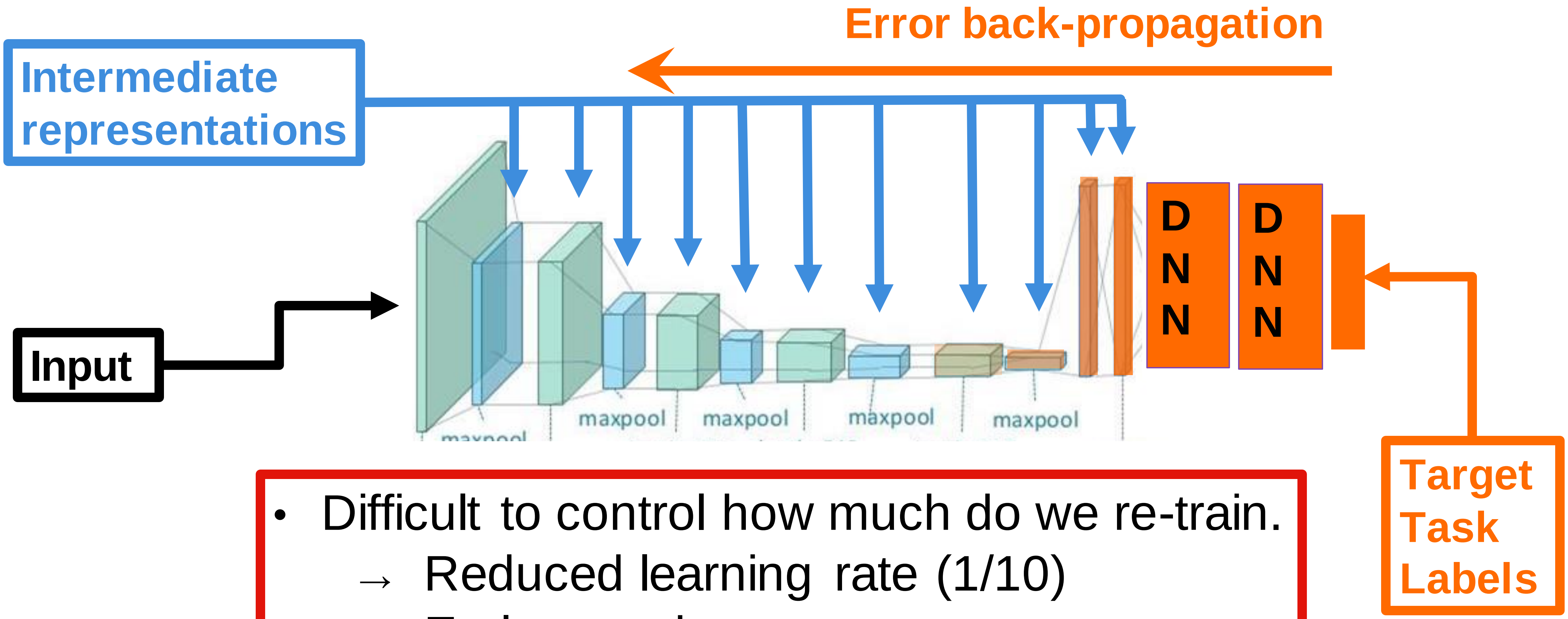

Error back-propagation
Intermediate representations
Input
maxpool
maxpool
maxpool
maxpool
maxpool
depth=64
3x3 conv
conv1_1
conv1_2
depth=128
3x3 conv
conv2_1
conv2_2
depth=256
3x3 conv
conv3_1
conv3_2
conv3_3
conv3_4
depth=512
3x3 conv
conv4_1
conv4_2
conv4_3
conv4_4
depth=512
3x3 conv
conv5_1
conv5_2
conv5_3
conv5_4
size=4096
FC1
FC2
size=1000
softmax
DNN
DNN
Target Task Labels

Fine tuning



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

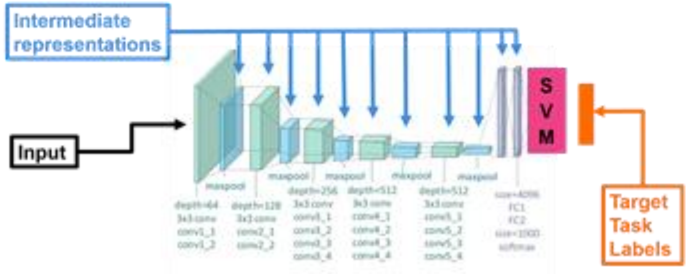
Fine tuning



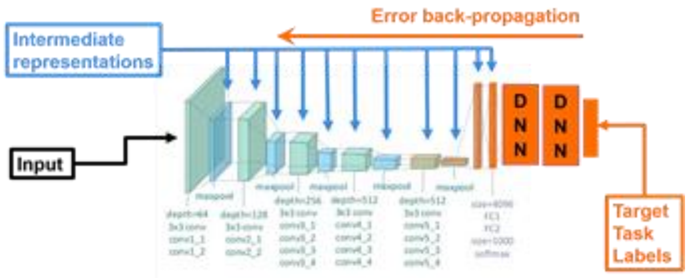
- Difficult to control how much do we re-train.
  - → Reduced learning rate (1/10)
  - → Early stopping
  - → Alternate source/target sampling

Simple solutions

- DNN last layer features + SVM
  (Feature extraction)
  - Similar task and domain



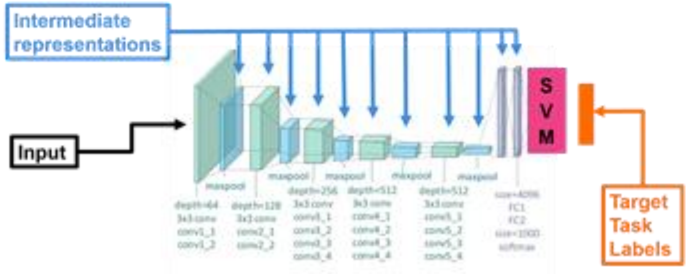
- Add one or several NN layers +
  Fine-tuning pre-trained layers

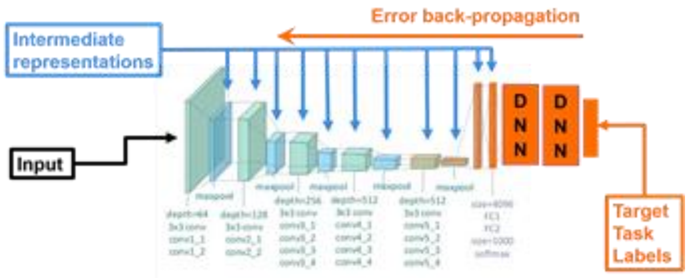


Simple solutions

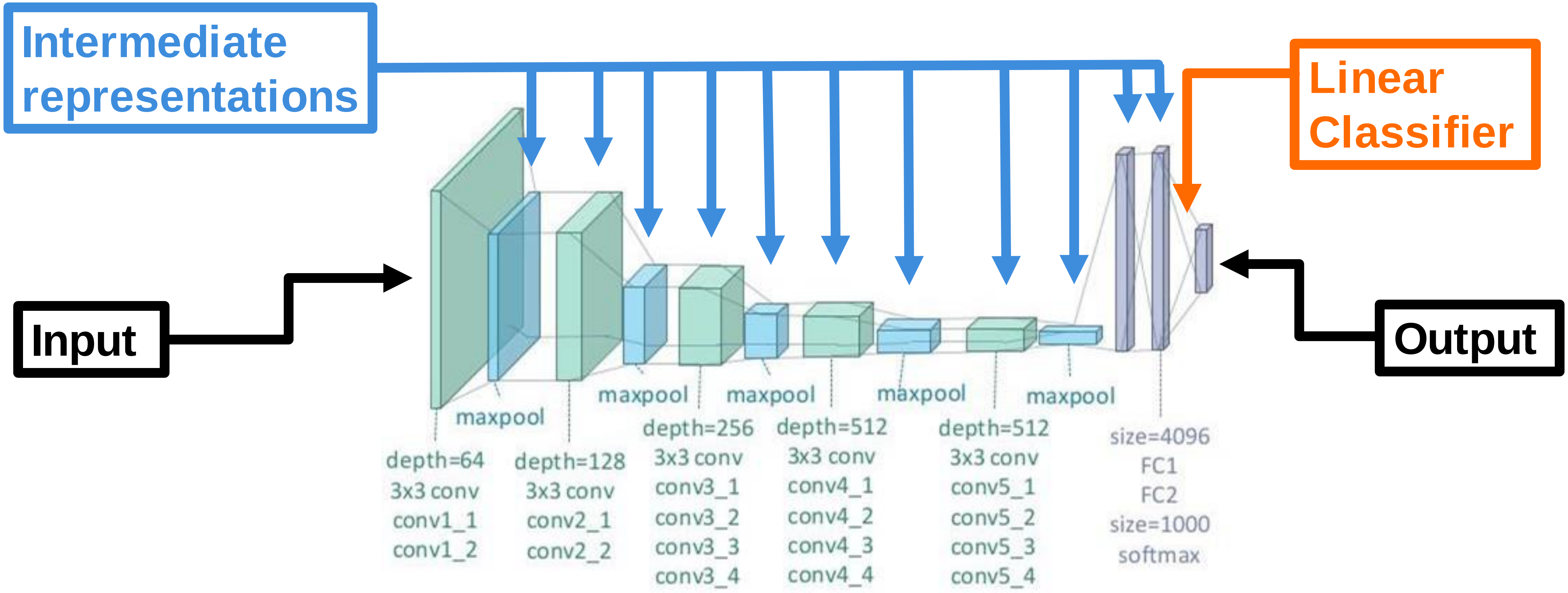
- DNN last layer features + SVM (Feature extraction)
  - **Similar task and domain**

- Add one or several NN layers + Fine-tuning pre-trained layers
  - **Enough data**

