Visual Learning (Tutorial C1) Introduction

Instructor - Simon Lucey

RVSS - 2024

THE UNIVERSITY of ADELAIDE

AUSTRALIAN INSTITUTE FOR MACHINE LEARNING

User: What is unusual about this image?

GPT4: The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

M. Oquab, et al. "DINOv2: Learning robust visual features without supervision." In arXiv 2023.

A. Kirillov, et al. "Segment Anything.” In CVPR 2023.

Visual Learning is Fundamental to Robotics!!!


GO
STOP
0 MPH
3 FEET
2 MPH
10 FEET

Today

- **Shallow Deep Visual Learning**
- Deep Visual Learning
- ConvNets and AlexNet

Visual Learning – A Sampling Problem?

$$[\mathbf{x}_1, \ldots, \mathbf{x}_N]$$

$$s(\mathbf{X})$$

$$[\mathbf{y}_1, \ldots, \mathbf{y}_N]$$

$$s(\mathbf{X}) \approx \mathcal{F}(\mathbf{X}; \boldsymbol{\theta})$$

$$\boldsymbol{\theta} = [\theta_1, \ldots, \theta_M]^T$$

Visual Learning – A Generalization Problem?

$\mathbf{x}$

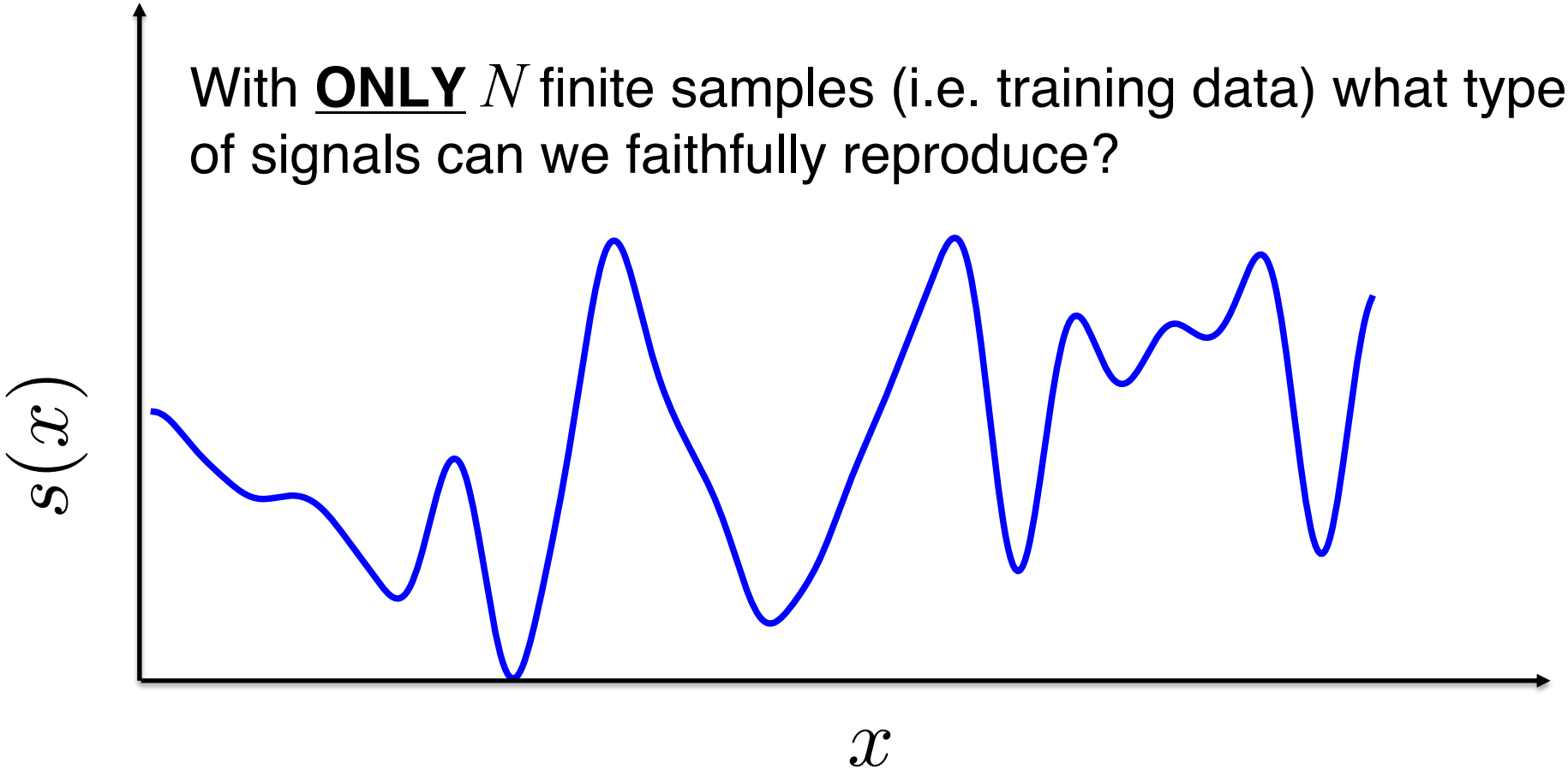
$\Rightarrow \mathcal{F}(\mathbf{x};\boldsymbol{\theta}) \Rightarrow$