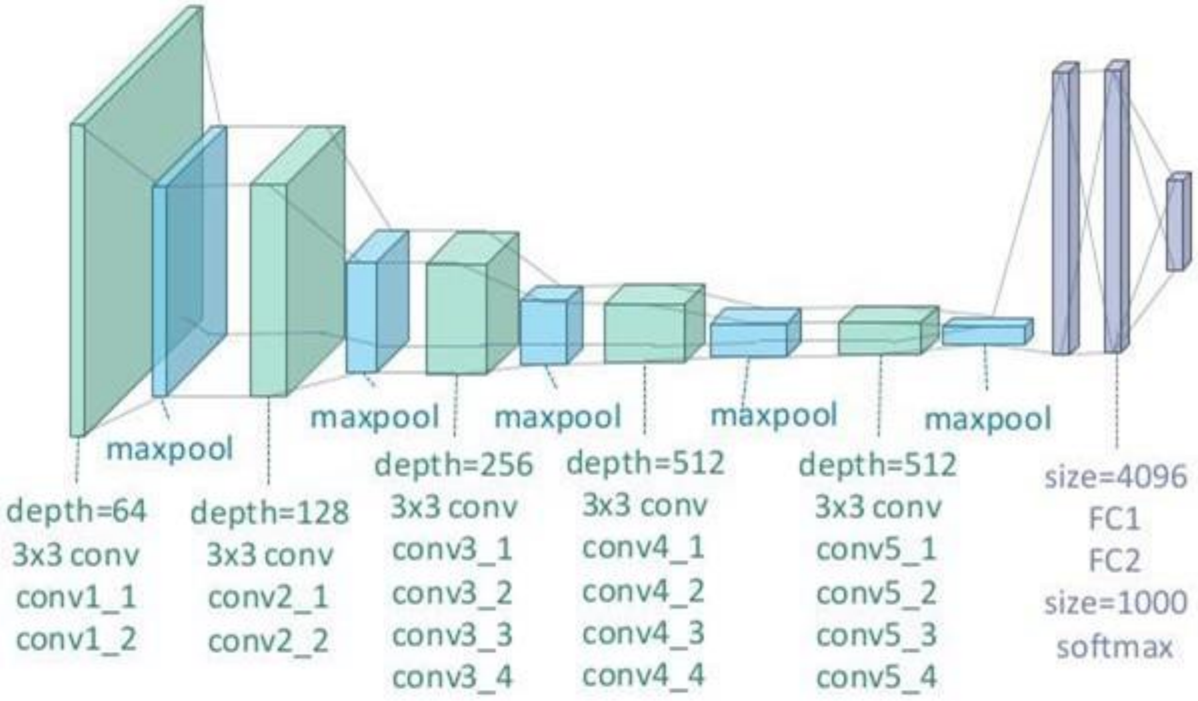




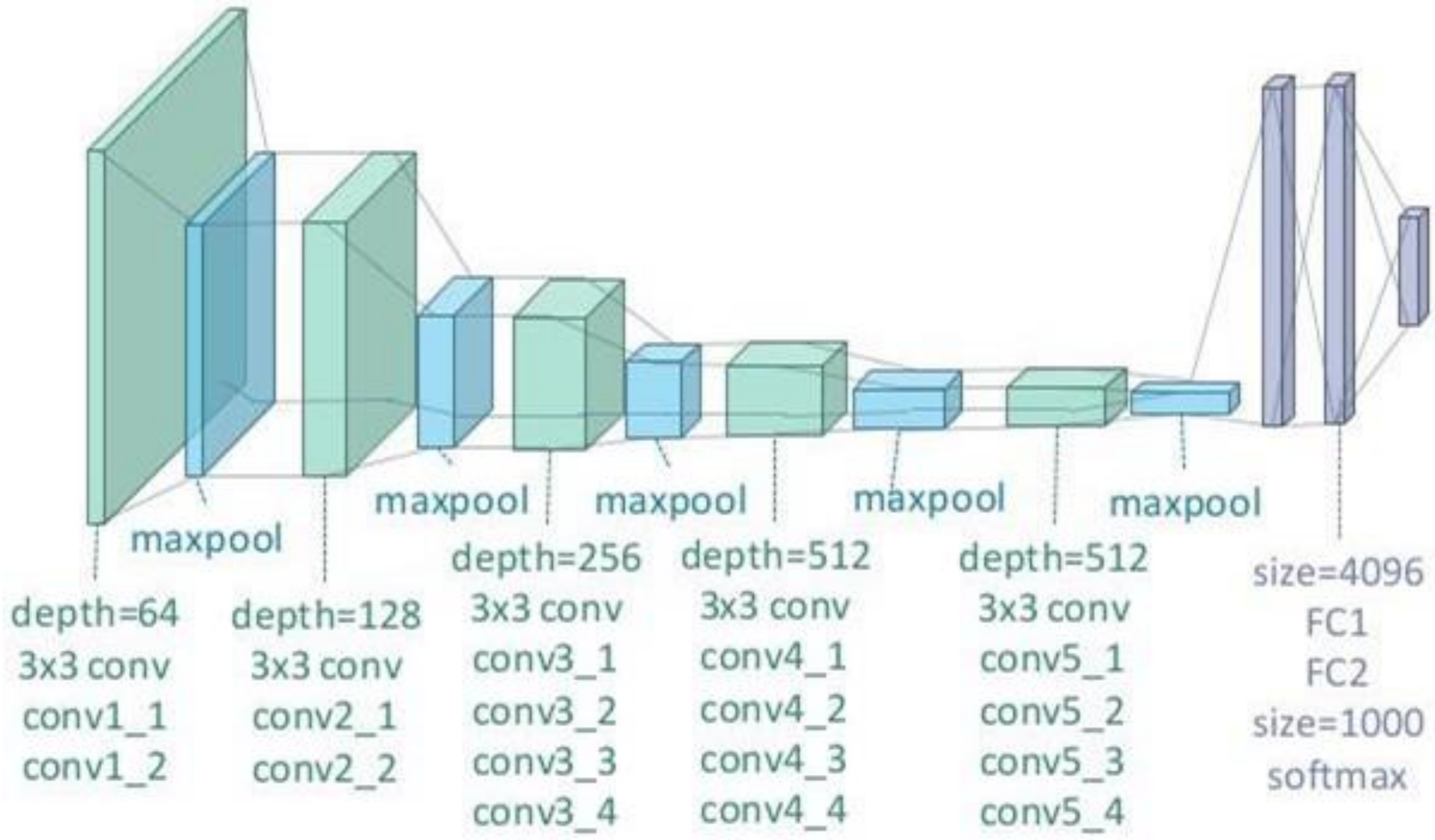
Beyond the last layer



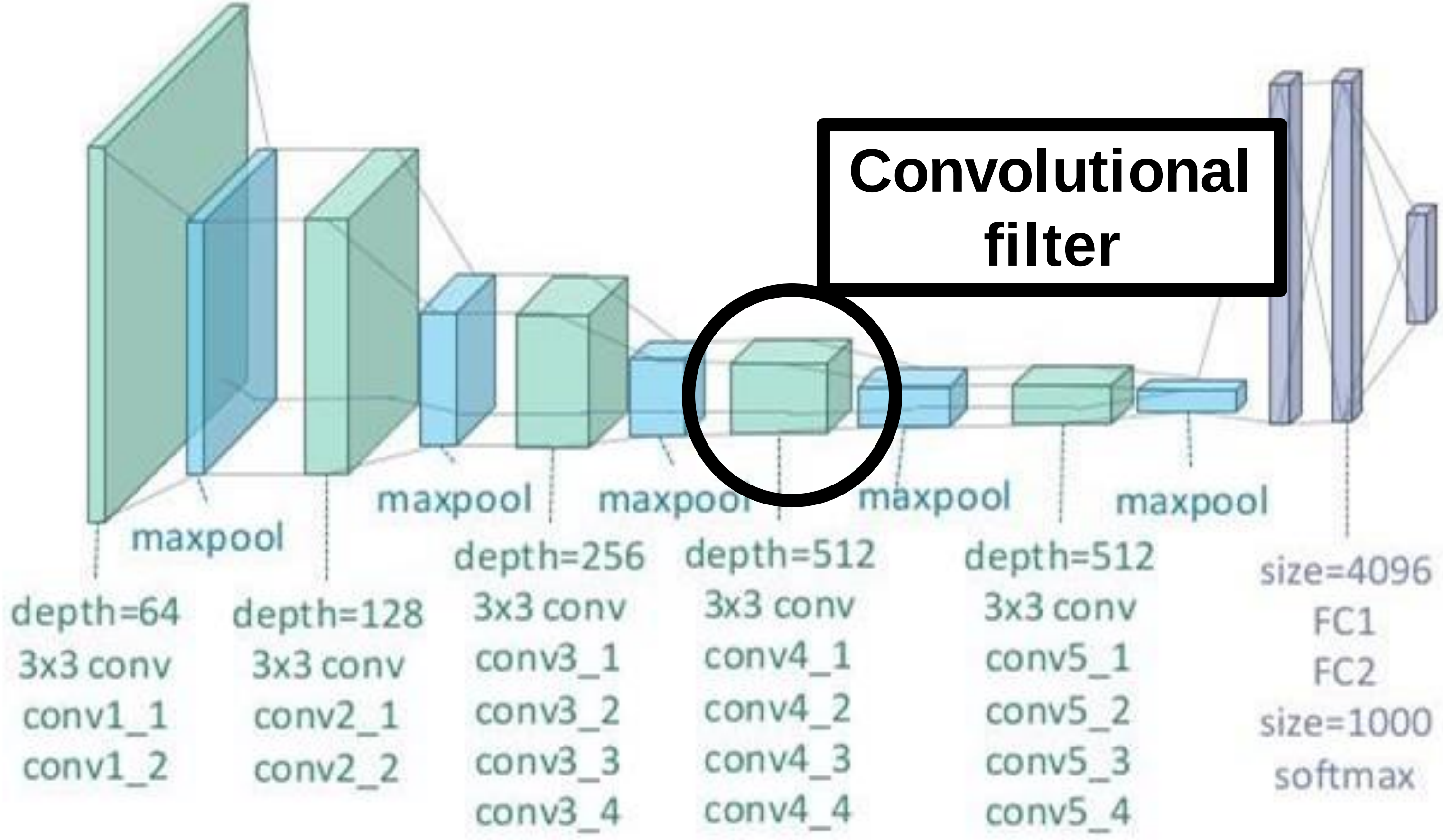
Knowledge inside DNN



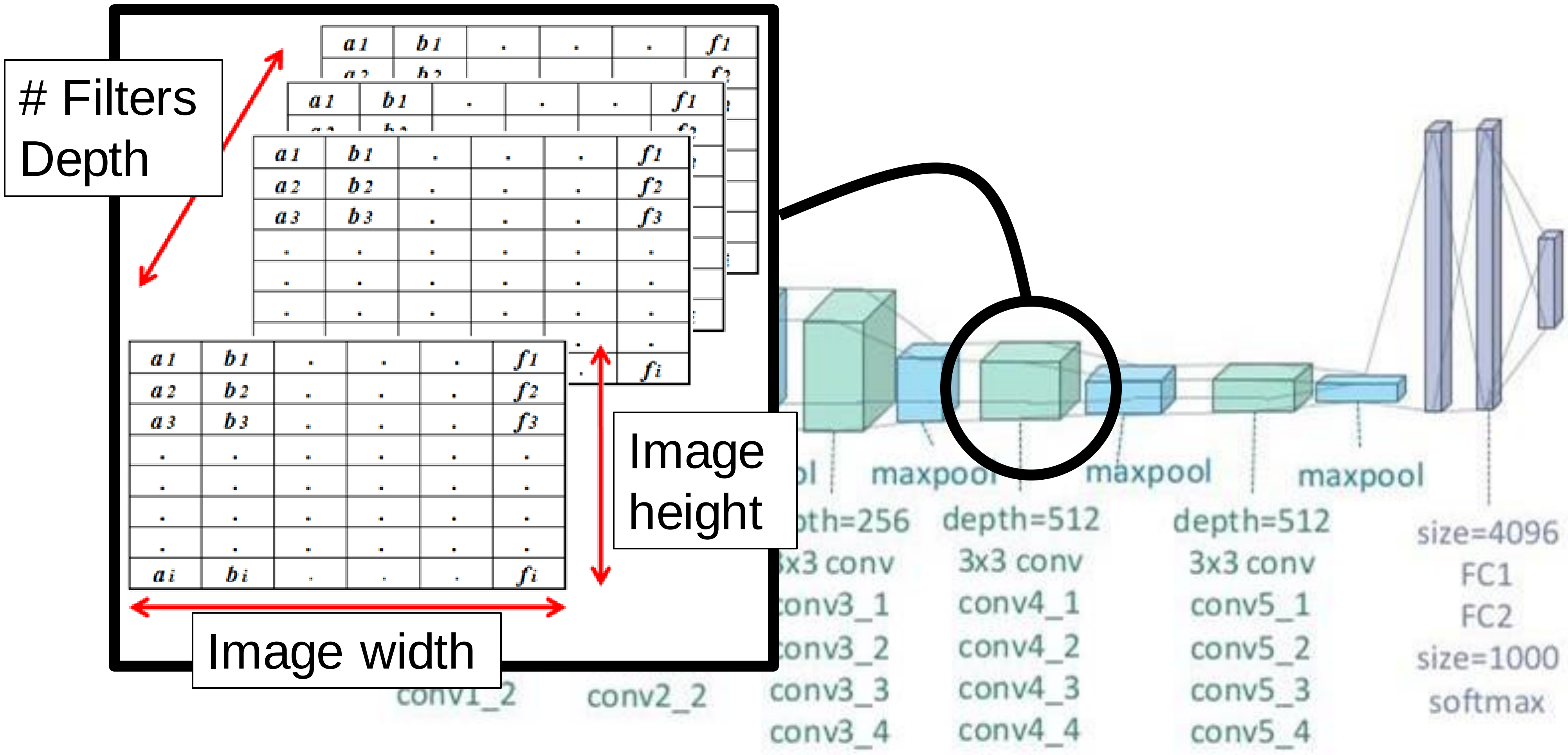
Knowledge inside DNN



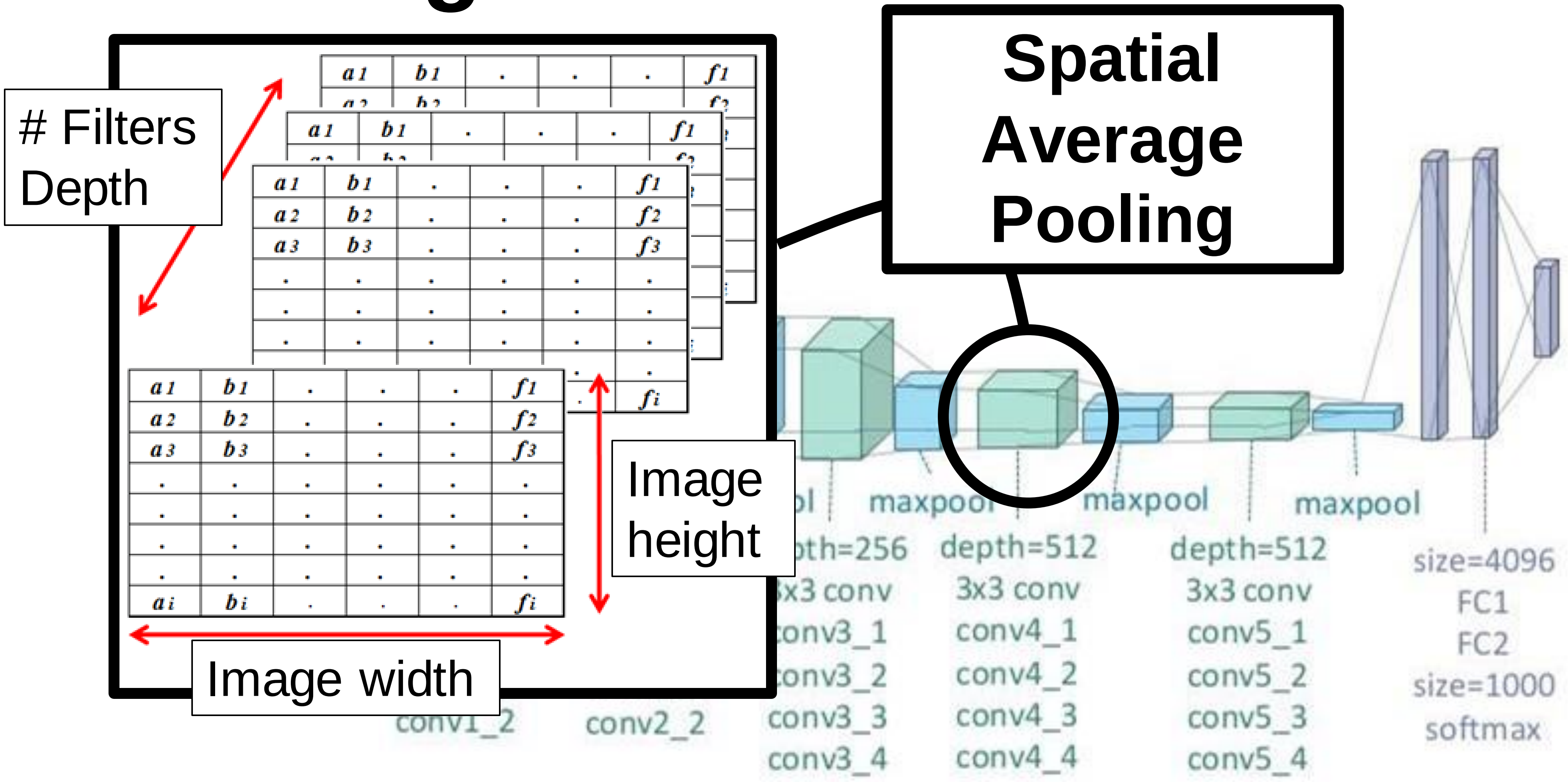
Knowledge inside DNN



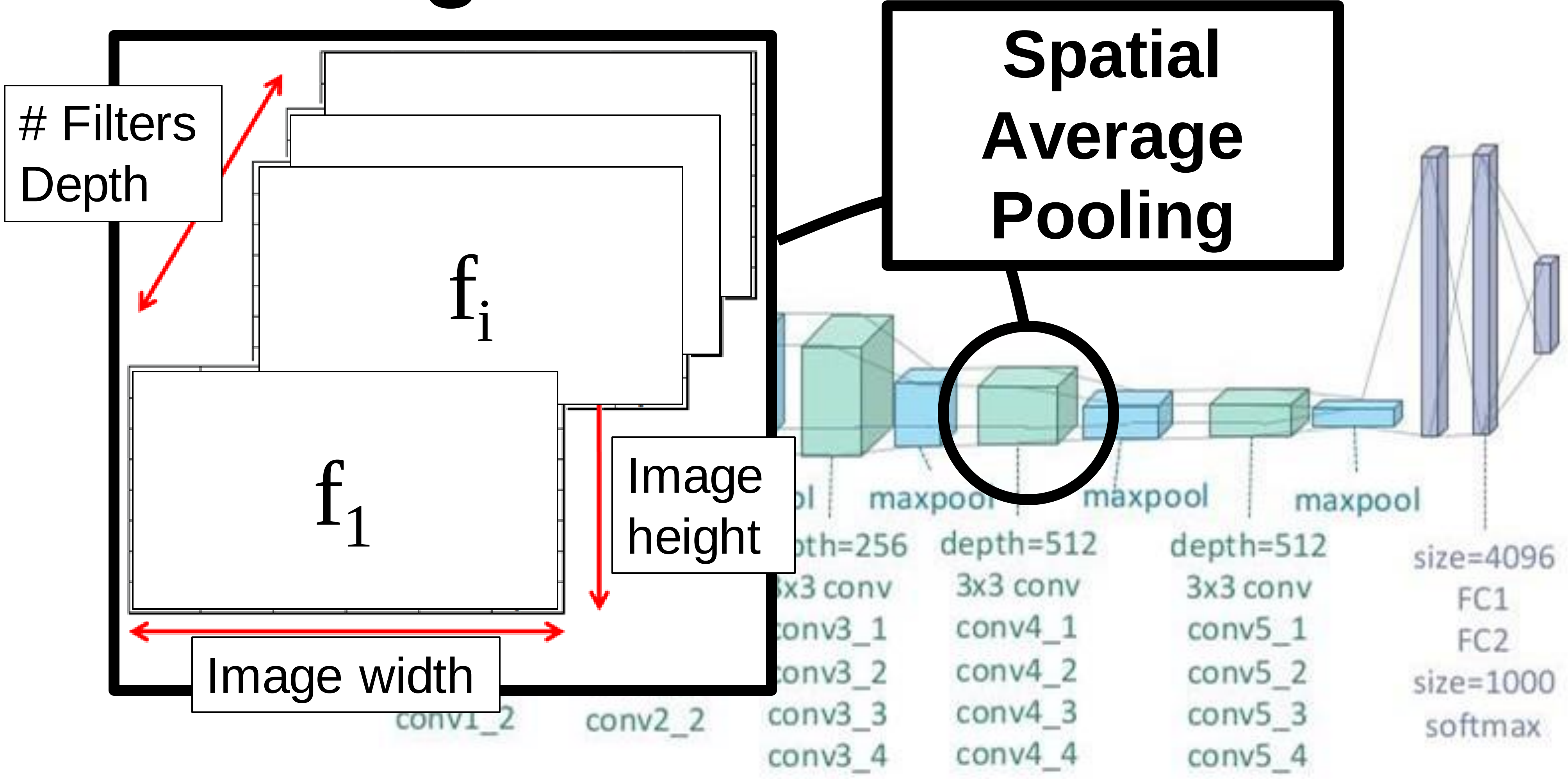
Knowledge inside DNN



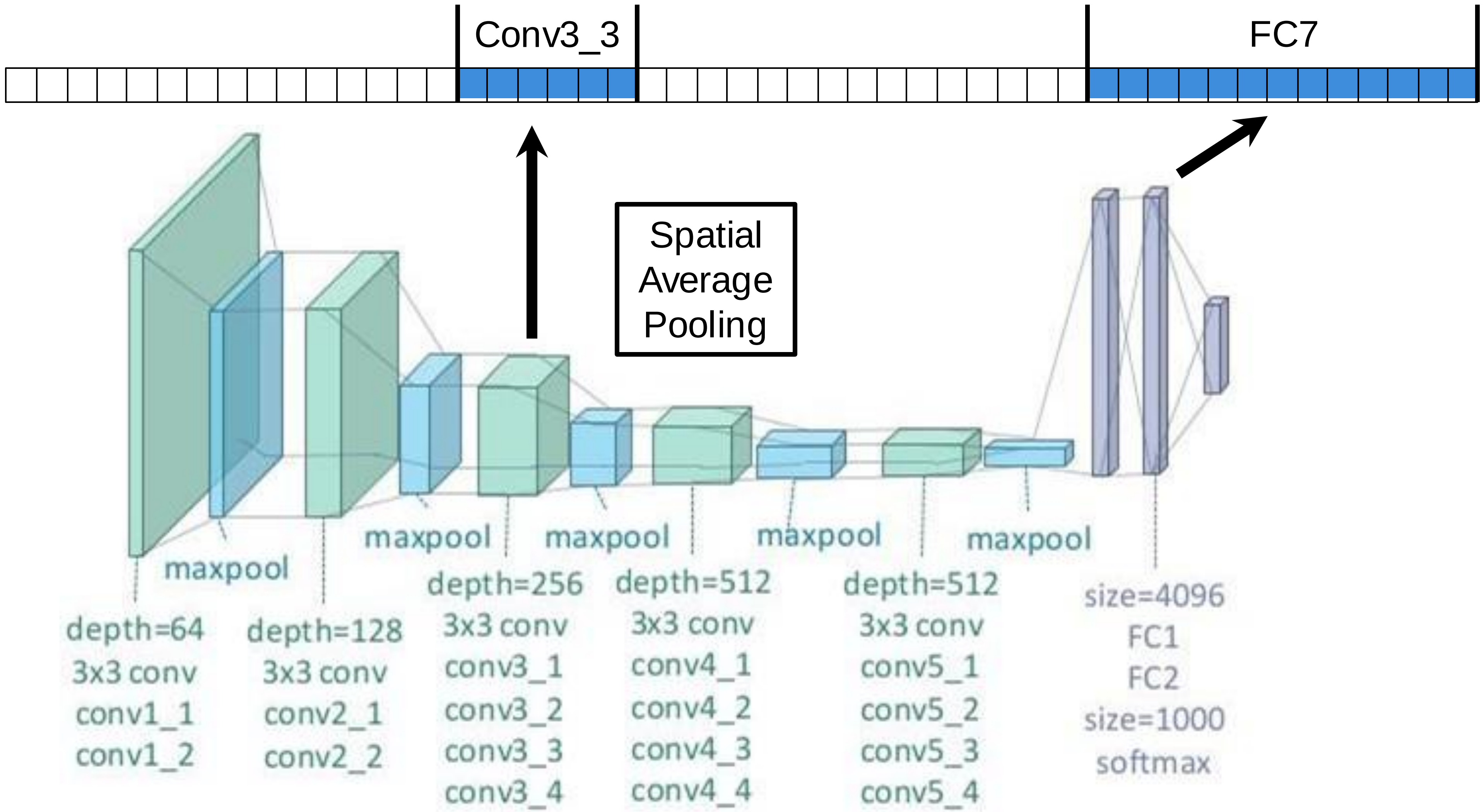
Knowledge inside DNN



Knowledge inside DNN

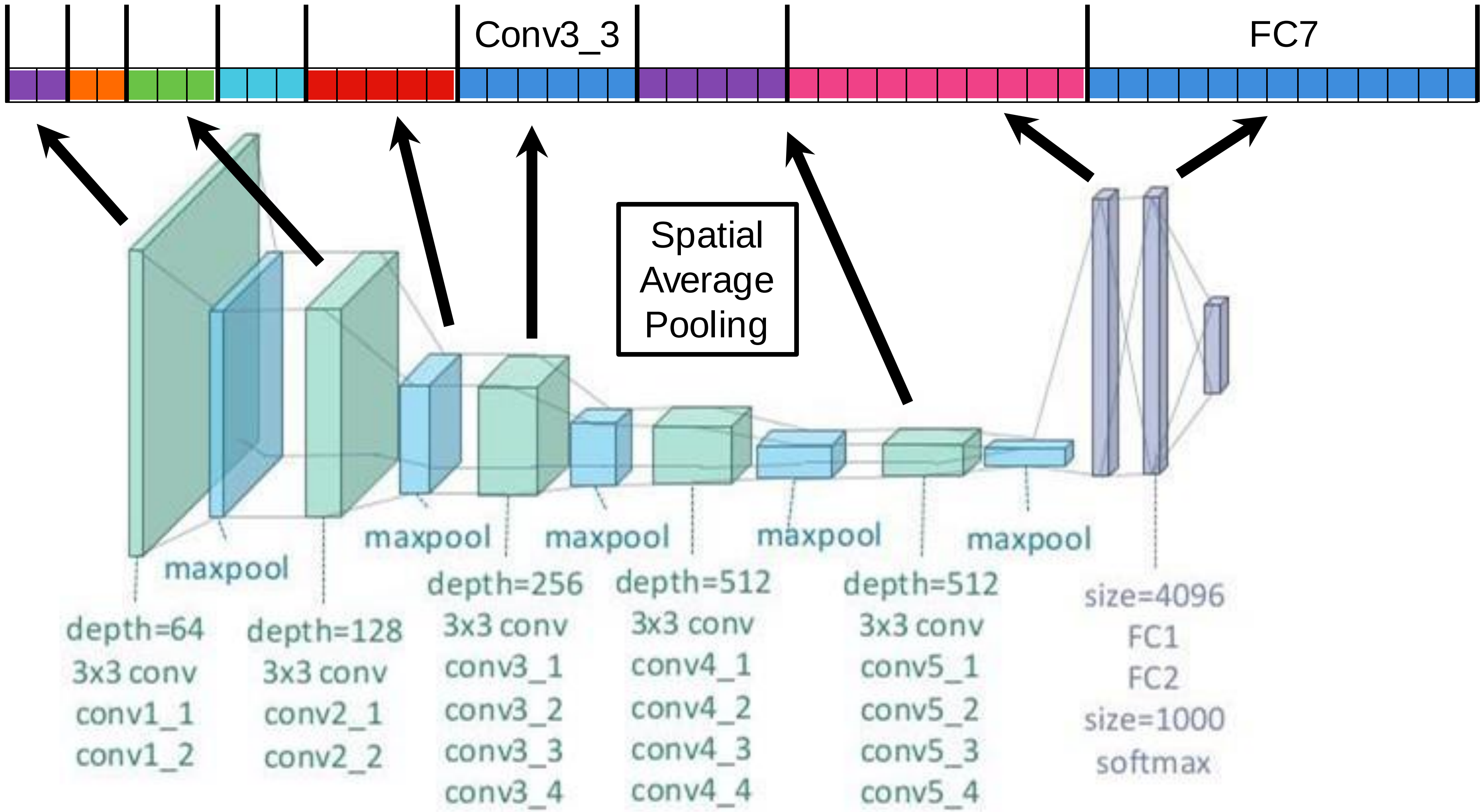


Knowledge inside DNN

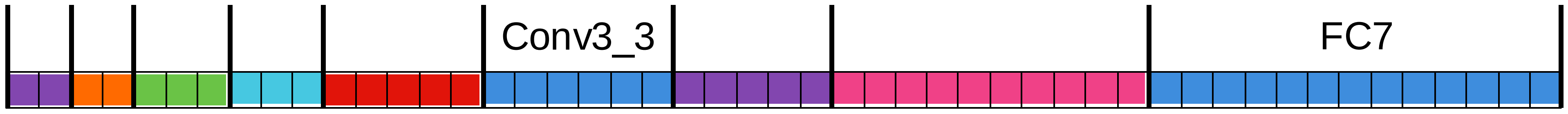


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

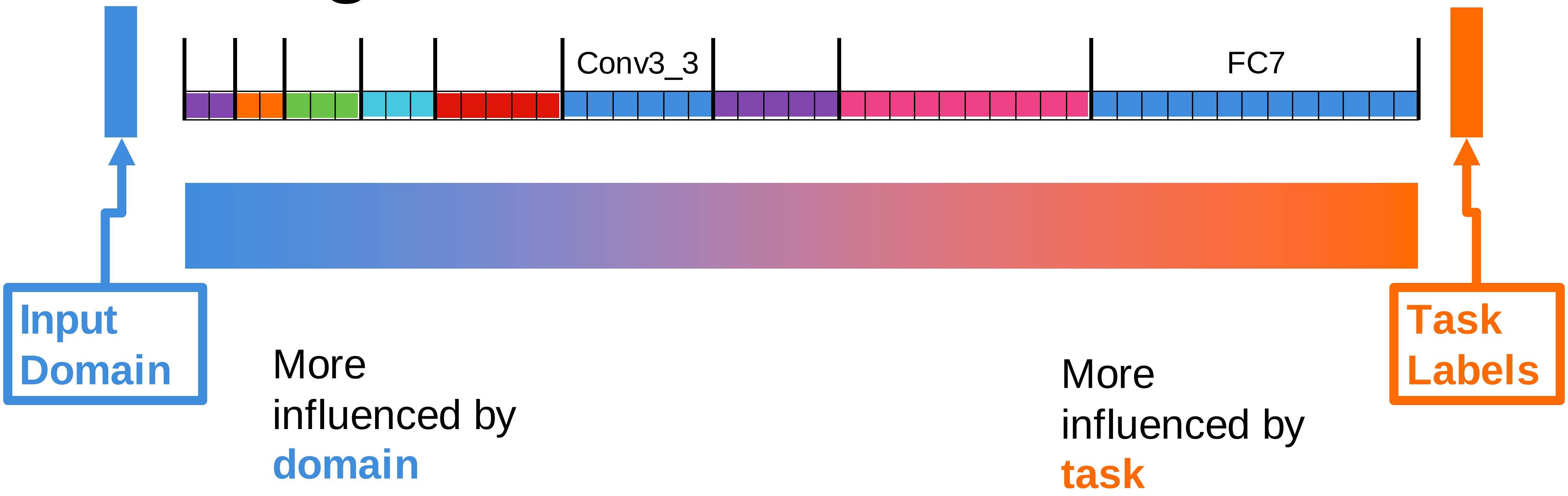
Knowledge inside DNN



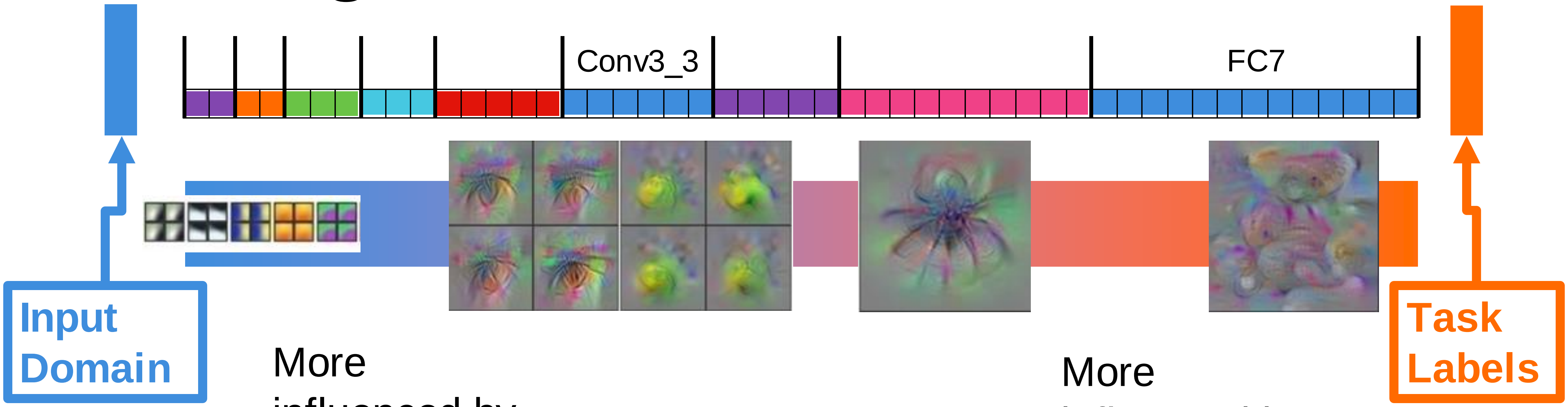
Knowledge inside DNN



Knowledge inside DNN



Knowledge inside DNN



More influenced by **domain**

Simple features

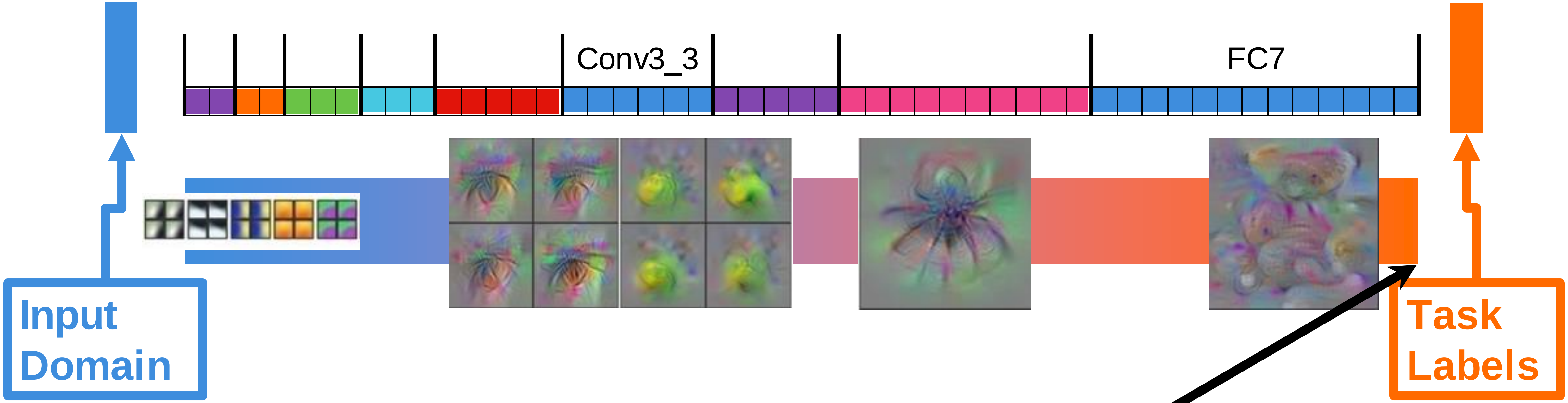
More influenced by **task**

Complex features

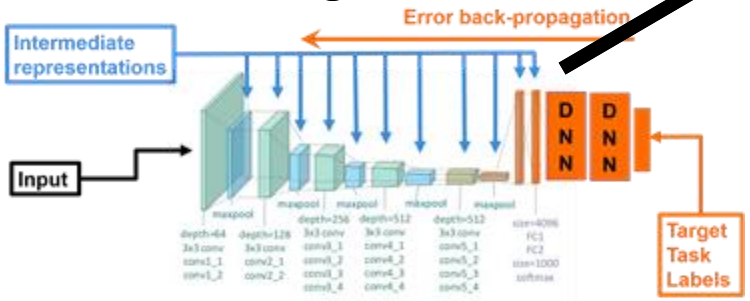
Visualizations from: Yosinski, Jason, et al. "Understanding neural networks through deep visualization." *arXiv preprint arXiv:1506.06579* (2015).

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

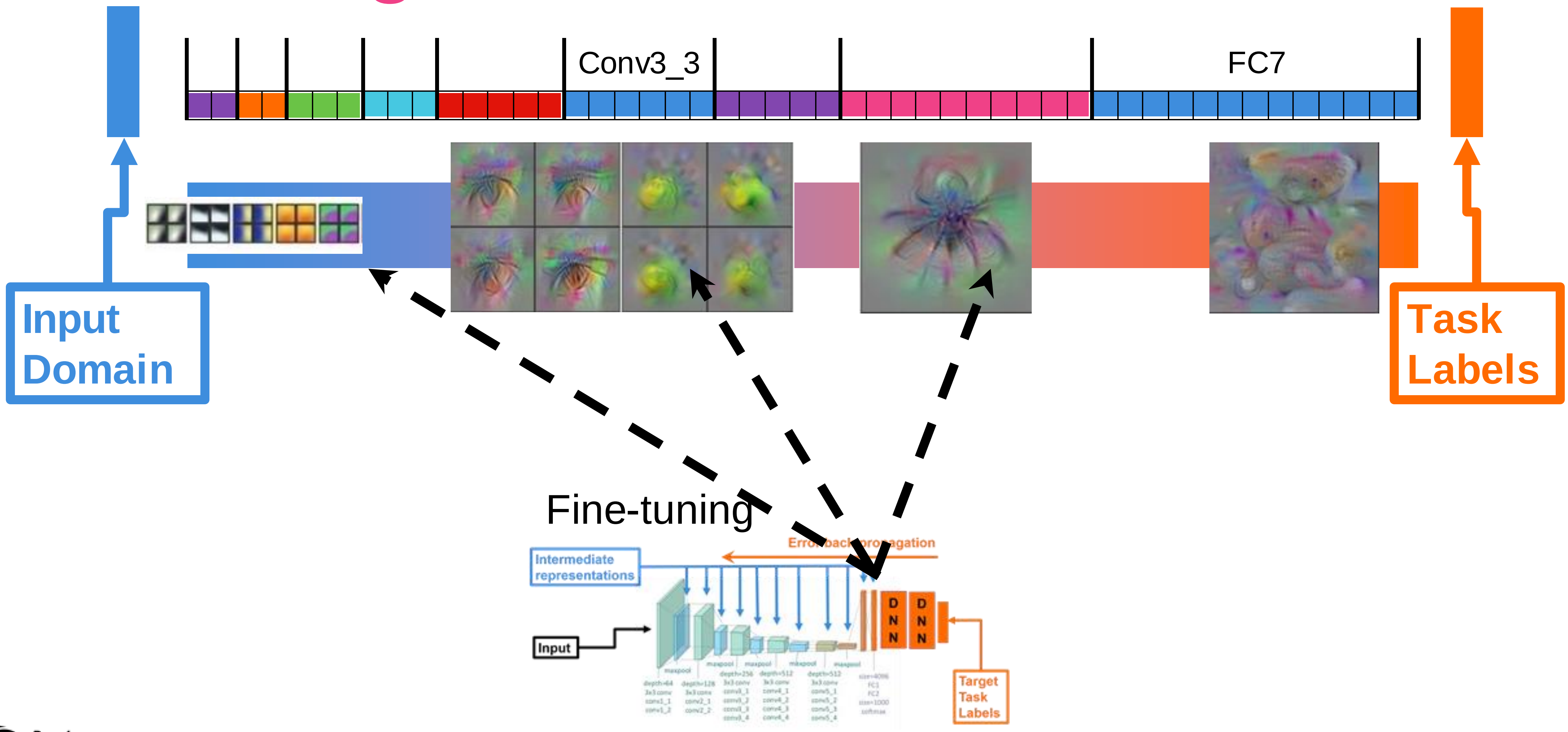
Fine tuning inside DNN?



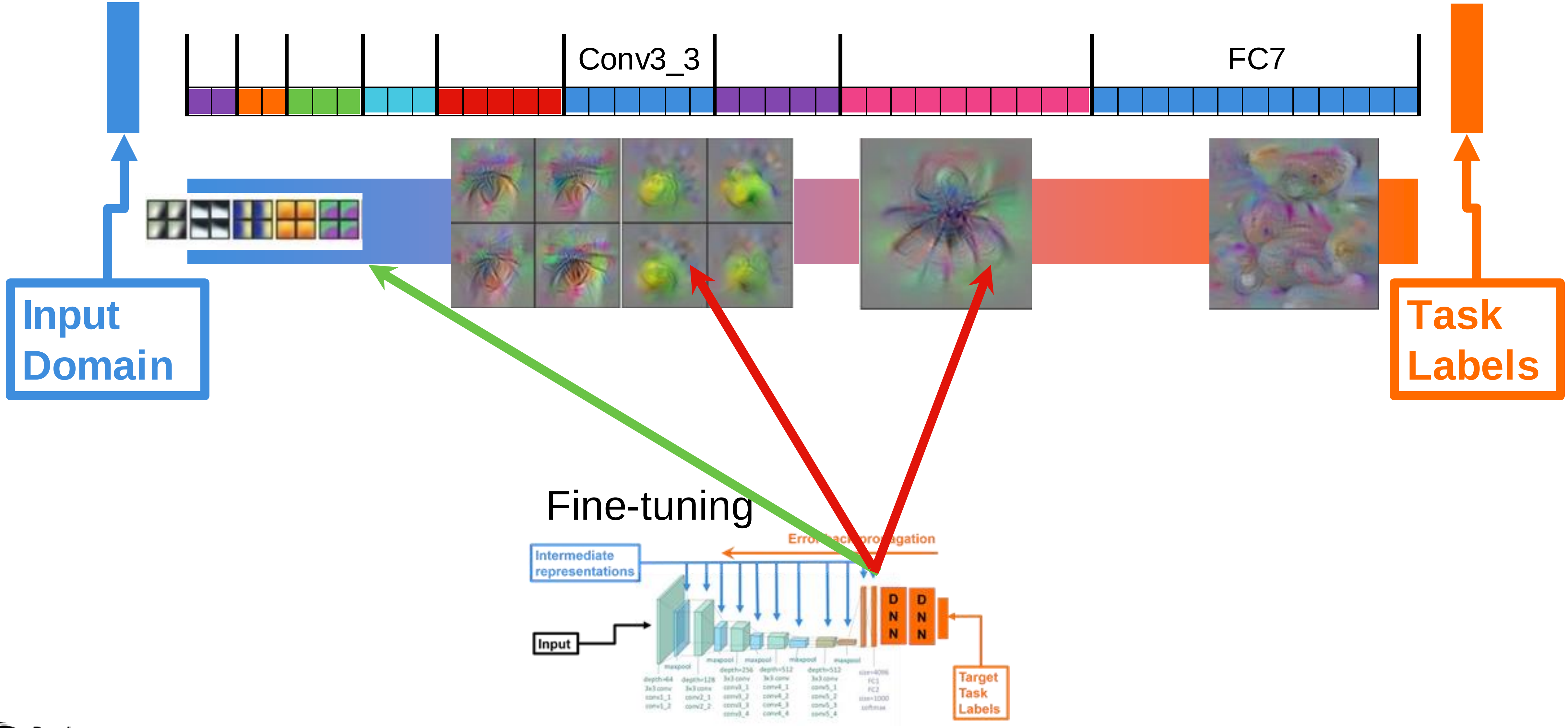
Fine-tuning



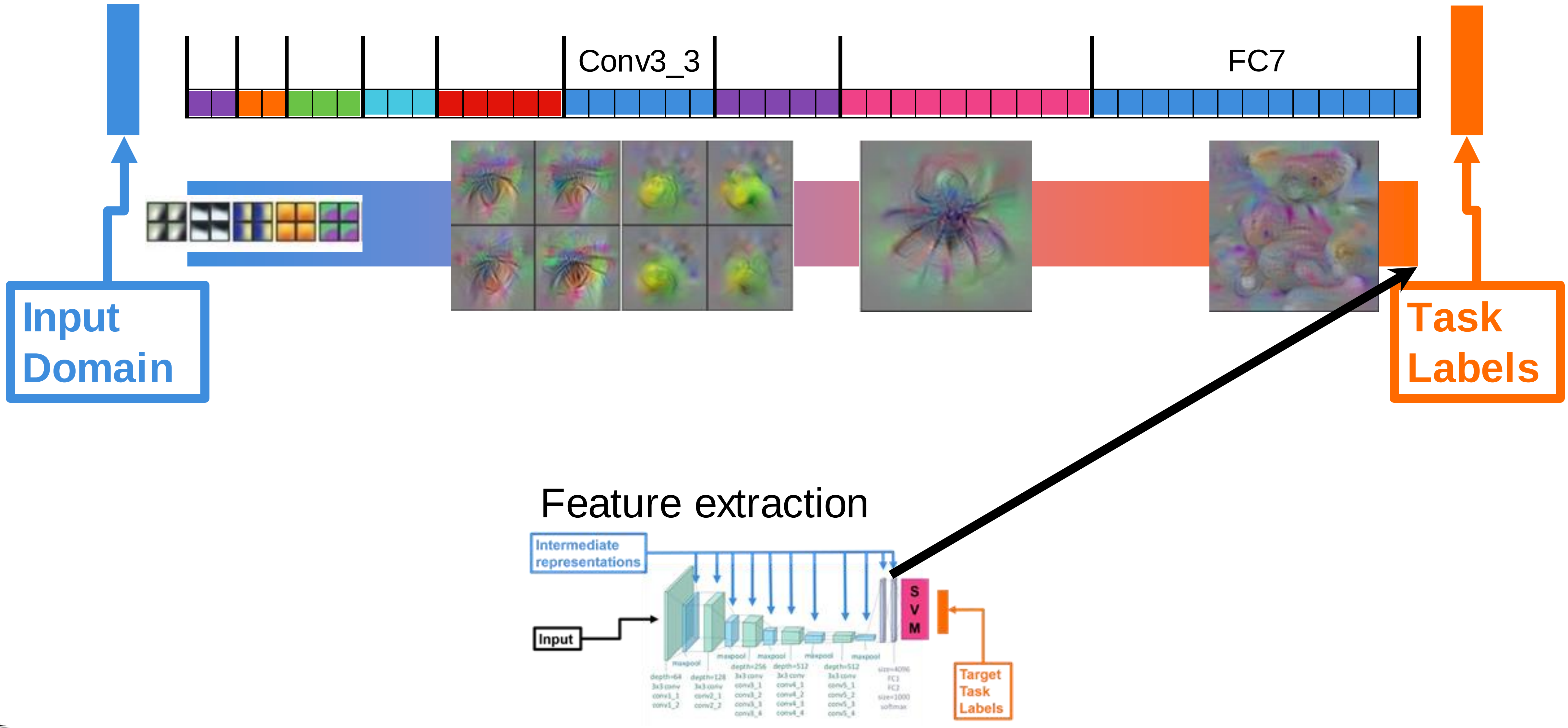
Fine tuning inside DNN?



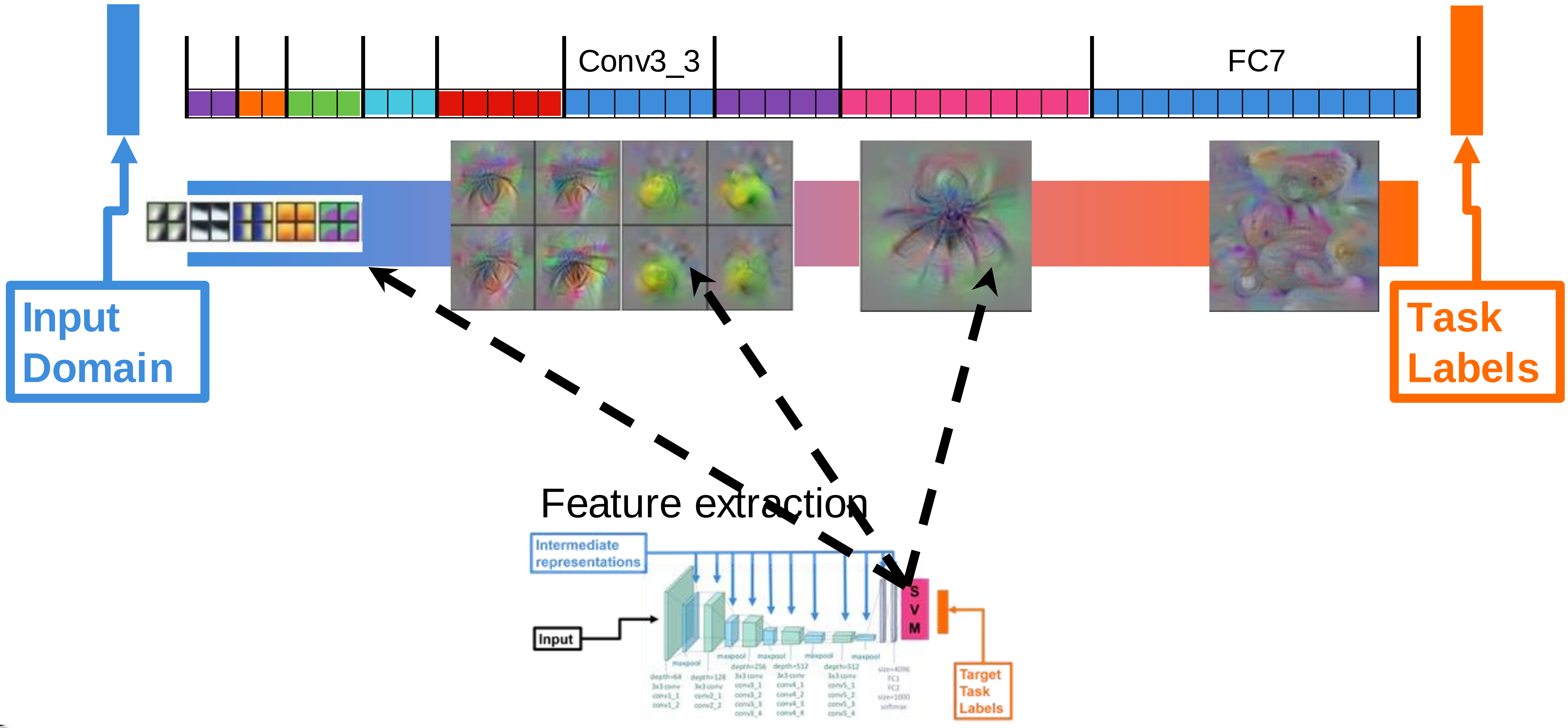
Fine tuning inside DNN?



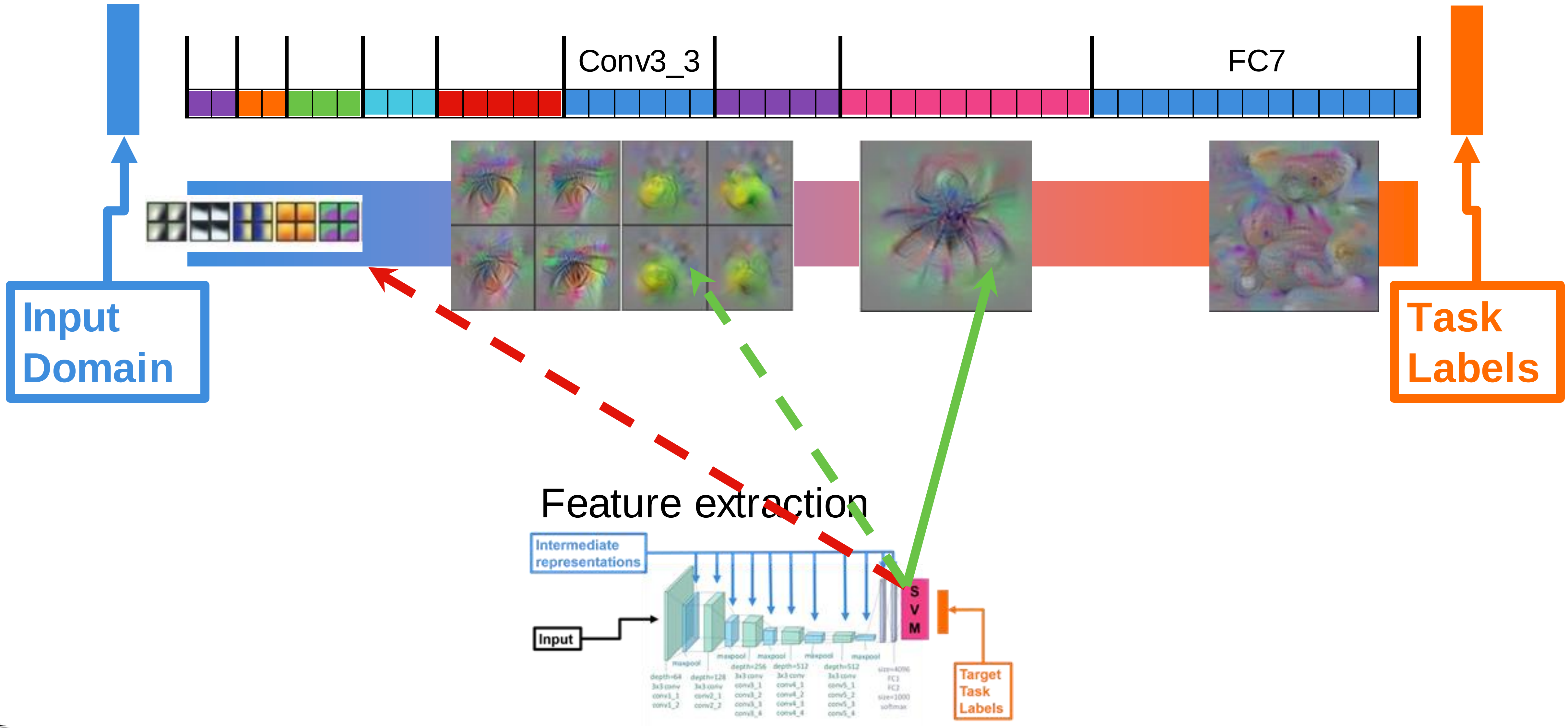
Feature extraction inside DNN?



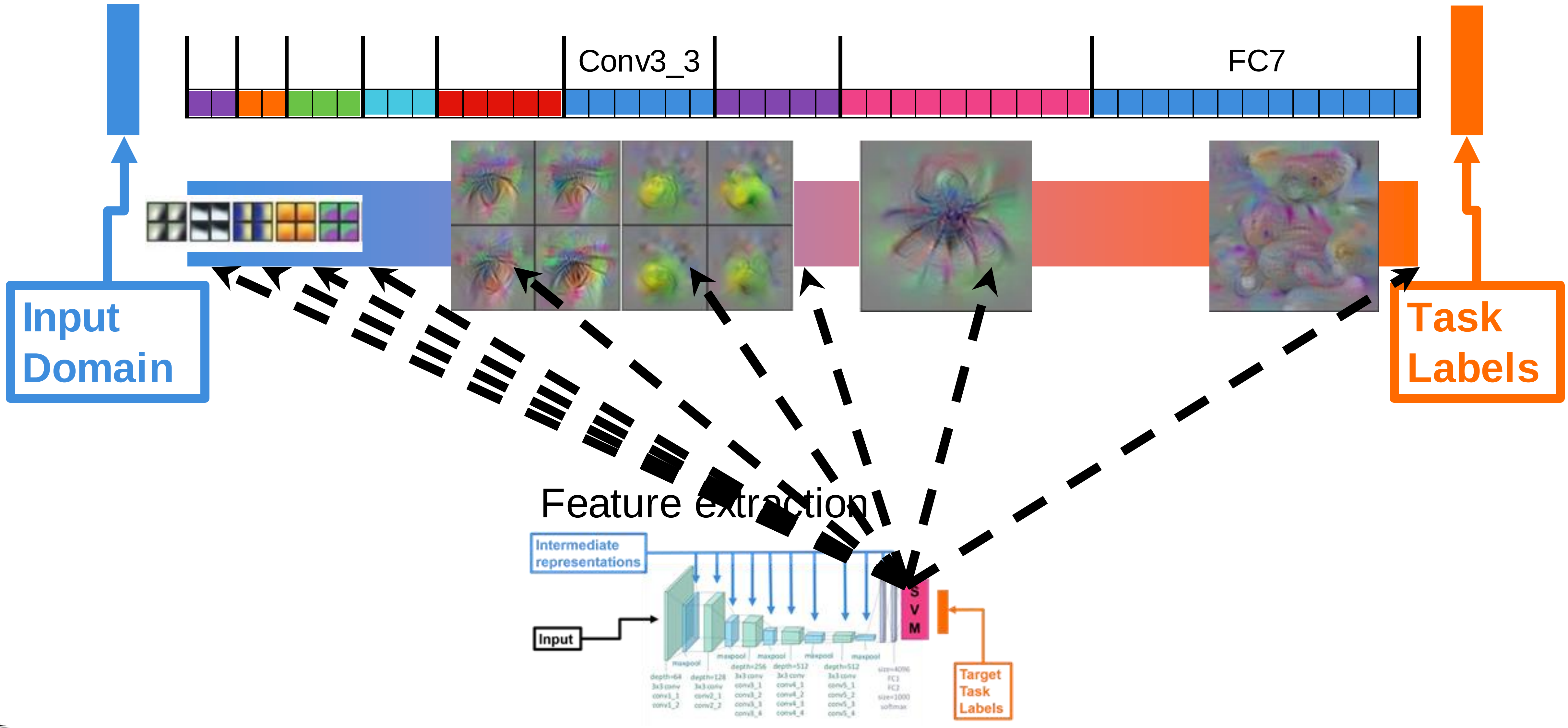
Feature extraction inside DNN?