?

$\mathbf{y}$

Sampling in 1D - Question?

With **ONLY** N finite samples (i.e. training data) what type of signals can we faithfully reproduce?



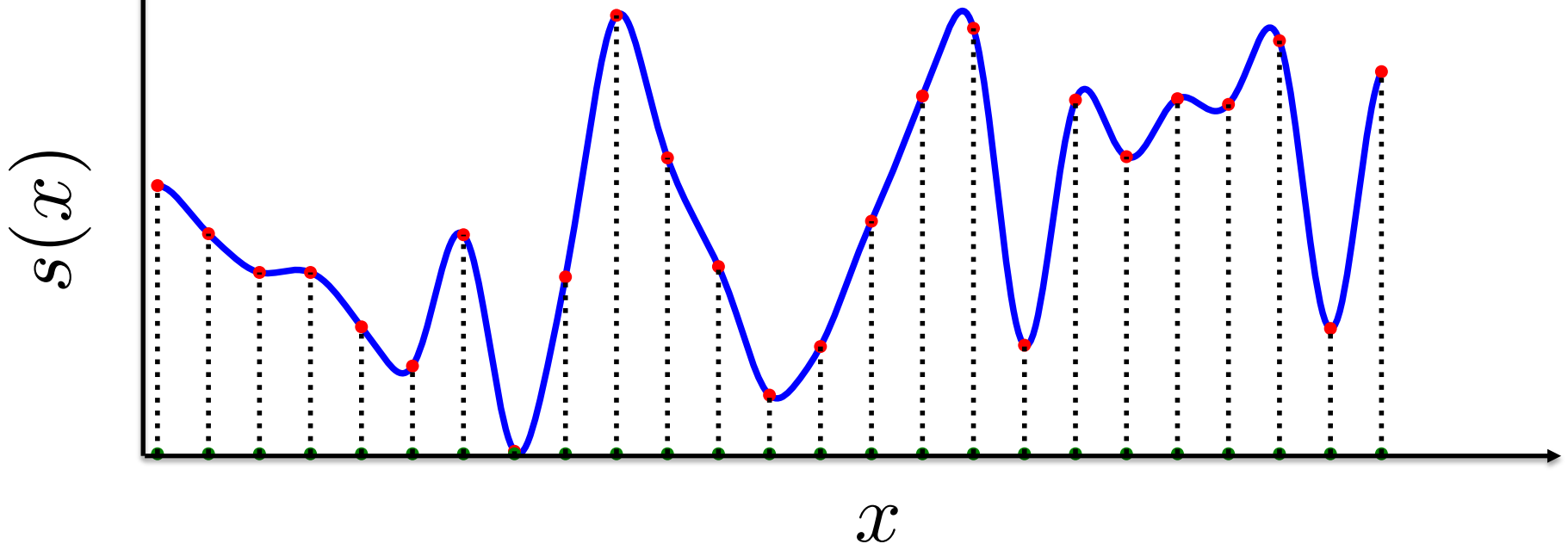
# of hidden units M

$$s(x) \approx \sum_{m=1}^{M} \theta_m \cdot \eta(x - \delta_m)$$