Feature extraction inside DNN?

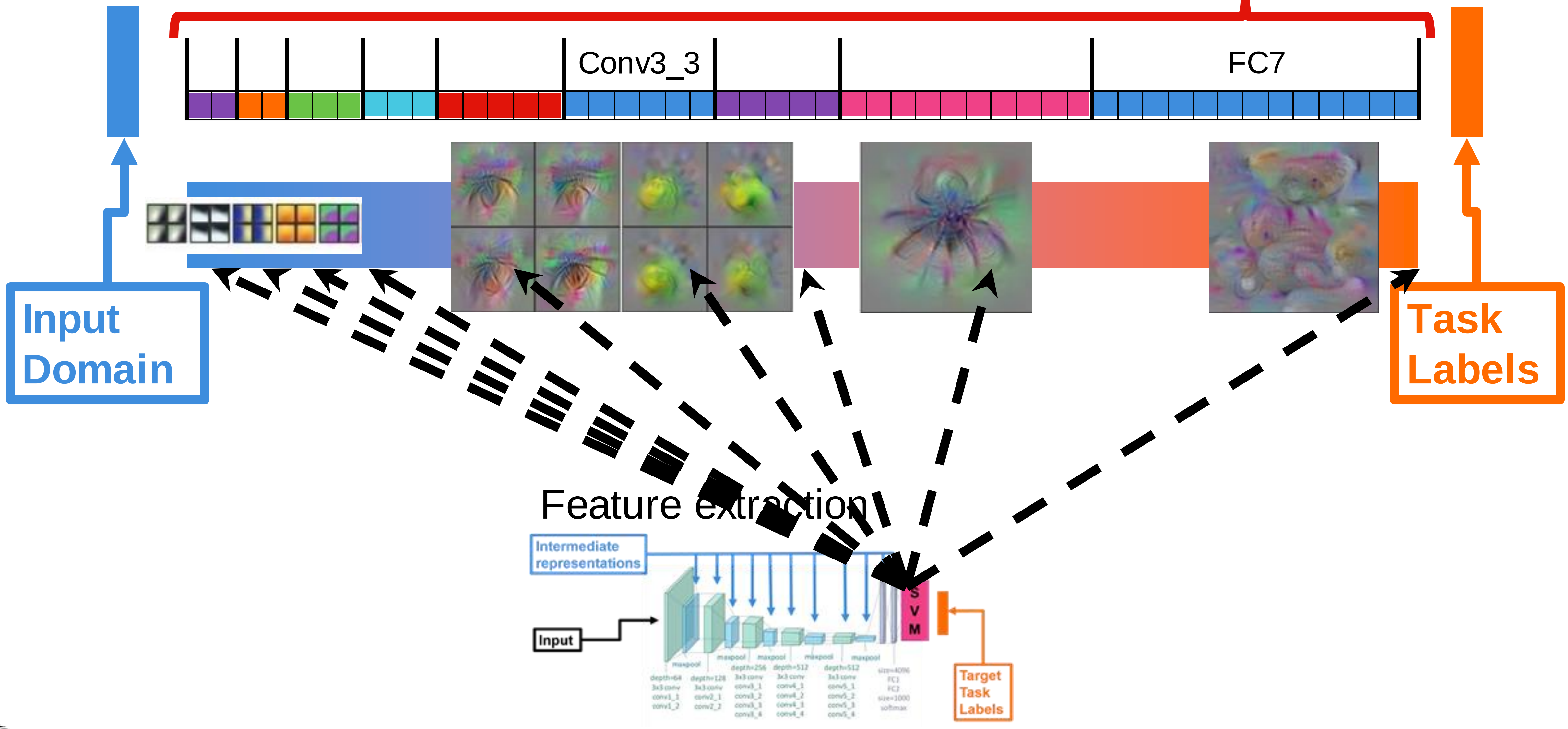


Feature extraction inside DNN?

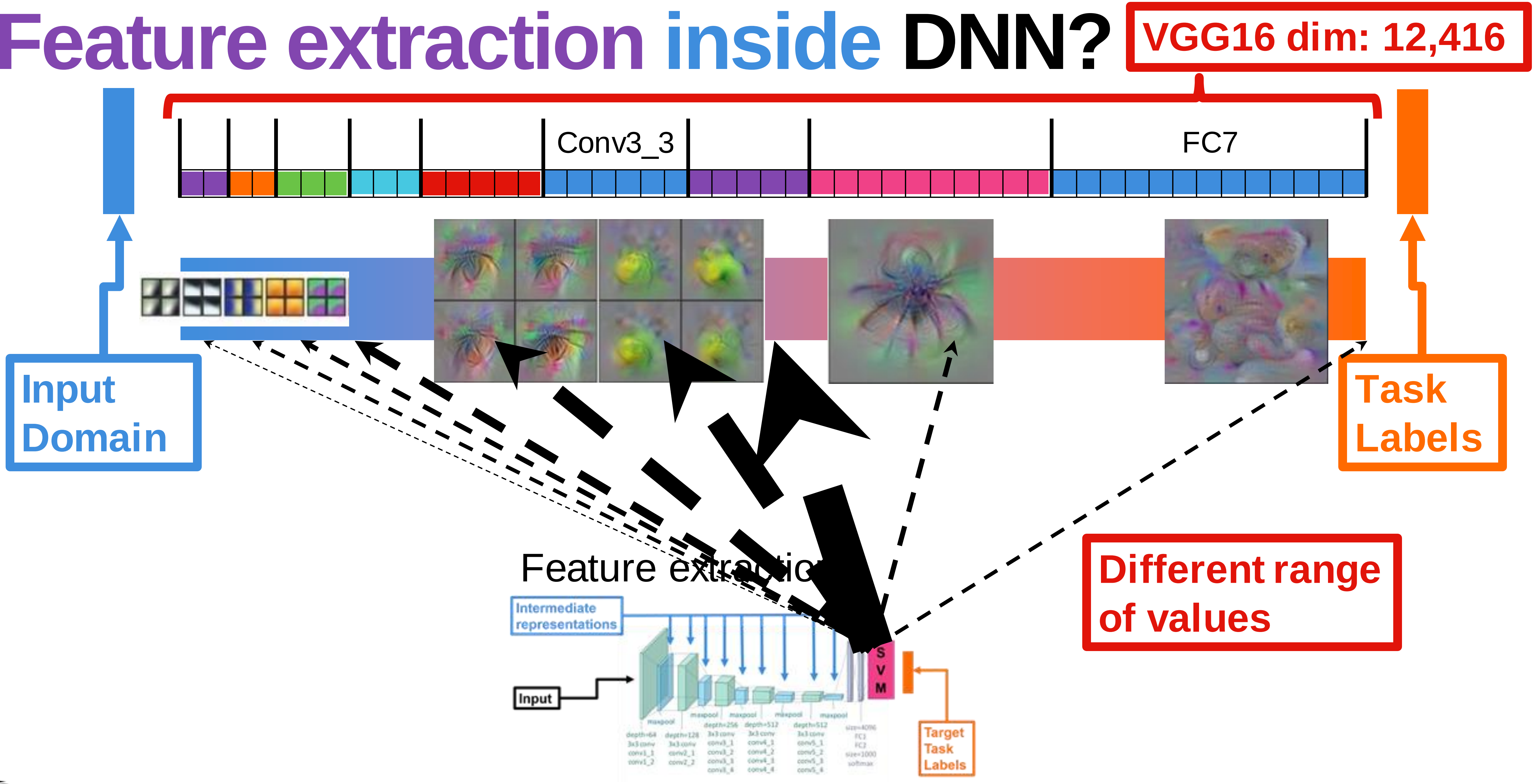


Feature extraction inside DNN?

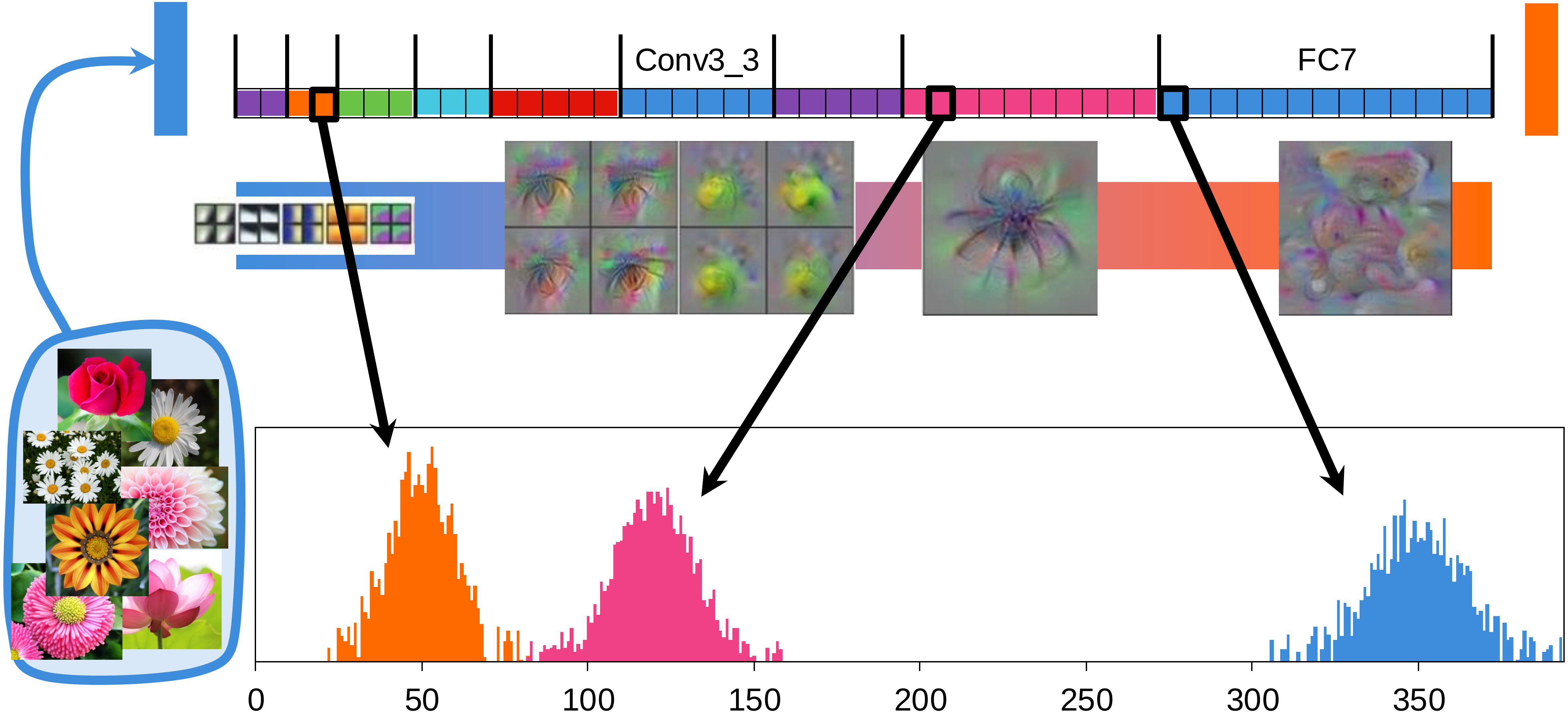
VGG16 dim: 12,416



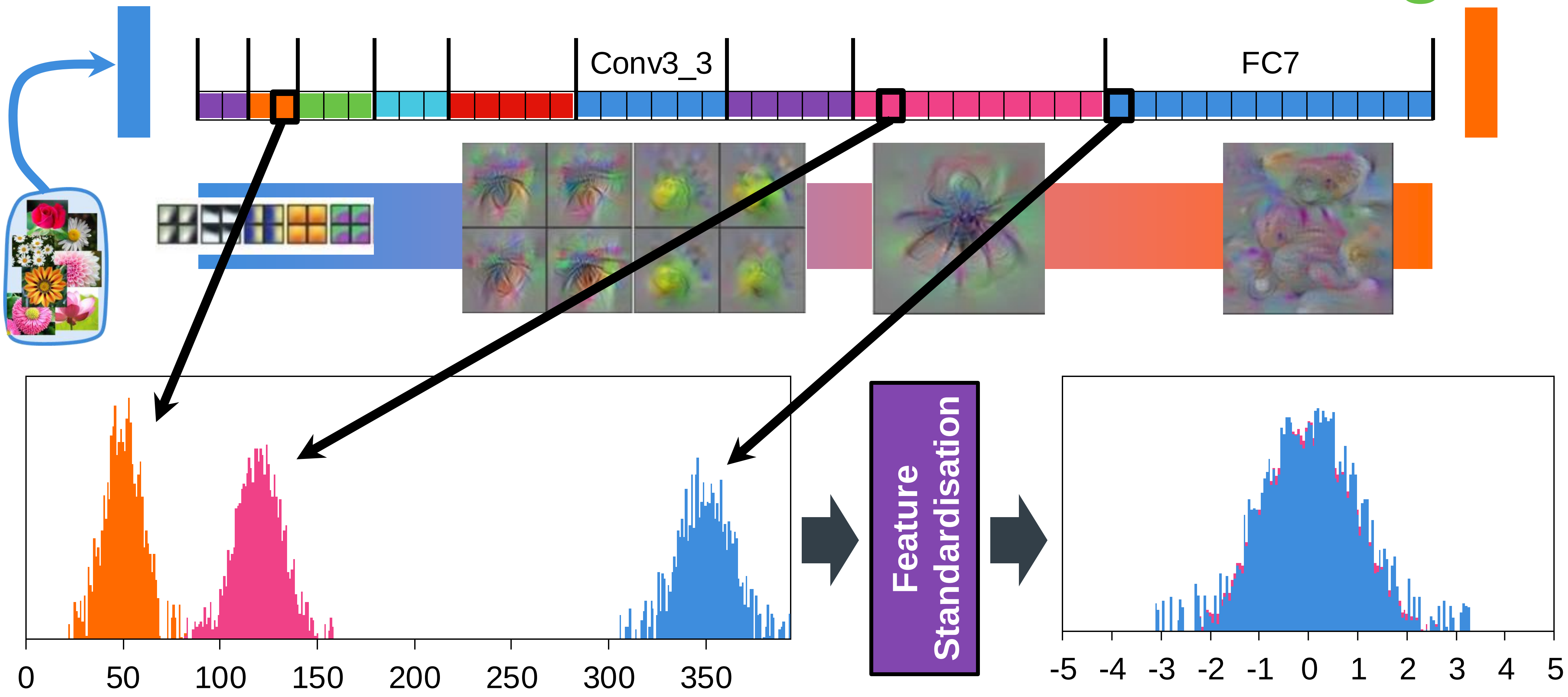
Feature extraction inside DNN?



Feature behavior in Transfer Learning



Standardization: features in same range

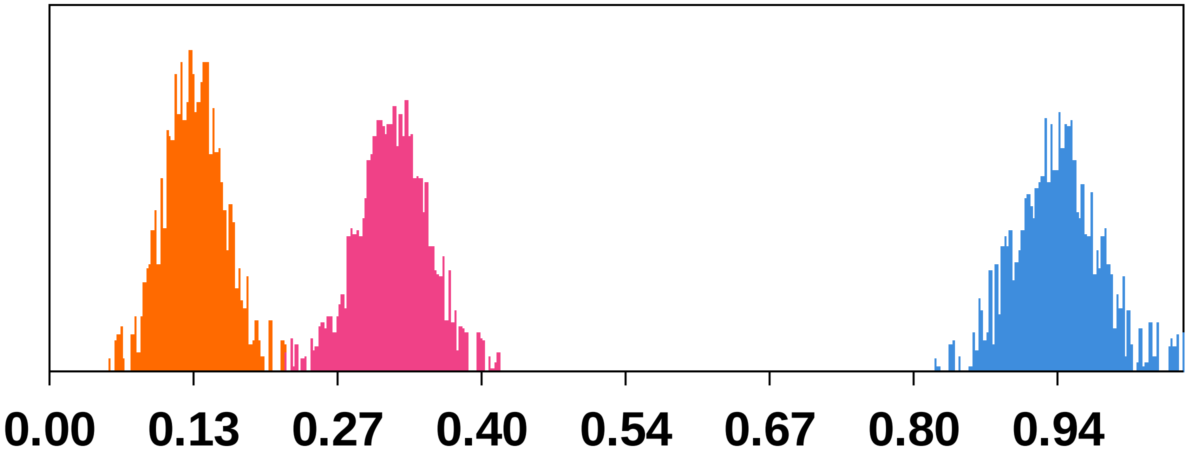


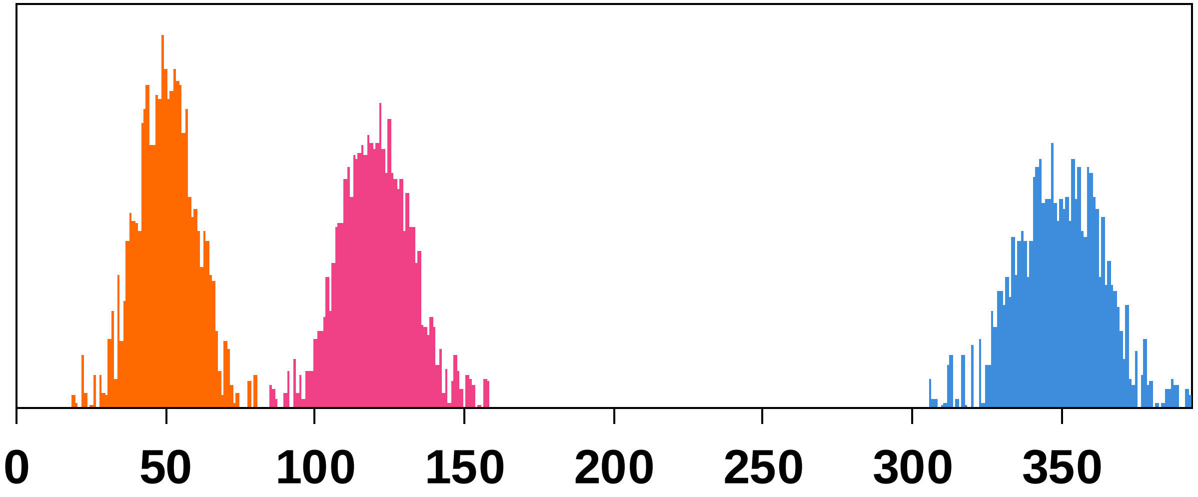
Each feature independently

Standardization: features in same range

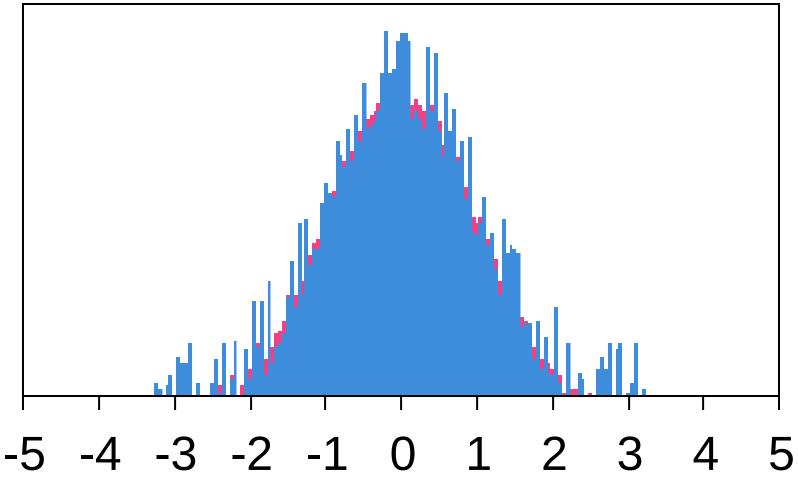
Typically, it is used a L2-normalization

L2 norm





Feature Standardisation

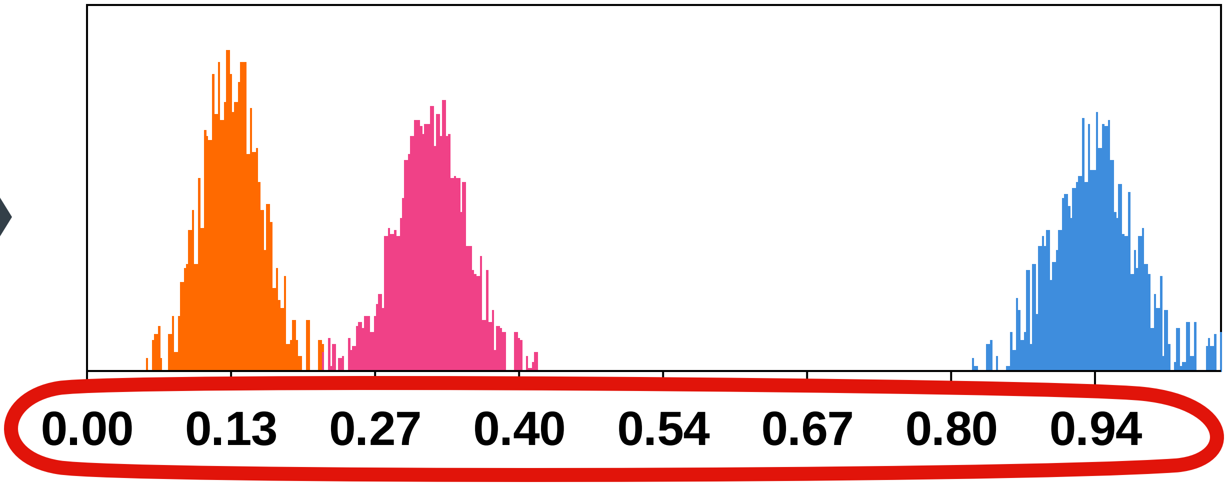


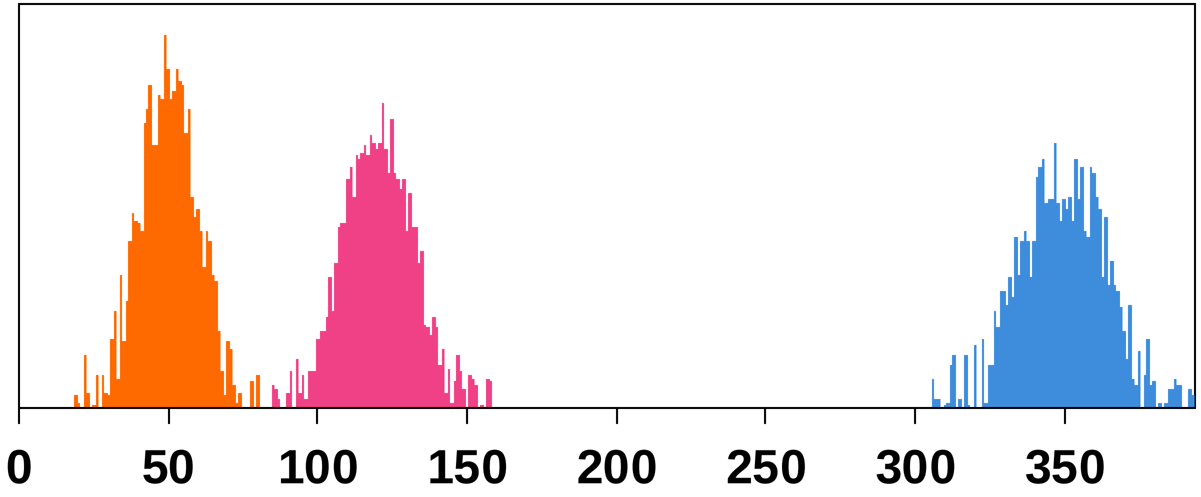
Each feature independently

Standardization: features in same range

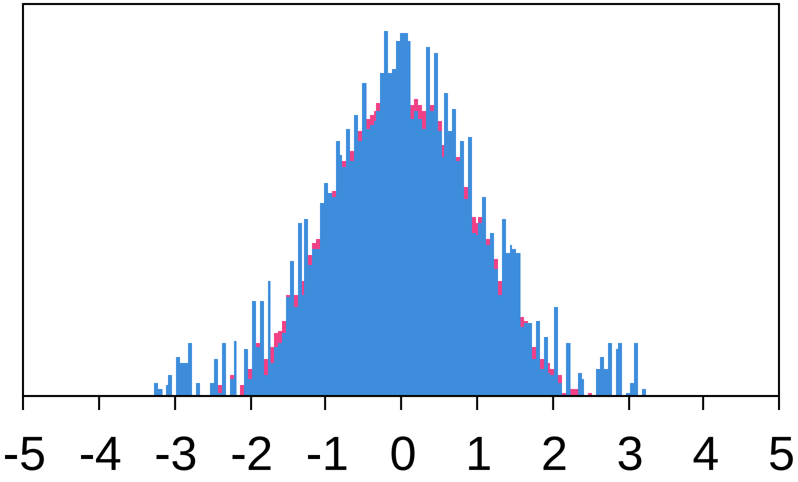
Typically, it is used a L2-normalization

L2 norm





Feature Standardisation



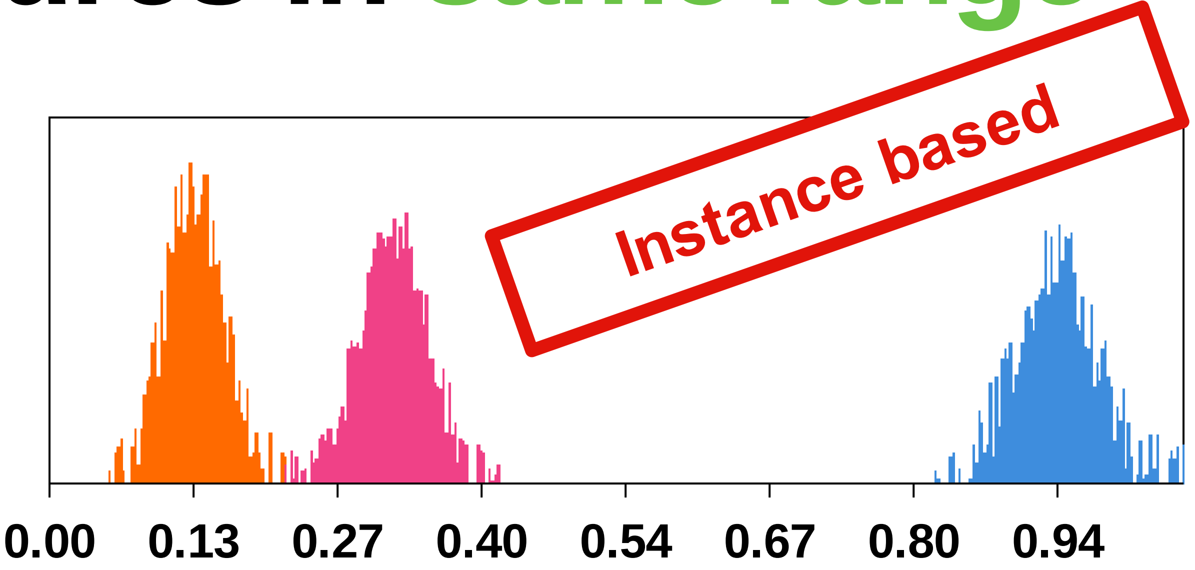
Each feature independently

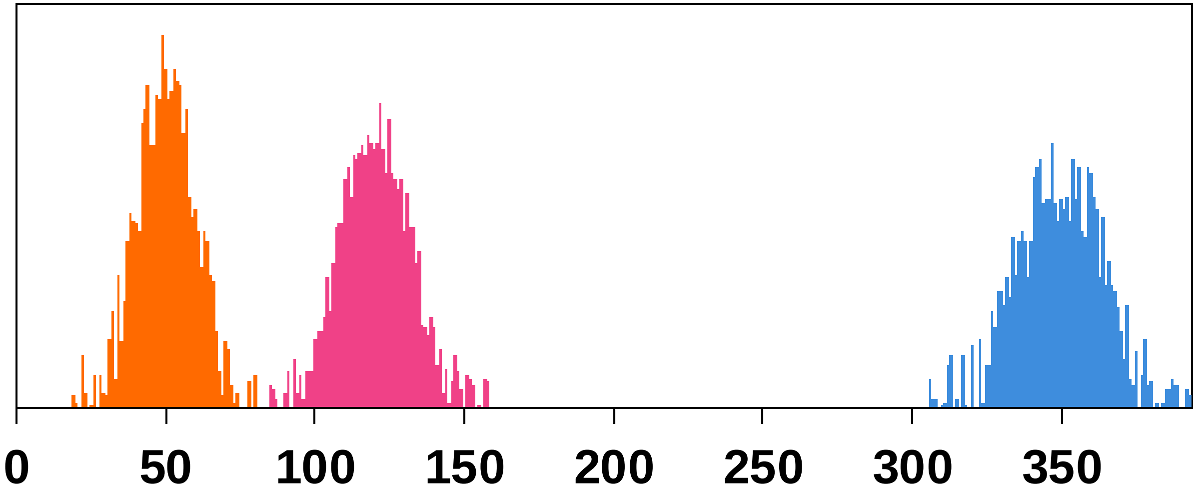
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Standardization: features in same range

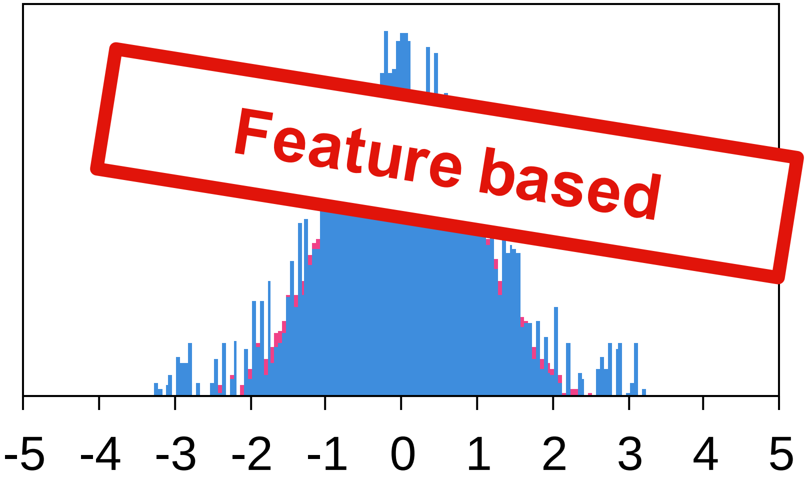
Typically, it is used a L2-normalization

L2 norm



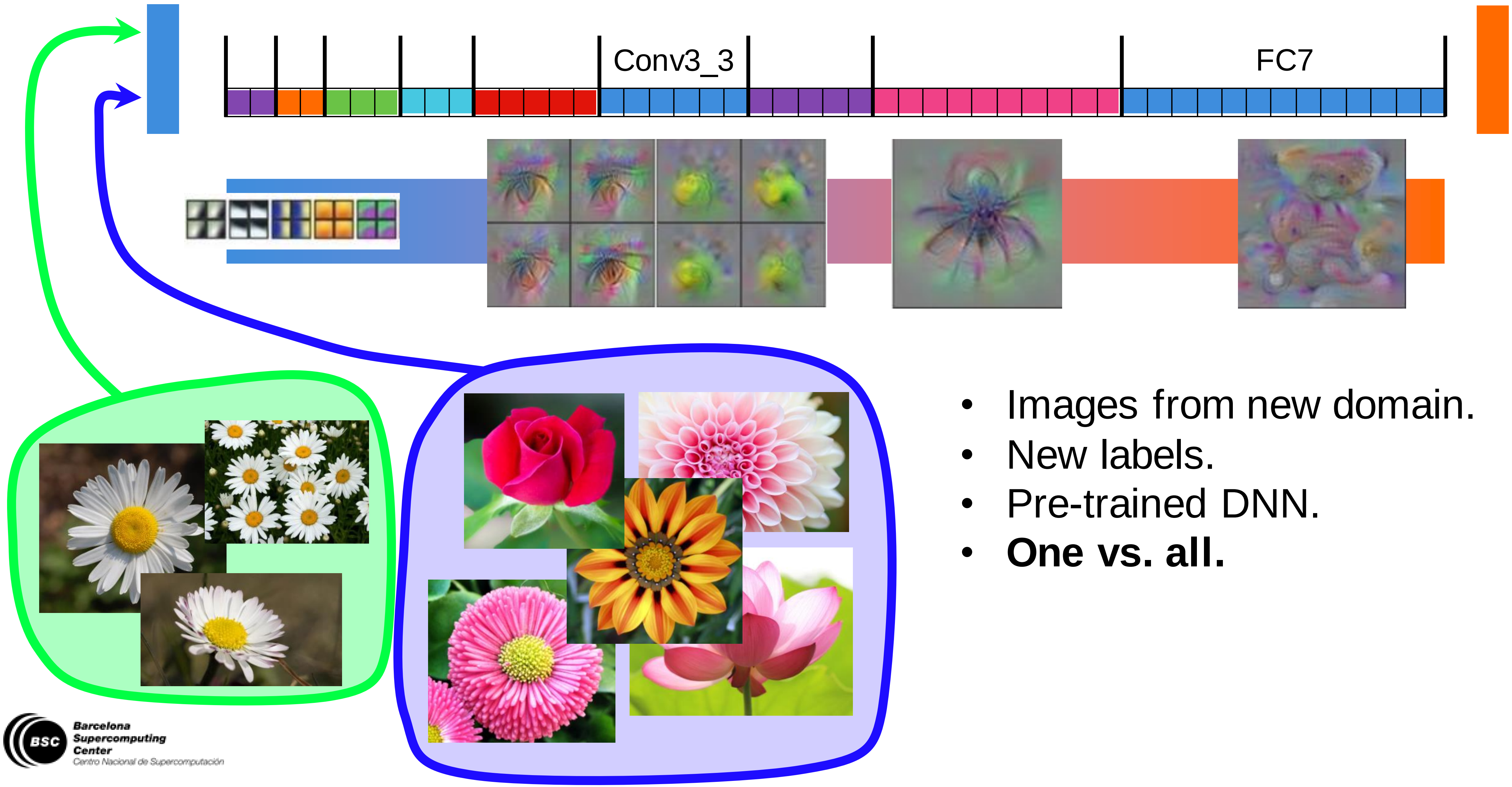


Feature Standardisation



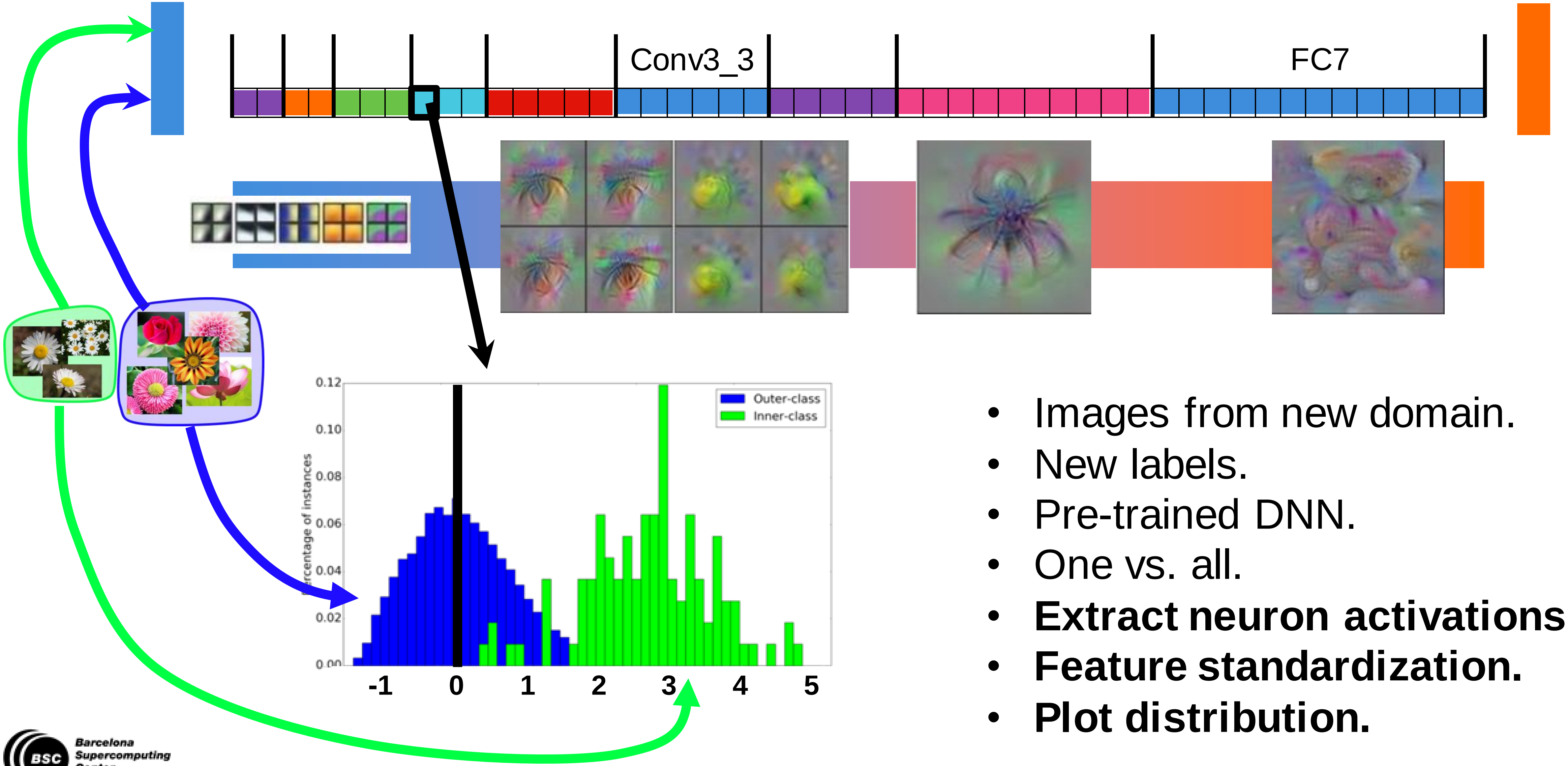
Each feature independently

Standardization: features in context



- Images from new domain.
- New labels.
- Pre-trained DNN.
- **One vs. all.**

Standardization: features in context

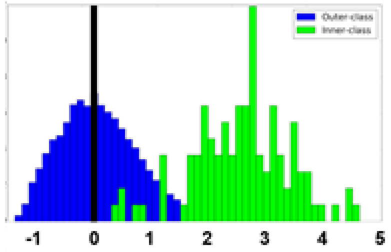


- Images from new domain.
- New labels.
- Pre-trained DNN.
- One vs. all.
- **Extract neuron activations.**
- **Feature standardization.**
- **Plot distribution.**

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Stardardization: features in context