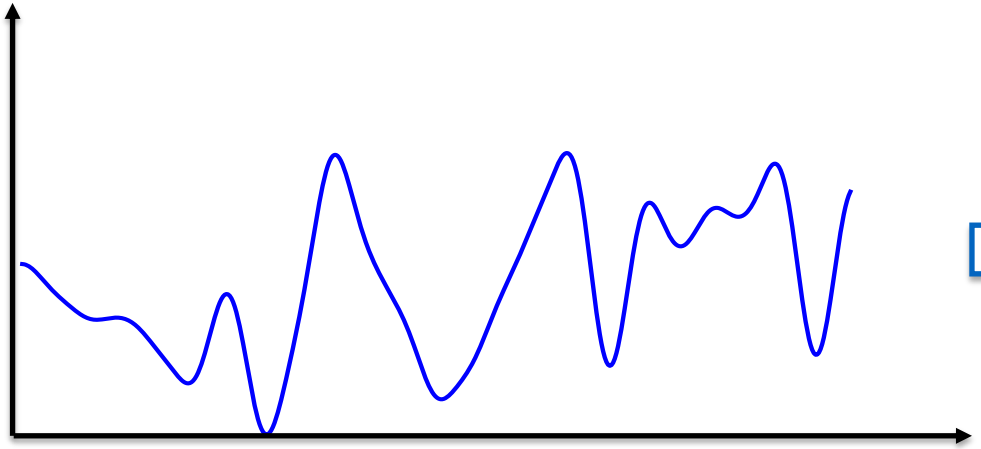
"learned" θ_m

"bias" δ_m

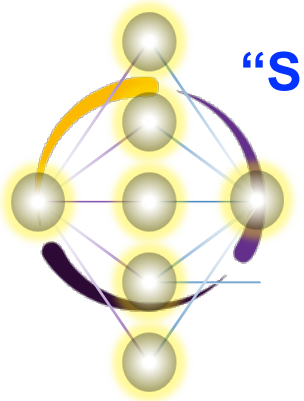
activation function η



C. E. Shannon. "Communication in the presence of noise." Proceedings of the IRE, 37(1):10–21, 1949.

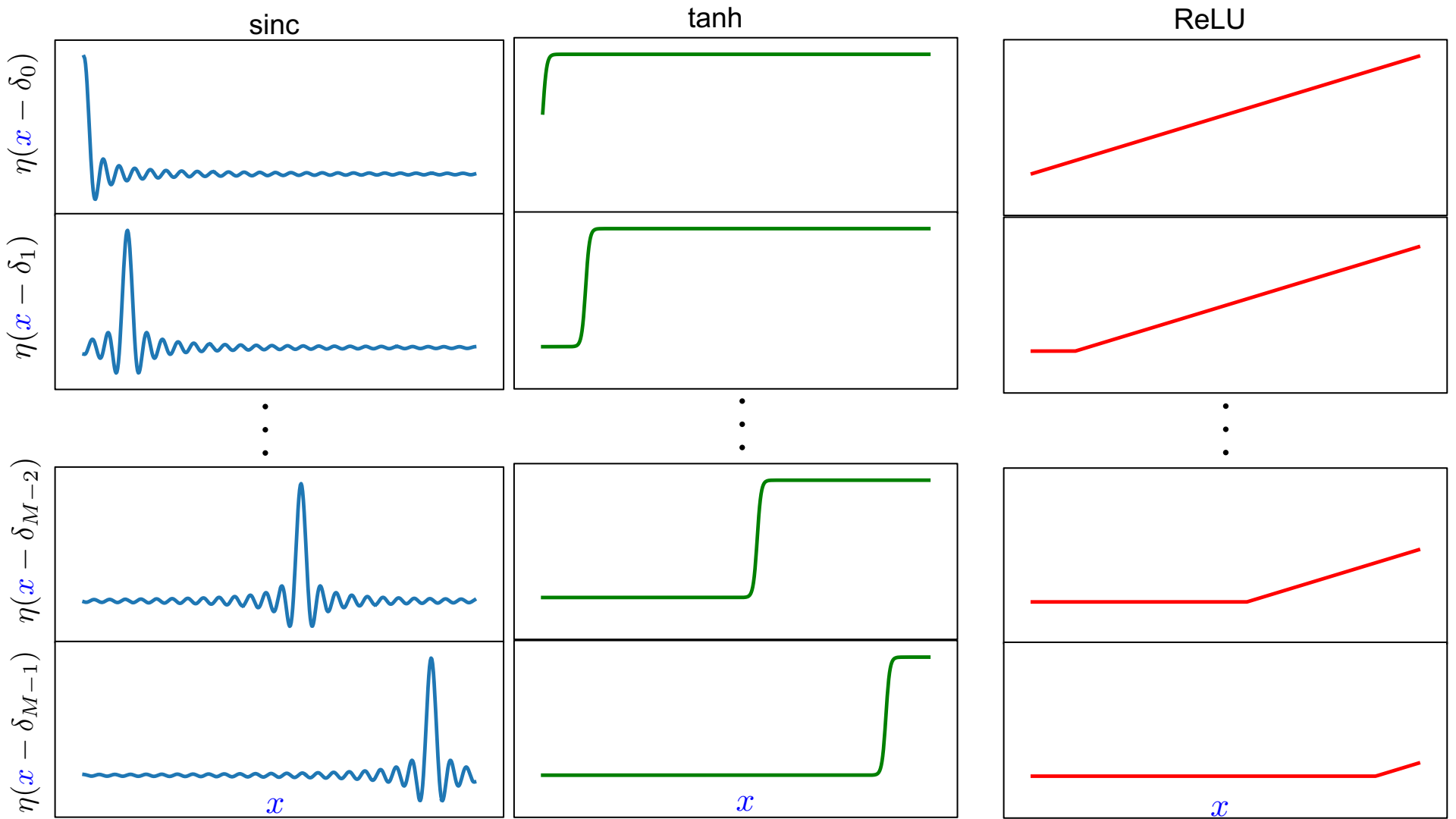


$$\Rightarrow \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_M \end{bmatrix}$$

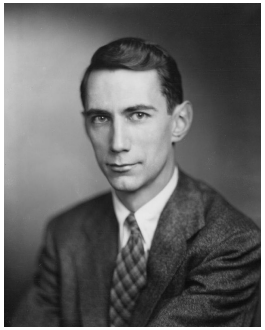


“Shallow Network”

x $\mathcal{F}(x;\boldsymbol{\theta})$


sinc
tanh
ReLU
$\eta(x-\delta_0)$
$\eta(x-\delta_1)$
$\eta(x-\delta_{M-2})$
$\eta(x-\delta_{M-1})$
x
x
x

Universal Function Approximator?