High activation features for a specific class



conv3_4 n202

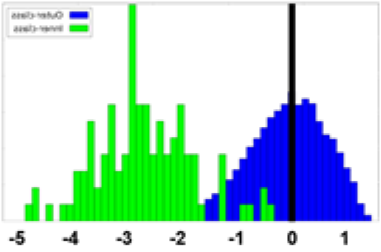
Greenhouse

fc7 n1779

Cloister

MIT-67 - **indoors**

Low activation features for a specific class



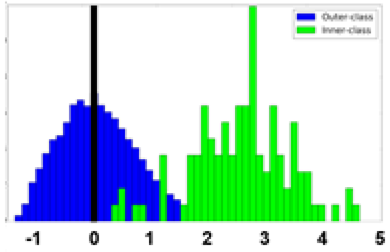
fc7 n1946

Gadwall Brown Pelican White Pelican Heermann Gull

CUB-200 - **birds**

Stardardization: features in context

High activation features for a specific class



conv3_4 n202

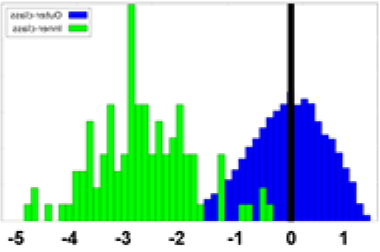
Greenhouse

fc7 n1779

Cloister

MIT-67 - **indoors**

Low activation features for a specific class



fc7 n1449

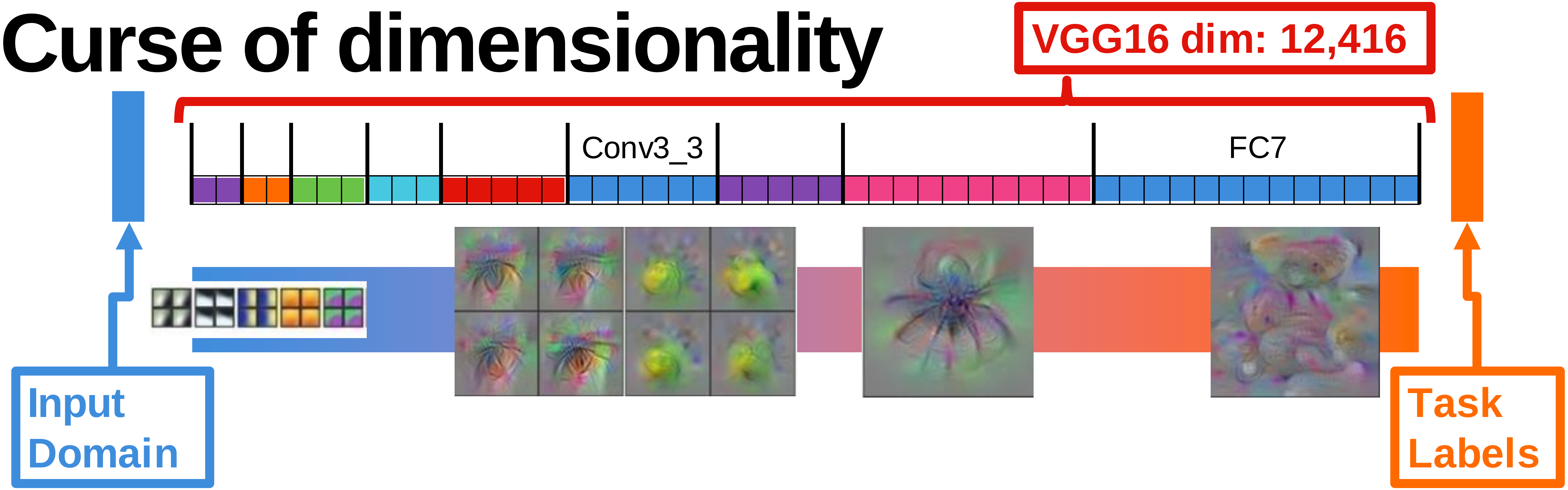
35 alpine sea holly

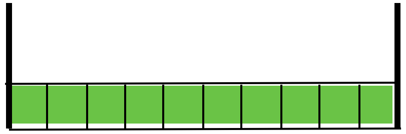
10 globe thistle

14 spear thistle

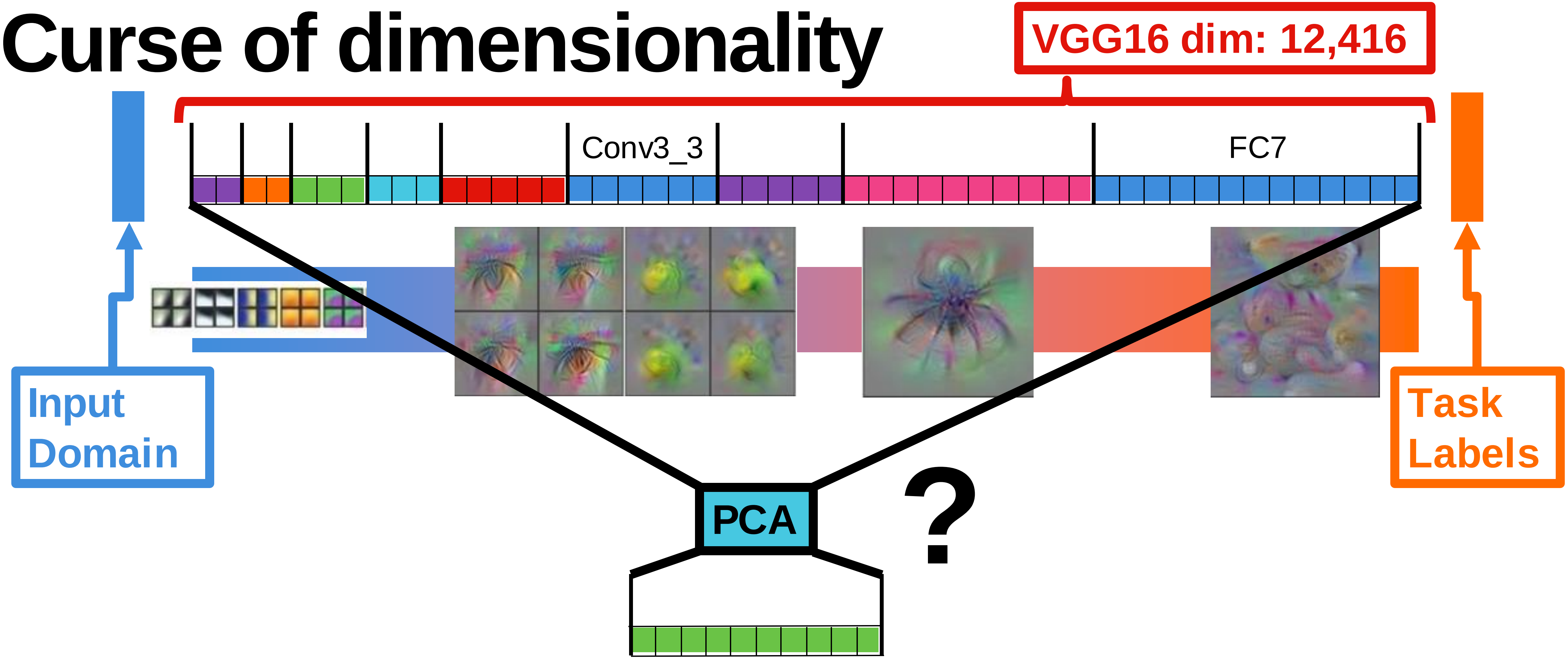
Flowers-102

Curse of dimensionality

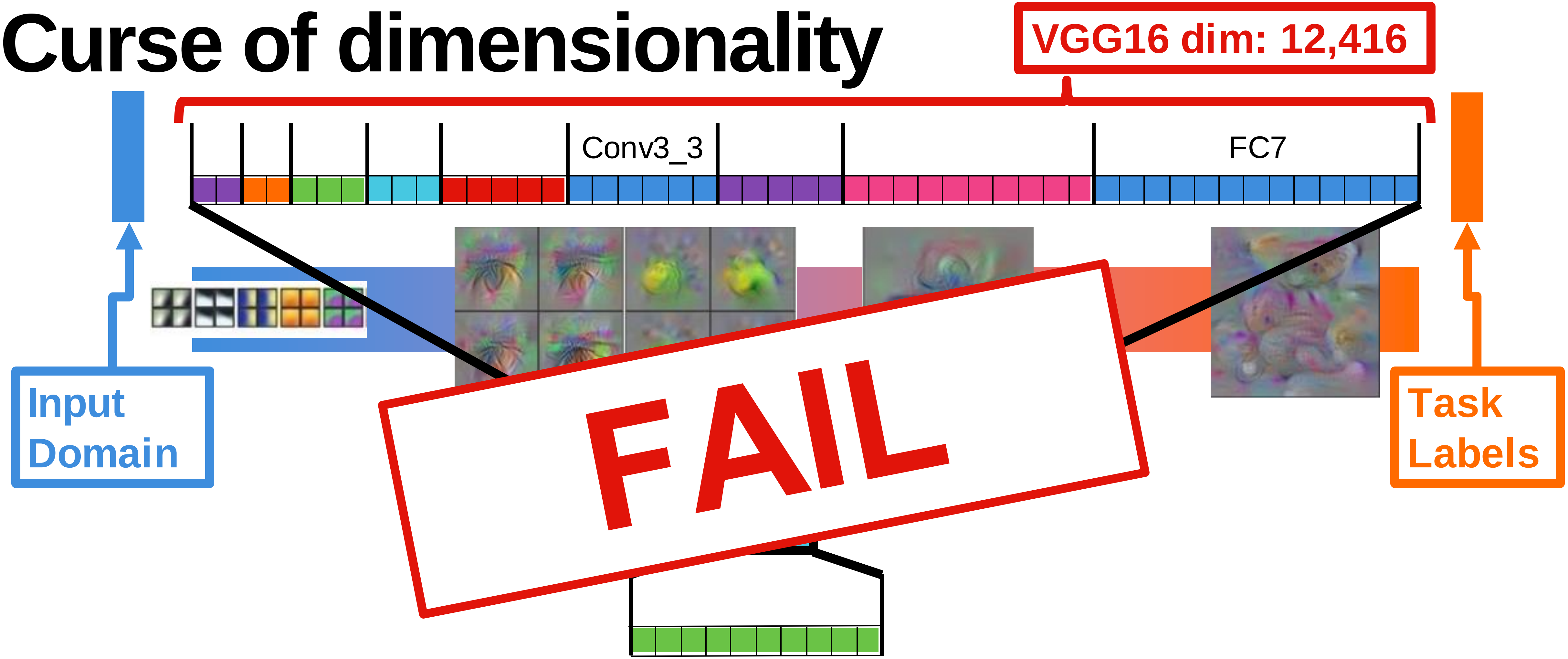




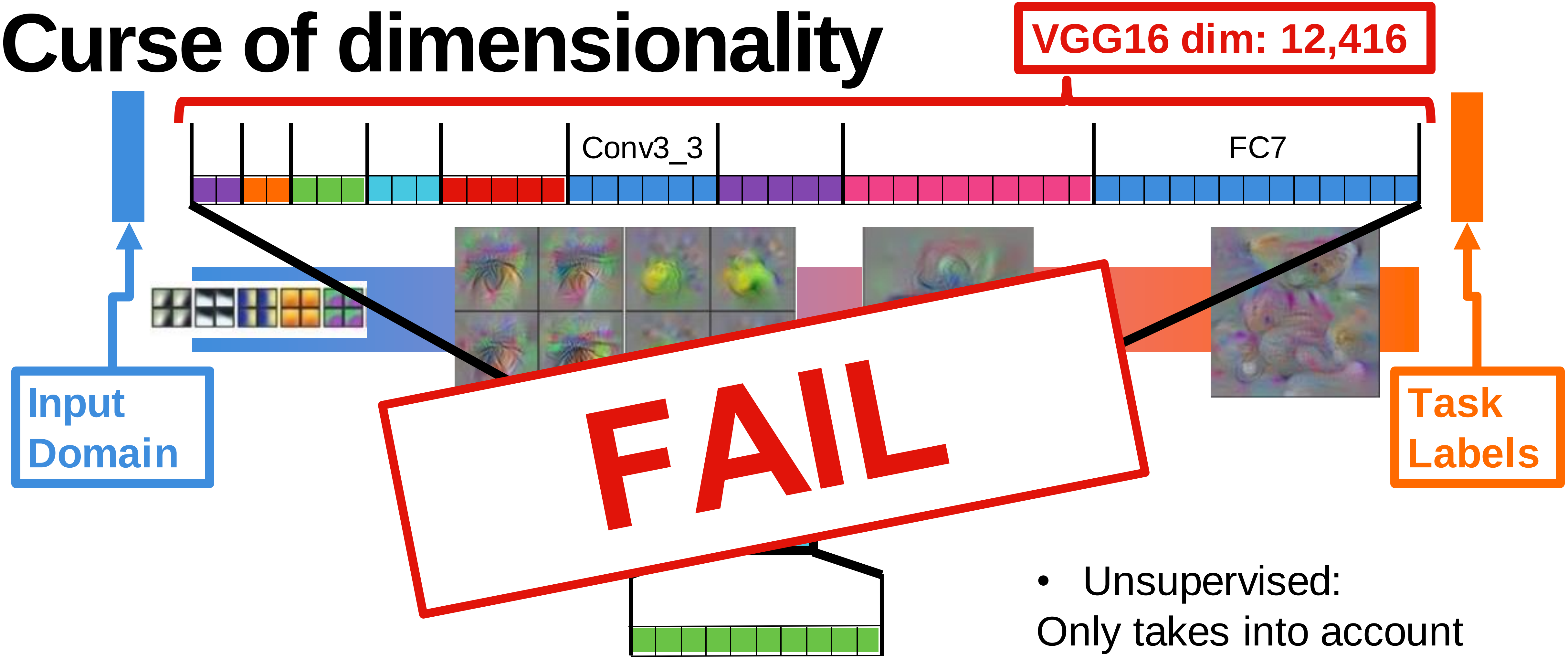
Curse of dimensionality



Curse of dimensionality

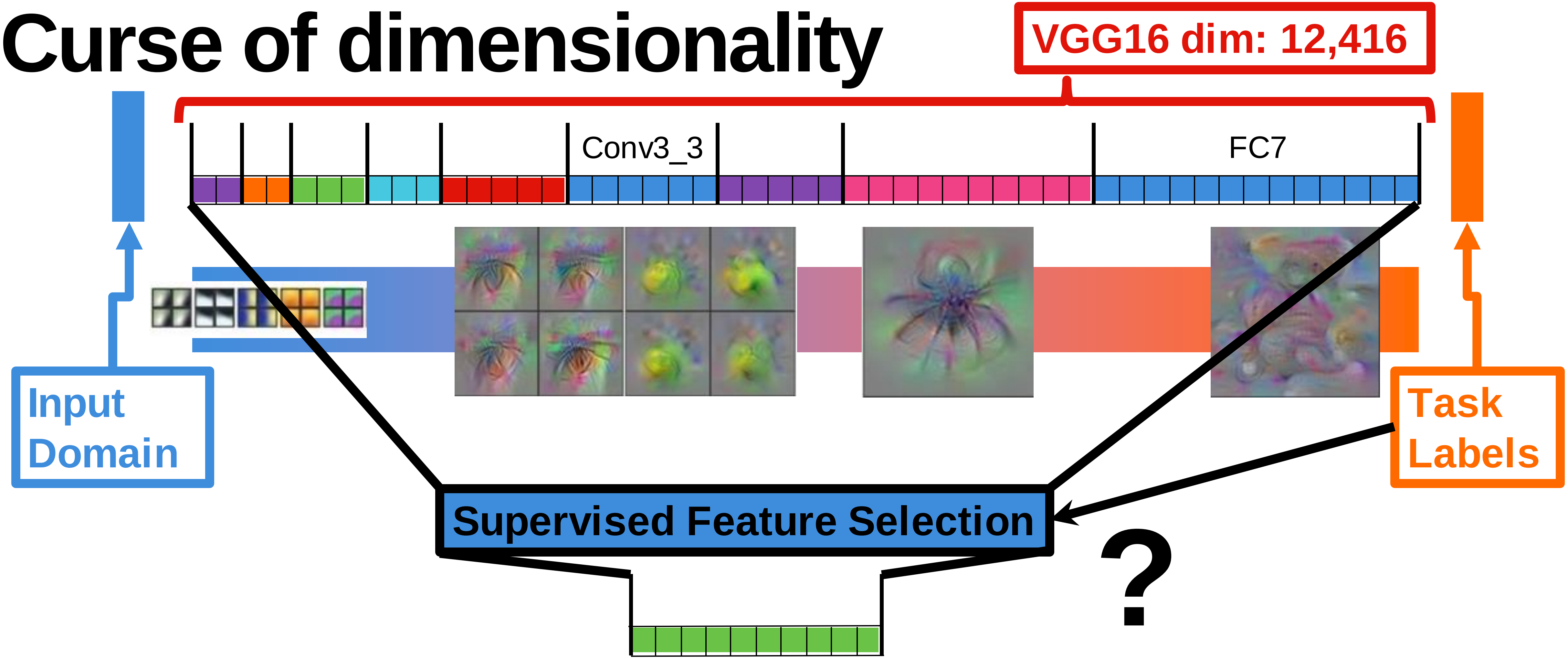


Curse of dimensionality

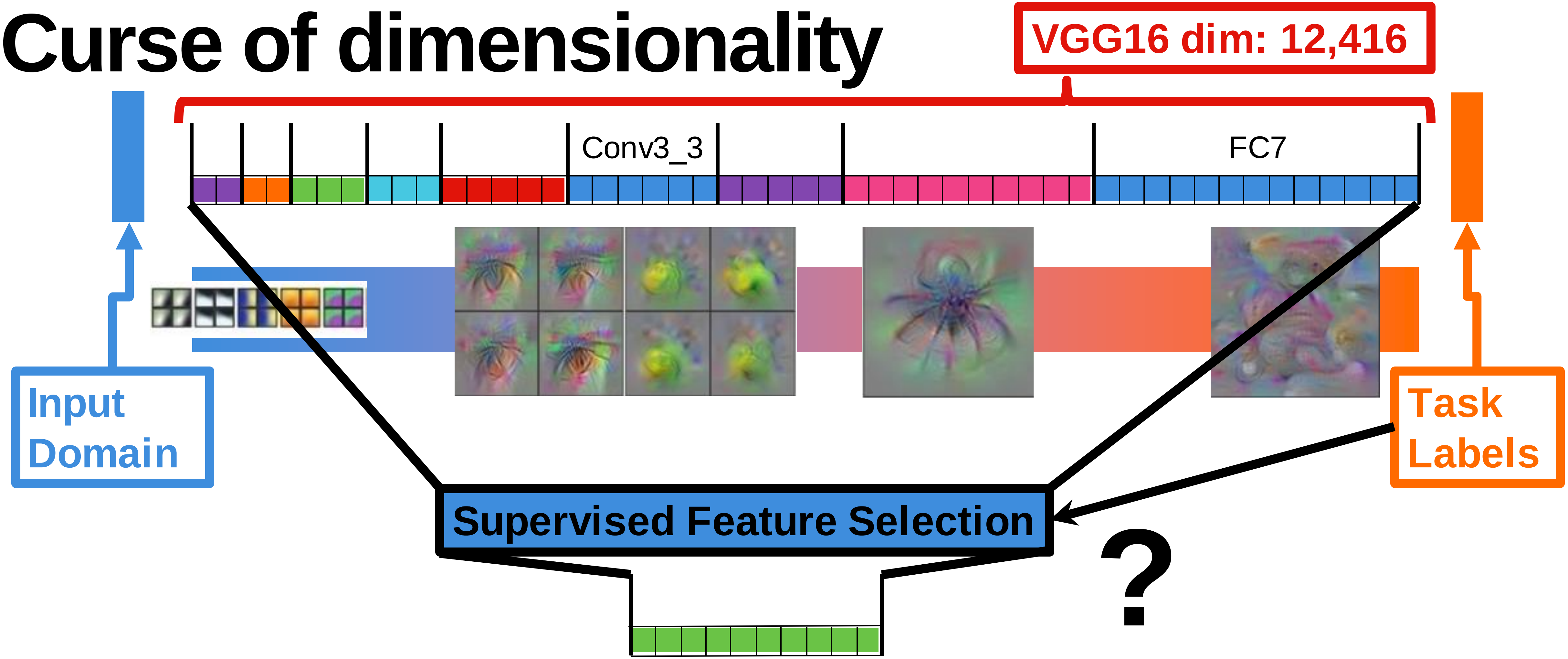


- Unsupervised:
Only takes into account the **domain**

Curse of dimensionality

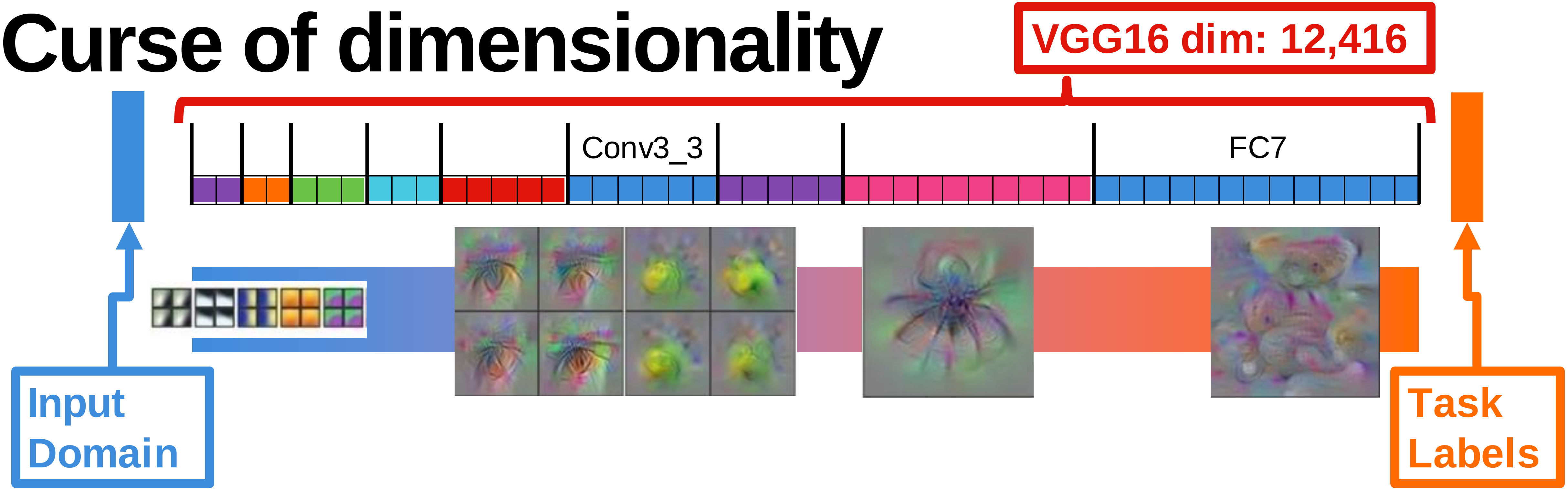


Curse of dimensionality

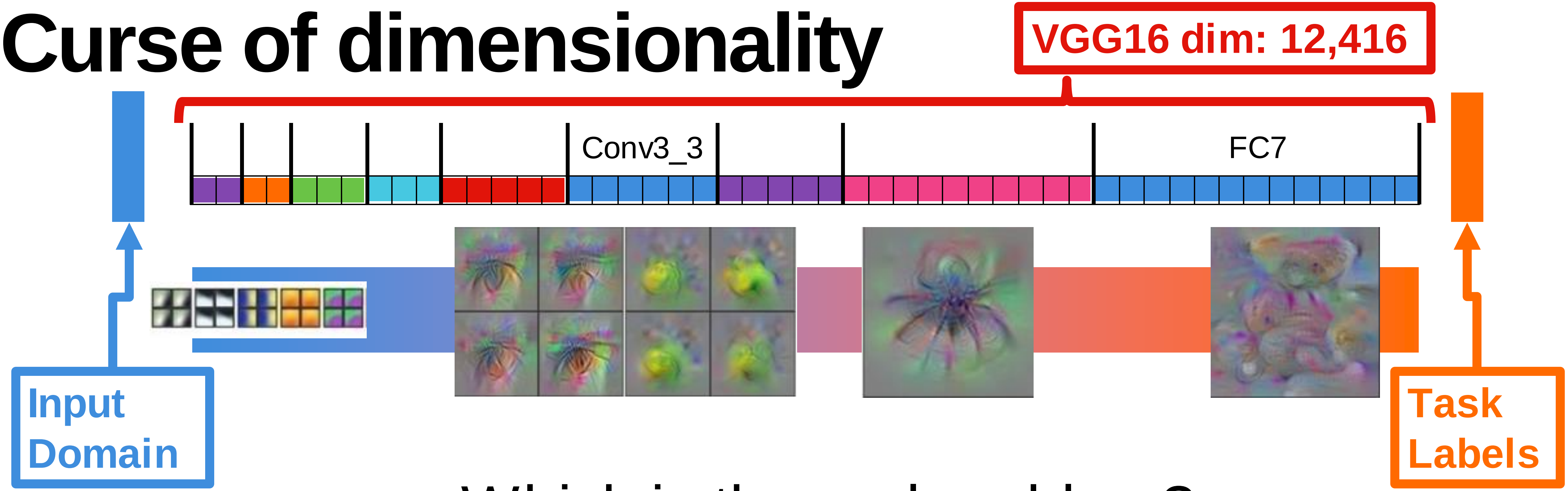


- High computational **cost!**

Curse of dimensionality

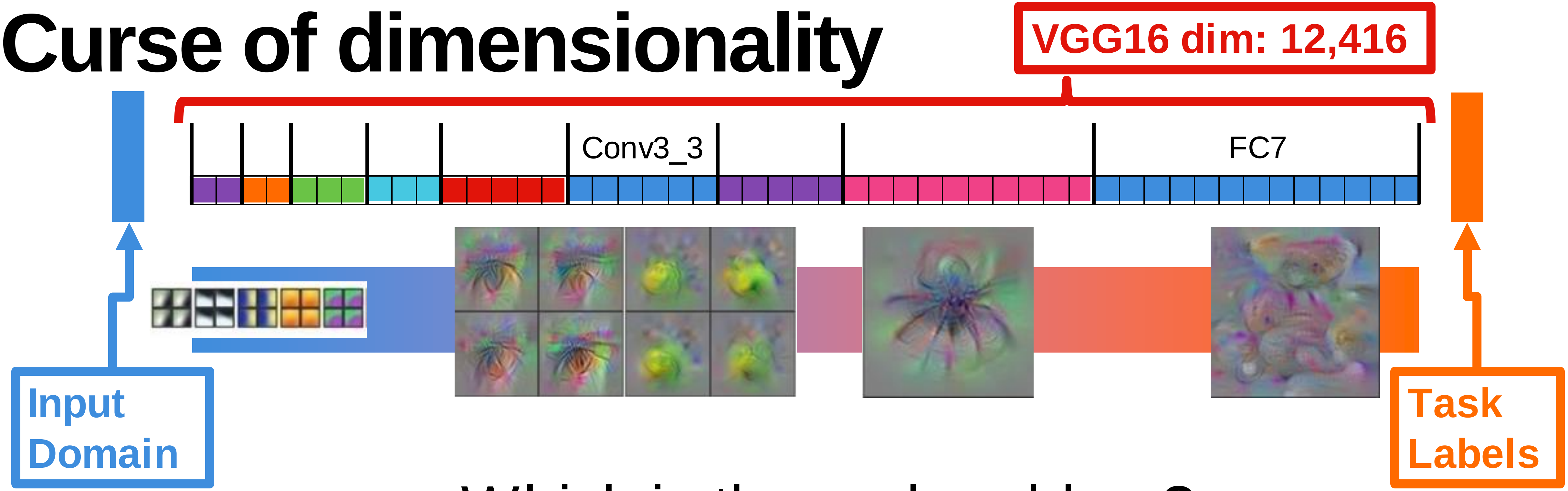


Curse of dimensionality



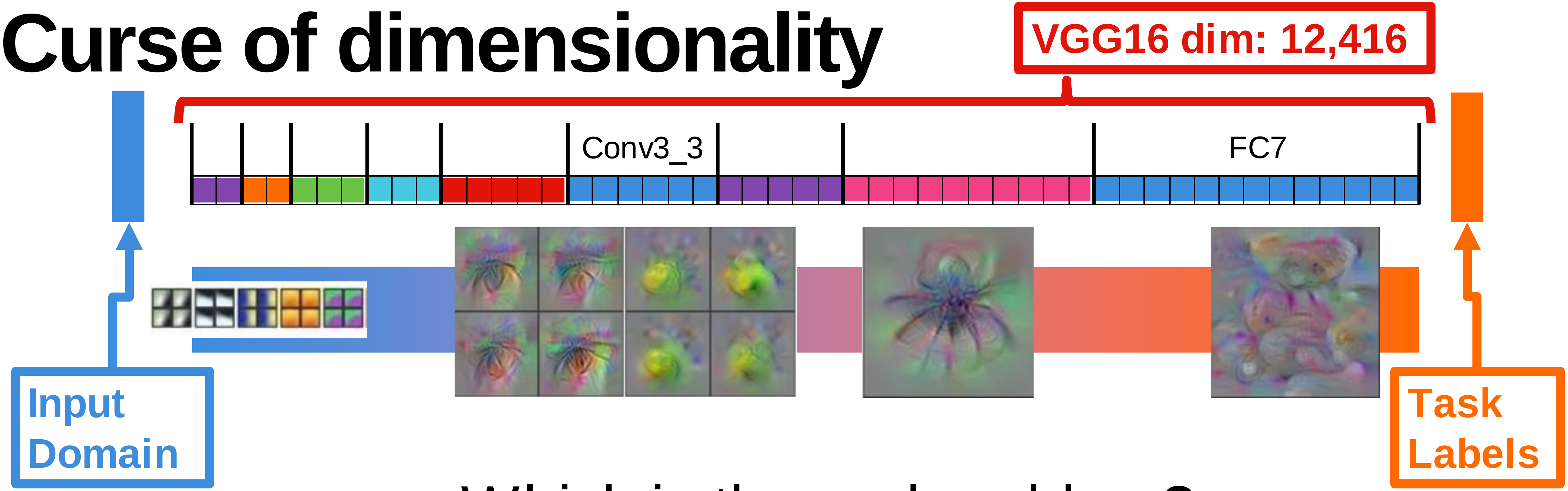
- Which is the real problem?