- To memorize N samples, one must satisfy,

"No. of hidden units"

$$\boxed{M} \geq \text{rank}(\mathbf{A}) \geq \boxed{N}$$

where,

"No. of samples"

$$\mathbf{A}[i,j] = \int_x \eta(x - \delta_i) \cdot \eta(x - \delta_j) \cdot \mathrm{d}x$$

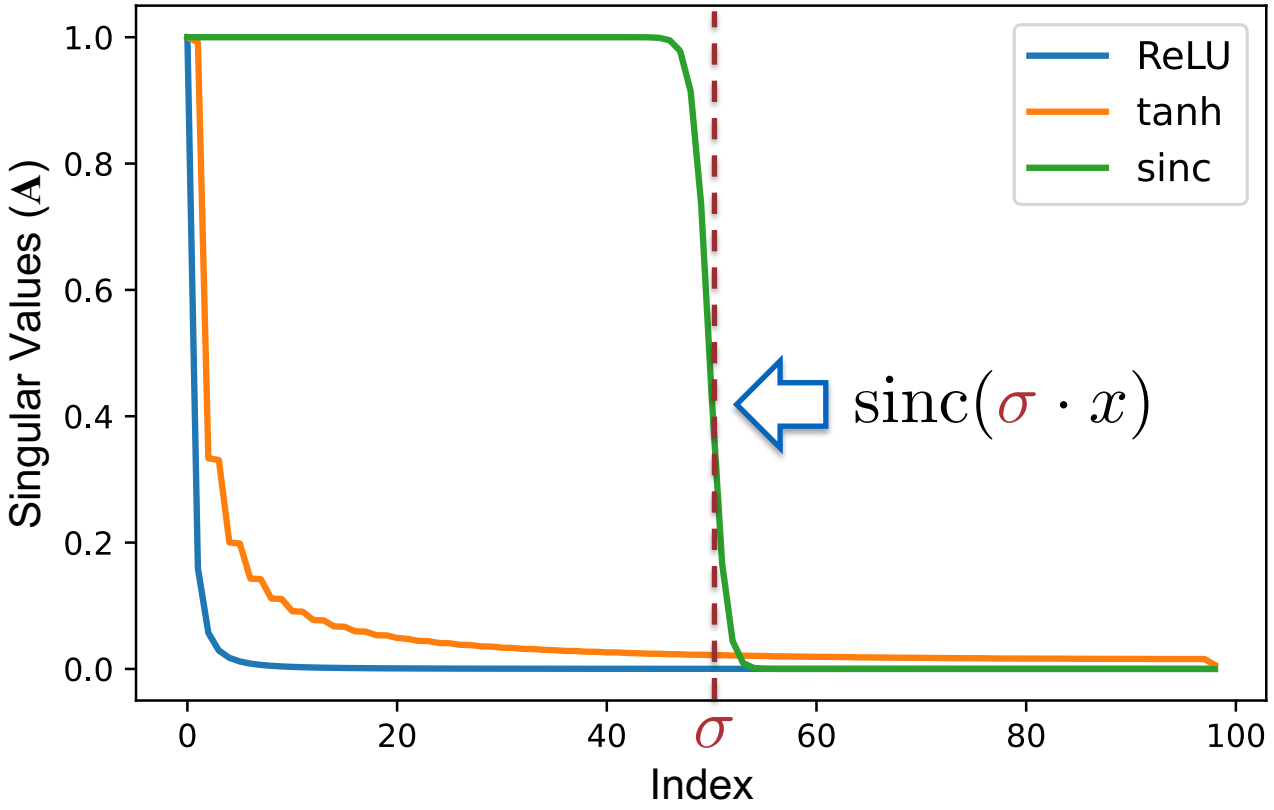
$$\eta(x) = \sinh(\boxed{\sigma} \cdot x)$$

"Hyper-parameter"

C. E. Shannon. "Communication in the presence of noise." Proceedings of the IRE, 37(1):10–21, 1949.

G. Cybenko "Approximation by superpositions of a sigmoidal function." Math. Control. Signals Syst. 2(4): 303-314 (1989)

Universal Function Approximator?