Curse of dimensionality



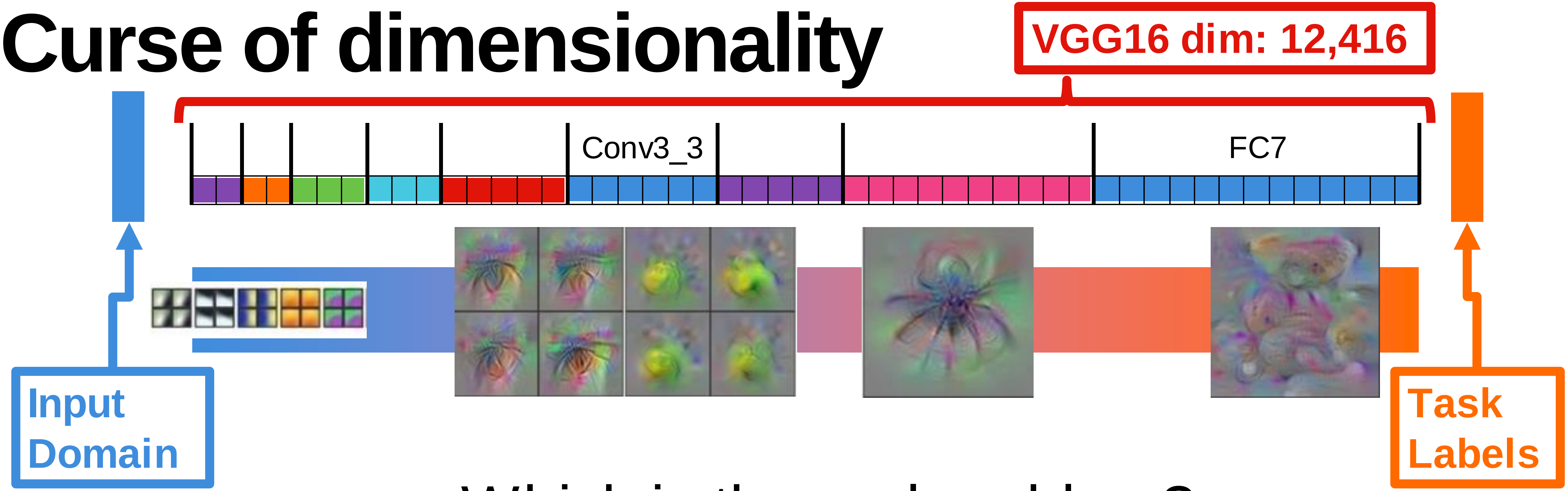
- Which is the real problem?
  - Too many features?
  - Too few images?

Curse of dimensionality



- Which is the real problem?
  - Too many features?
  - ~~Too few images?~~ **A requirement**

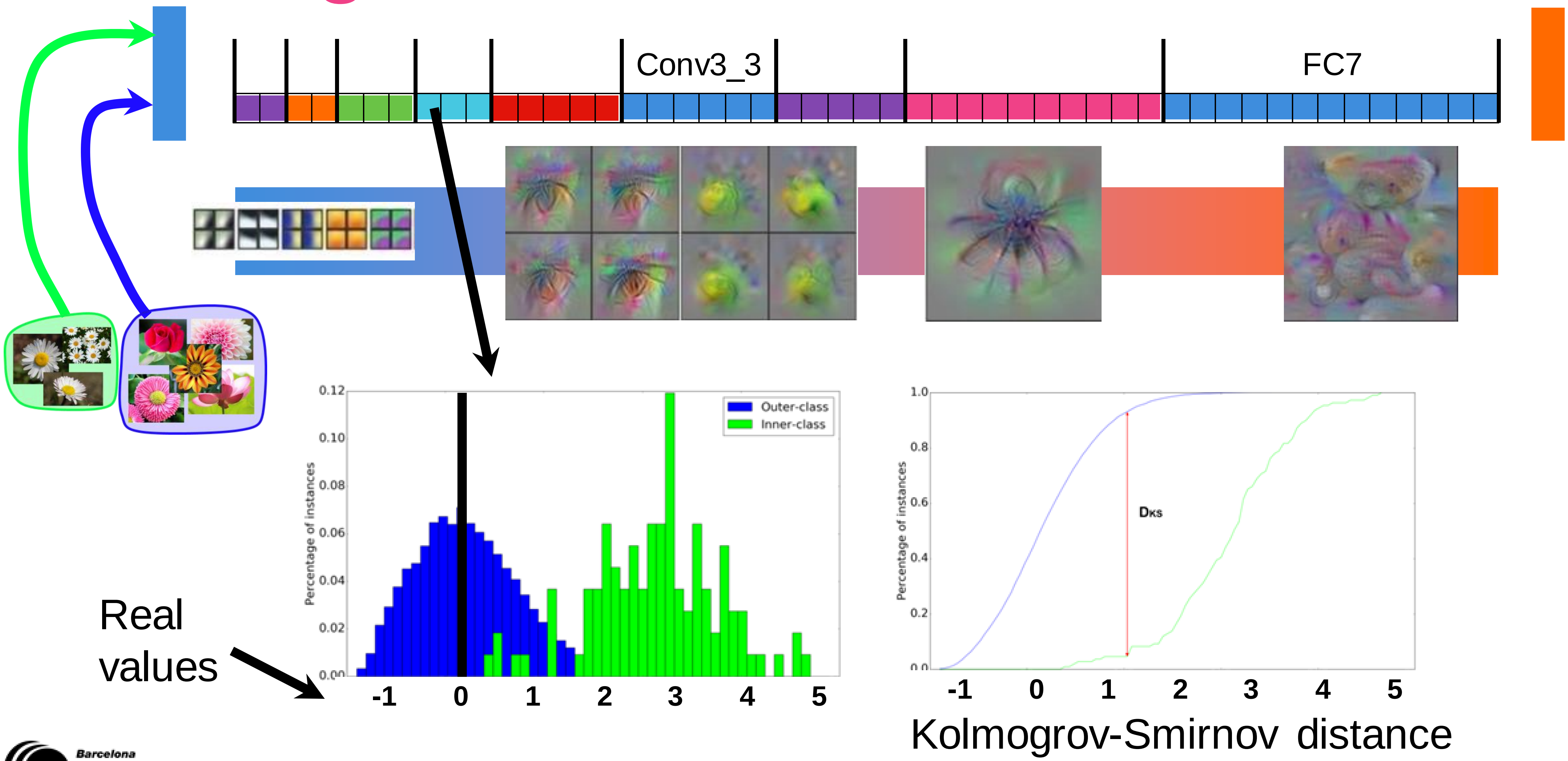
Curse of dimensionality



- Which is the real problem?
  - ~~Too many features?~~
  - **Too much information!**
  - ~~Too few images?~~

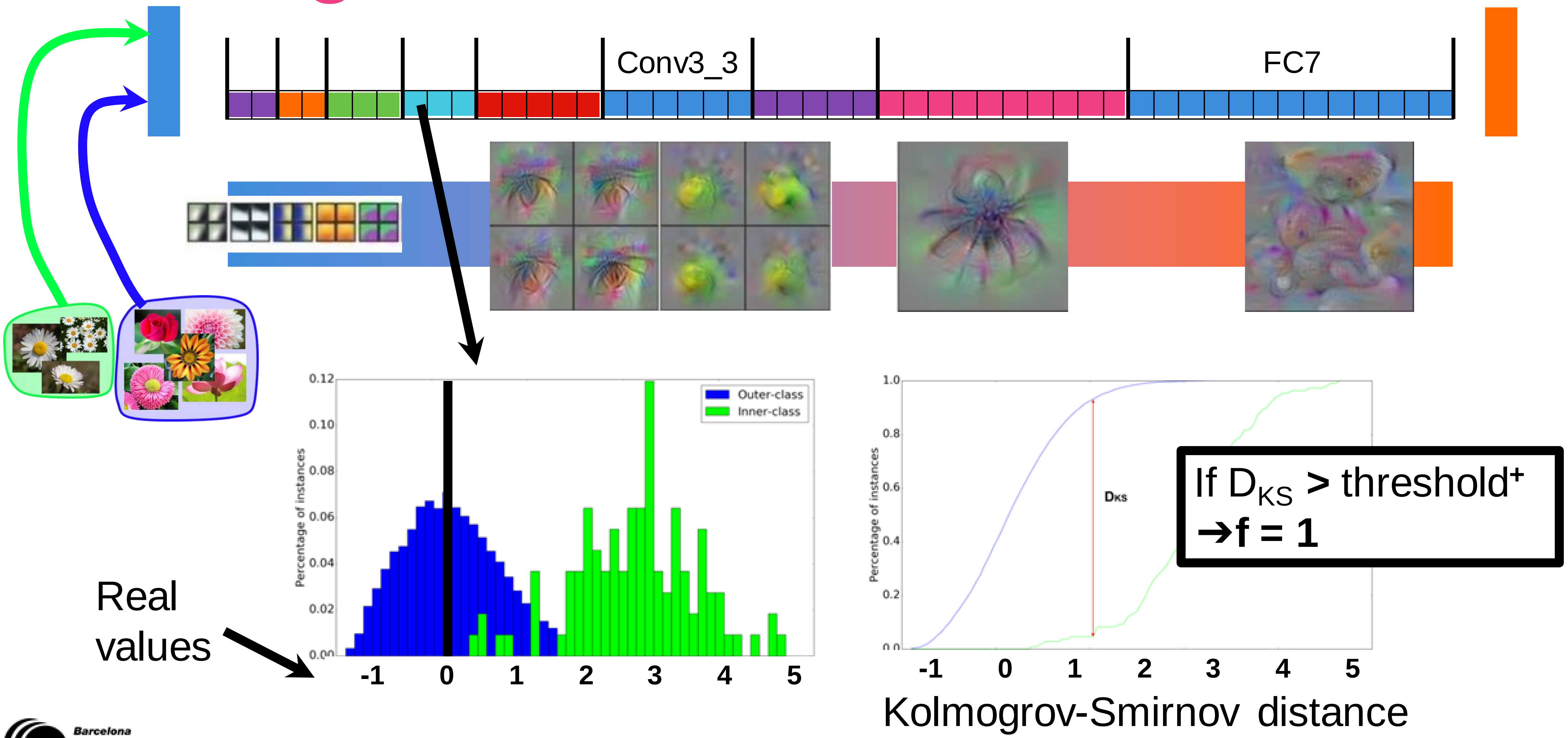
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Meaningful discretization

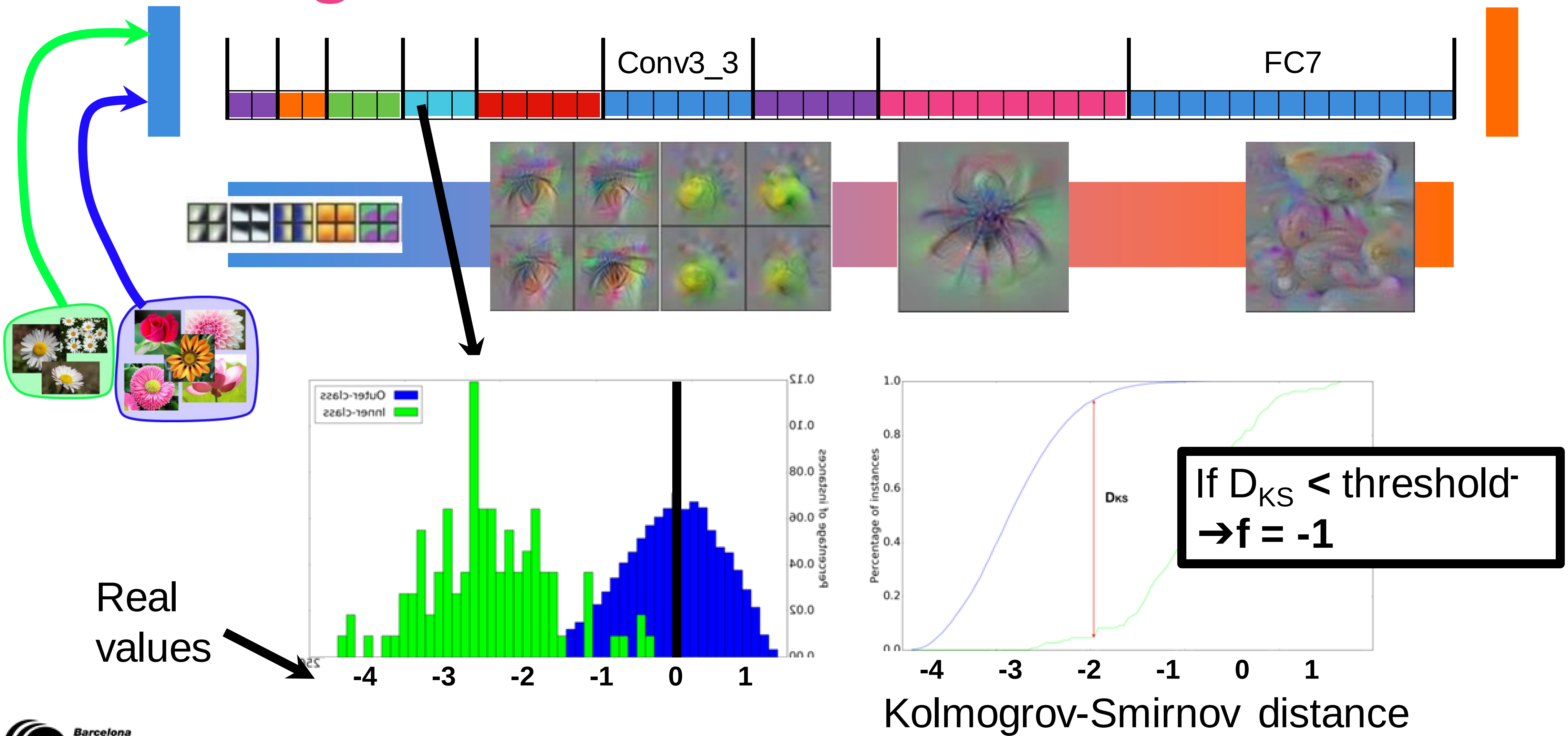


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

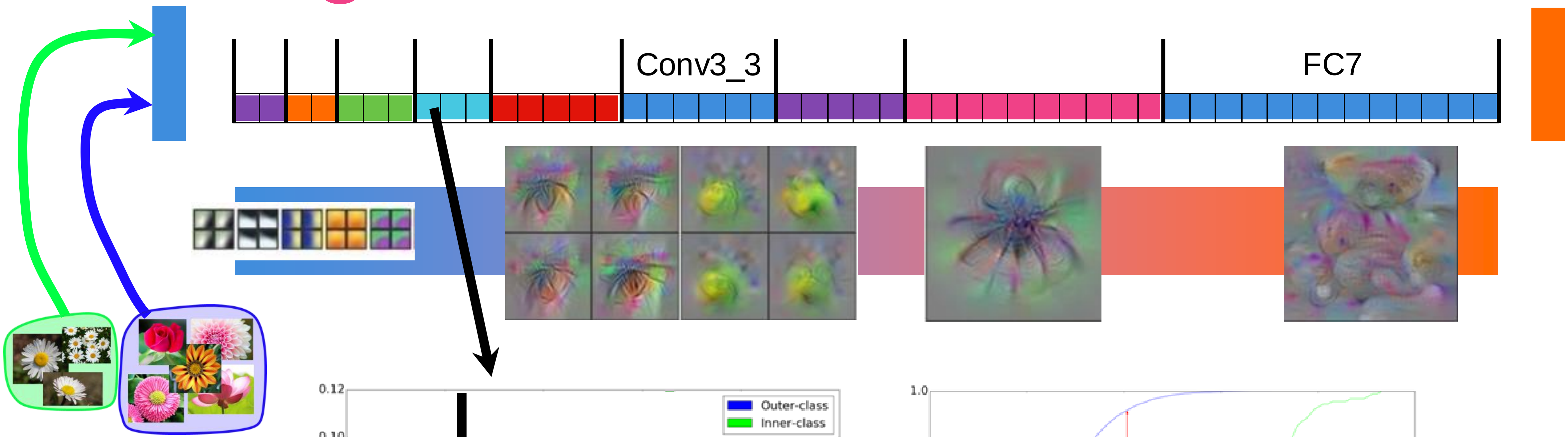
Meaningful discretization

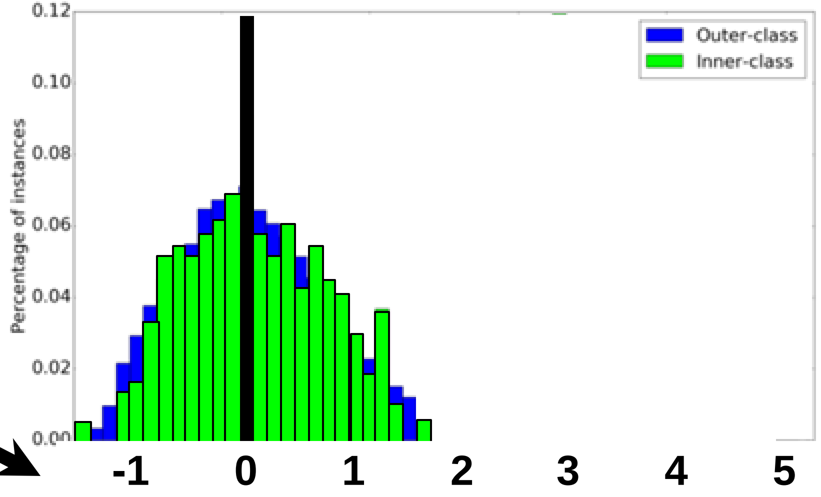


Meaningful discretization

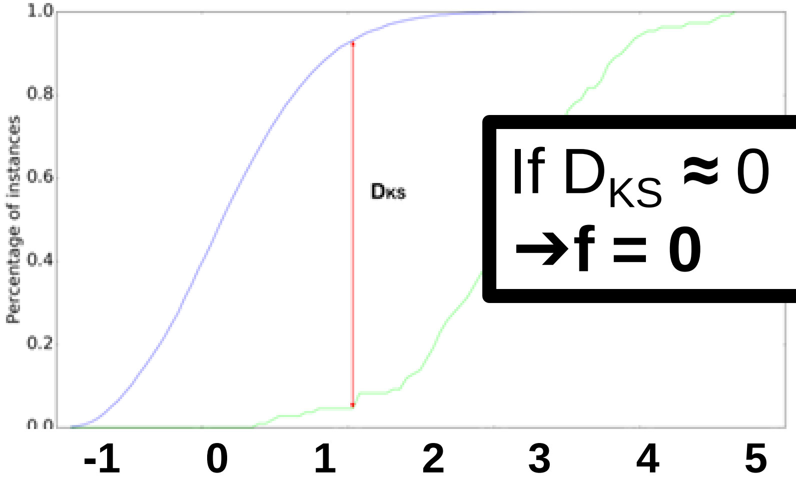


Meaningful discretization





Real values



If $D_{KS} \approx 0$

→f = 0

Komogrov-Smirnov distance

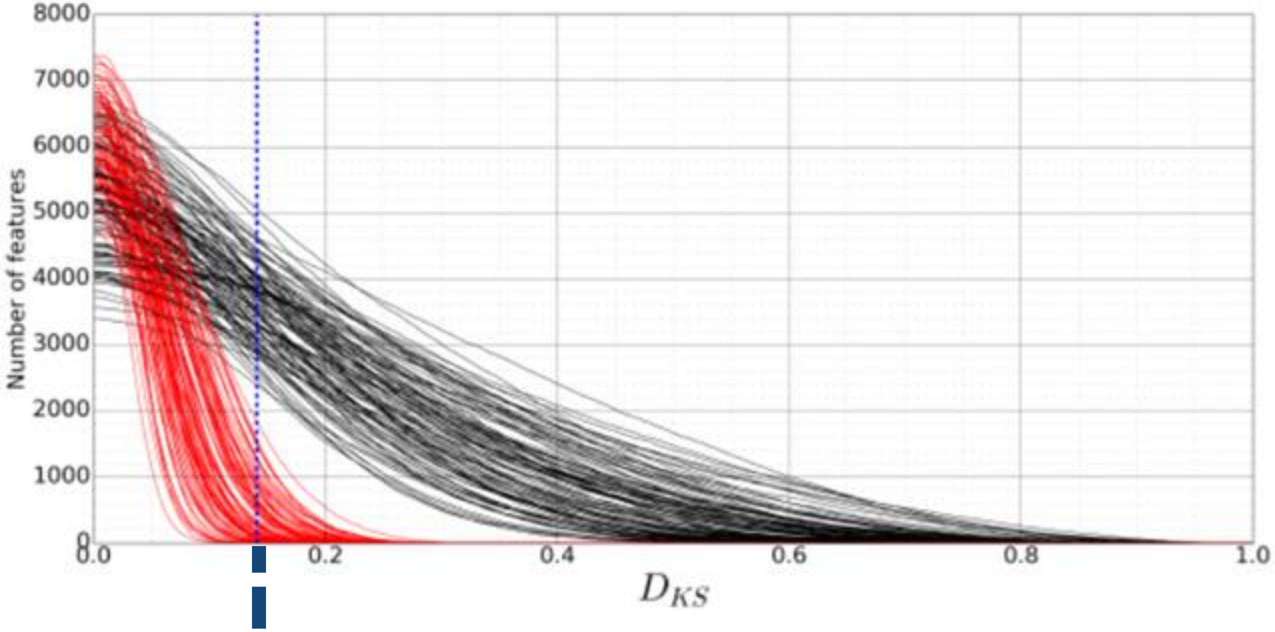
Finding appropriate D_{KS} thresholds

Probabilistically

Feature-class pairs above D_{KS}

Target labels

Randomized target labels



f = 0 ← Feature is more likely to behave randomly

Feature is more likely to have meaning for some class → f = 1

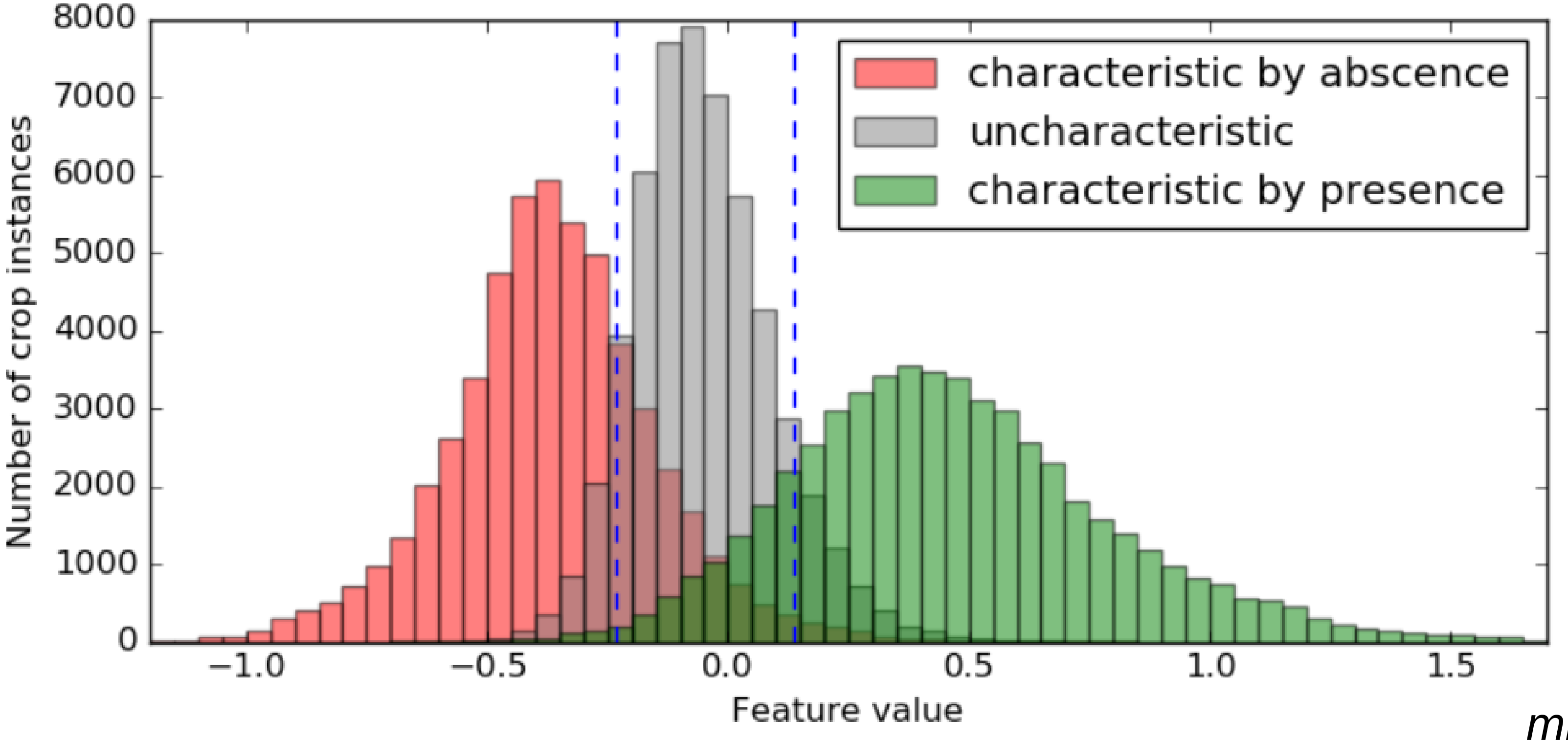
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Using D_{KS} thresholds

- We have the thresholds, but…
  - They are **expensive** to compute
  - They are **class dependent. Supervised.**
- We would like to have a threshold on **feature value**, not on D_{KS} (feature-class).
- We would like the threshold to be **unsupervised**.

Using D_{KS} thresholds

- **Probabilistically**
  - Find distributions of **feature values** conditioned to D_{KS} thresholds.
  - Find the thresholds that best separate them.



mit67

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Using D_{KS} thresholds

- **Probabilistically**
  - Find distributions of feature values conditioned to D_{KS} thresholds.
  - Find the thresholds that best separate them.
  - Solution:

Dataset	ft^+	ft^-
mit67	0.14	-0.23
cub200	0.20	-0.24
flowers102	0.15	-0.24

Negative feature threshold = -0.25

Positive feature threshold = 0.15

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Using D_{KS} thresholds

- **Probabilistically**
  - Find distributions of feature values conditioned to D_{KS} thresholds.
  - Find the thresholds that best separate them.
  - Solution:

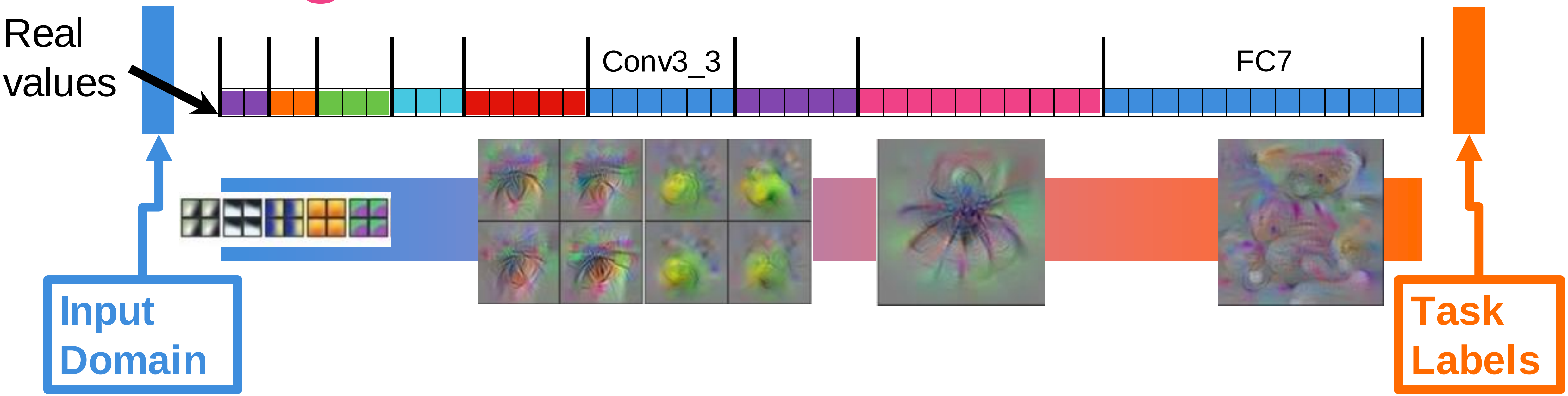
Dataset	ft^+	ft^-
mit67	0.14	-0.23
cub200	0.20	-0.24
flowers102	0.15	-0.24

Negative feature threshold = -0.25

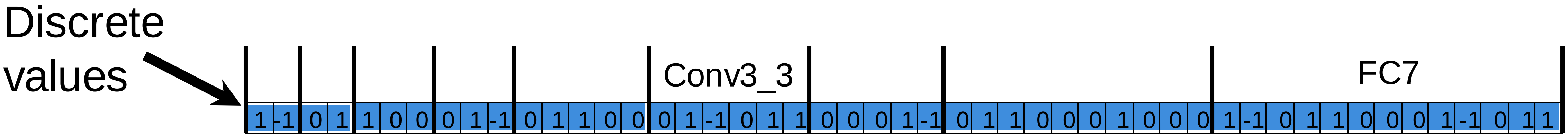
Positive feature threshold = 0.15

Good for all datasets!

Meaningful discretization



Negative feature threshold = -0.25
Positive feature threshold = 0.15

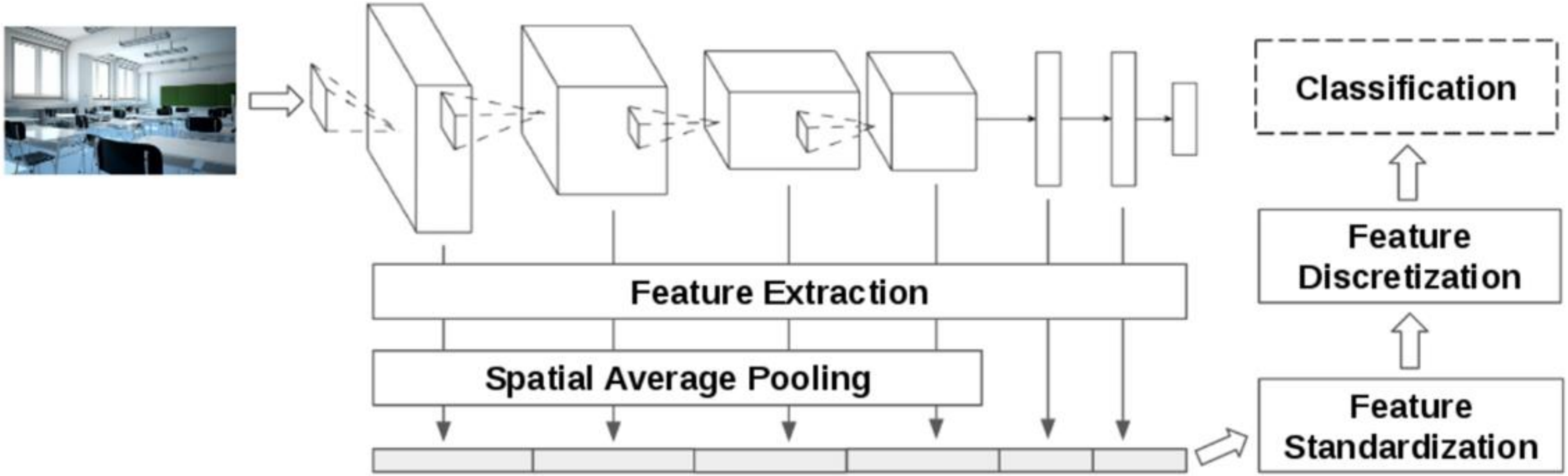


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Full-Network Embedding Recipe

1. Spatial Average Pooling
2. Standardisation
3. Discretization

Full Network embedding



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

FNE – Small datasets

Dataset	#Images	#Classes	#Images (train)	#Images (test)	#Images per class	#Images per class (train)	#Images per class (test)
mit67	6,700	67	5,360	1,340	100	77 - 83	17 - 23
cub200	11,788	200	5,994	5,794	41 - 60	29 - 30	12 - 30
flowers102	8,189	102	2,040	6,149	40 - 258	20	20 - 238
cats-dogs	7,349	37	3,680	3,669	184 - 200	93 - 100	88 - 100
sdogs	20,580	120	12,000	8,580	150 - 200	100	50 - 100
caltech101	9,146	101	3,060	2,995	31 - 800	30	1 - 50
food101	25,250	101	20,200	5,050	250	200	50
textures	5,640	47	3,760	1,880	120	80	40
wood	438	7	350	88	14 - 179	10 - 142	3 - 37

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

FNE – Small datasets

Dataset	#Images	#Classes	#Images (train)	#Images (test)	#Images per class	#Images per class (train)	#Images per class (test)
mit67	6,700	67	5,360	1,340	100	77 - 83	17 - 23
cub200	11,788	200	5,994	5,794	41 - 60	29 - 30	12 - 30
flowers102	8,189	102	2,040	6,149	40 - 258	20	20 - 238
cats-dogs	7,349	37	3,680	3,669	184 - 200	93 - 100	88 - 100
sdogs	20,580	120	12,000	8,580	150 - 200	100	50 - 100
caltech101	9,146	101	3,060	2,995	31 - 800	30	1 - 50
food101	25,250	101	20,200	5,050	250	200	50
textures	5,640	47	3,760	1,880	120	80	40
wood	438	7	350	88	14 - 179	10 - 142	3 - 37

10 – 200 images/class

FNE - Results: Similar source task

Network pre-trained on **Places2** for mit67 and on **ImageNet** for the rest.

Dataset	mit67	cub200	flowers102	cats-dogs	sdogs	caltech101	food101	textures	wood
Baseline `fc6`	80.0	65.8	89.5	89.3	78.0	91.4±0.6	61.4±0.2	69.6	70.8±6.6
Baseline `fc7`	81.7	63.2	87.0	89.6	79.3	89.7±0.3	59.1±0.6	69.0	68.9 ±6.8
Full-network	83.6	65.5	93.3	89.2	78.8	91.4±0.6	67.0±0.7	73.0	74.1±6.9
SotA	86.9 [5]	92.3 [10]	97.0 [5]	91.6 [6]	90.3 [5]	93.4 [31]	77.4 [4]	75.5 [17]	- -
ED	✓	✓	✓	✗	✓	✗	✗	✗	-
FT	✓	✓	✓	✓	✓	✓	✓	✗	-

FNE - Results: Similar source task

Best case scenario!

Network pre-trained on **Places2** for mit67 and on **Image**

Dataset	mit67	cub200	flowers102	cats-dogs	sdogs	caltech101	food101	textures	wood
Baseline `fc6`	80.0	65.8	89.5	89.3	78.0	91.4±0.6	61.4±0.2	69.6	70.8±6.6
Baseline `fc7`	81.7	63.2	87.0	89.6	79.3	89.7±0.3	59.1±0.6	69.0	68.9 ±6.8
Full-network	83.6	65.5	93.3	89.2	78.8	91.4±0.6	67.0±0.7	73.0	74.1±6.9
SotA	86.9 [5]	92.3 [10]	97.0 [5]	91.6 [6]	90.3 [5]	93.4 [31]	77.4 [4]	75.5 [17]	- -
ED	✓	✓	✓	✗	✓	✗	✗	✗	-
FT	✓	✓	✓	✓	✓	✓	✓	✗	-

FNE - Results: Similar source task

Network pre-trained on **Places2** for mit67 and on **ImageNet** for the rest.

Dataset	mit67	cub200	flowers102	cats-dogs	sdogs	caltech101	food101	textures	wood
Baseline fc6	80.0	65.8	89.5	89.3	78.0	91.4±0.6	61.4±0.2	69.6	70.8±6.6
Baseline fc7	81.7	63.2	87.0	89.6	79.3	89.7±0.3	59.1±0.6	69.0	68.9 ±6.8
Full-network	83.6	65.5	93.3	89.2	78.8	91.4±0.6	67.0±0.7	73.0	74.1±6.9
SotA	86.9 [5]	92.3 [10]	97.0 [5]	91.6 [6]	90.3 [5]	93.4 [31]	77.4 [4]	75.5 [17]	- -
ED	✓	✓	✓	✗	✓	✗	✗	✗	-
FT	✓	✓	✓	✓	✓	✓	✓	✗	-

+2.9

+4.2

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

FNE - Results: Similar source task

Network pre-trained on **Places2** for mit67 and on **ImageNet** for the rest.

Dataset	mit67	cub200	flowers102	cats-dogs	sdogs	caltech101	food101	textures	wood	
Baseline fc6	80.0	65.8	89.5	89.3	78.0	91.4±0.6	61.4±0.2	69.6	70.8±6.6	+2.9
Baseline fc7	81.7	63.2	87.0	89.6	79.3	89.7±0.3	59.1±0.6	69.0	68.9 ±6.8	+4.2
Full-network	83.6	-0.3	93.3	-0.4	-0.5	91.4±0.6	67.0±0.7	73.0	74.1±6.9	
SotA	86.9 [5]	92.3 [10]	97.0 [5]	91.6 [6]	90.3 [5]	93.4 [31]	77.4 [4]	75.5 [17]	- -	
ED	✓	✓	✓	✗	✓	✗	✗	✗	-	
FT	✓	✓	✓	✓	✓	✓	✓	✗	-	

FNE - Results: Similar source task

Network pre-trained on **Places2** for mit67 and on **ImageNet** for the rest.

Dataset	mit67	cub200	flowers102	cats-dogs	sdogs	caltech101	food101	textures	wood	
Baseline fc6	80.0	65.8	89.5	89.3	78.0	91.4±0.6	61.4±0.2	69.6	70.8±6.6	+2.9
Baseline fc7	81.7	63.2	87.0	89.6	79.3	89.7±0.3	59.1±0.6	69.0	68.9 ±6.8	+4.2
Full-network	83.6	-0.3	93.3	-0.4	-0.5	91.4±0.6	67.0±0.7	73.0	74.1±6.9	
SotA	86.9 [5]	92.3 [10]	97.0 [5]	91.6 [6]	90.3 [5]	93.4 [31]	77.4 [4]	75.5 [17]	- -	
ED	✓	✓	✓	✗	✓	✗	✗	✗	-	
FT	✓	✓	✓	✓	✓	✓	✓	✗	-	

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

FNE - Results: Dissimilar source task

Network pre-trained on **ImageNet** for mit67 and on **Places2** for the rest.

Dataset	mit67	cub200	flowers102	cats-dogs	caltech101	textures	wood
Baseline `fc7`	72.2	23.6	73.3	38.7	72.0	55.8	65.3
Full-network	75.5	35.5	88.7	56.2	80.0	65.1	74.0

FNE - Results: Dissimilar source task

Most frequent real-world scenario!

Network pre-trained on **ImageNet** for mit67 and on **Places2** for the rest.

Dataset	mit67	cub200	flowers102	cats-dogs	caltech101	textures	wood
Baseline fc7	72.2	23.6	73.3	38.7	72.0	55.8	65.3
Full-network	75.5	35.5	88.7	56.2	80.0	65.1	74.0

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

FNE - Results: Dissimilar source task

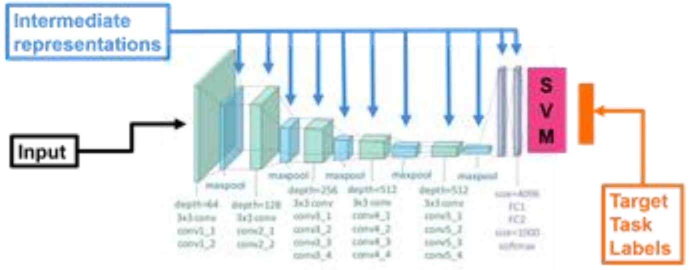
Network pre-trained on **ImageNet** for mit67 and on **Places2** for the rest.

Dataset	mit67	cub200	flowers102	cats-dogs	caltech101	textures	wood
Baseline fc7	72.2	23.6	73.3	38.7	72.0	55.8	65.3
Full-network	75.5	35.5	88.7	56.2	80.0	65.1	74.0
	+3.3	+11.9	+15.4	+17.5	+8.0	+9.3	+10.6

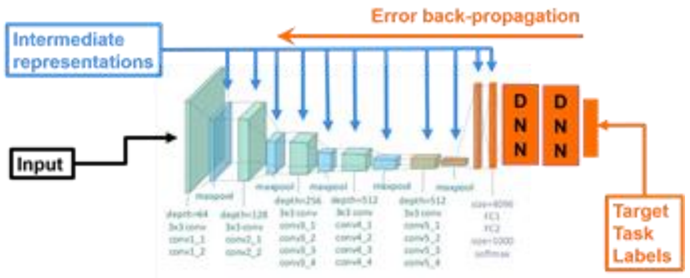
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Simple solutions

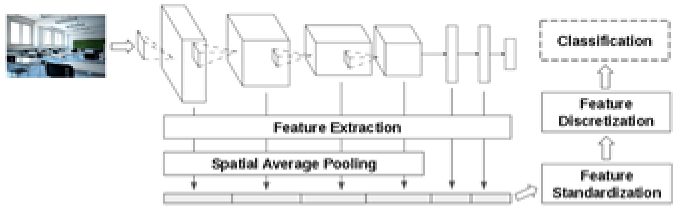
- DNN last layer features + SVM
  (Feature extraction)
  - Similar task and domain



- Add one or several NN layers +
  Fine-tuning pre-trained layers
  - Enough data



- Full Network Embedding
  - Robust to different task and domain
  - Works with little data



thanks.

armand.vilalta@bsc.es