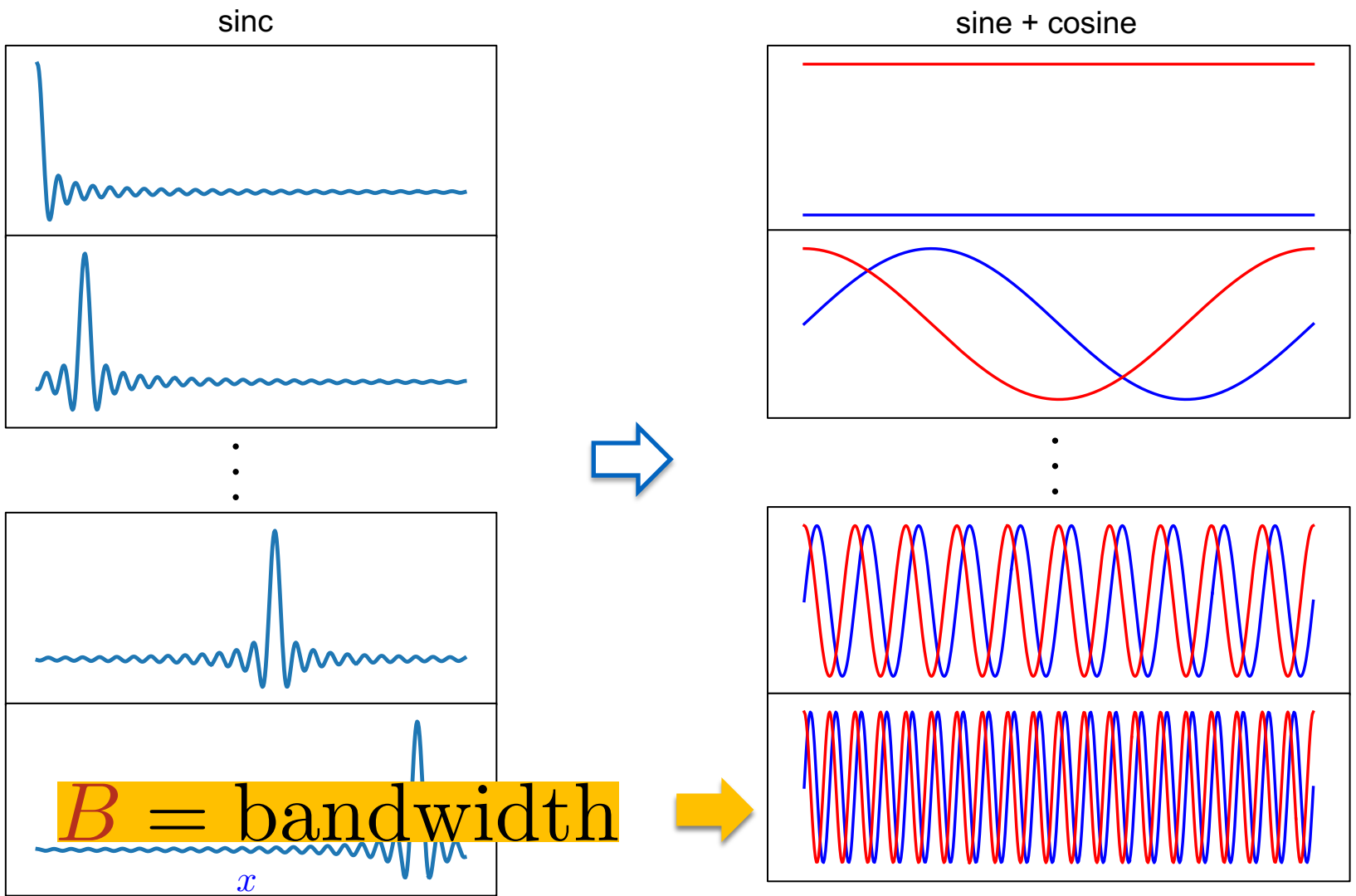
Only for sinc,

$$\sigma = 2 \times B$$

$$B = \text{bandwidth}$$

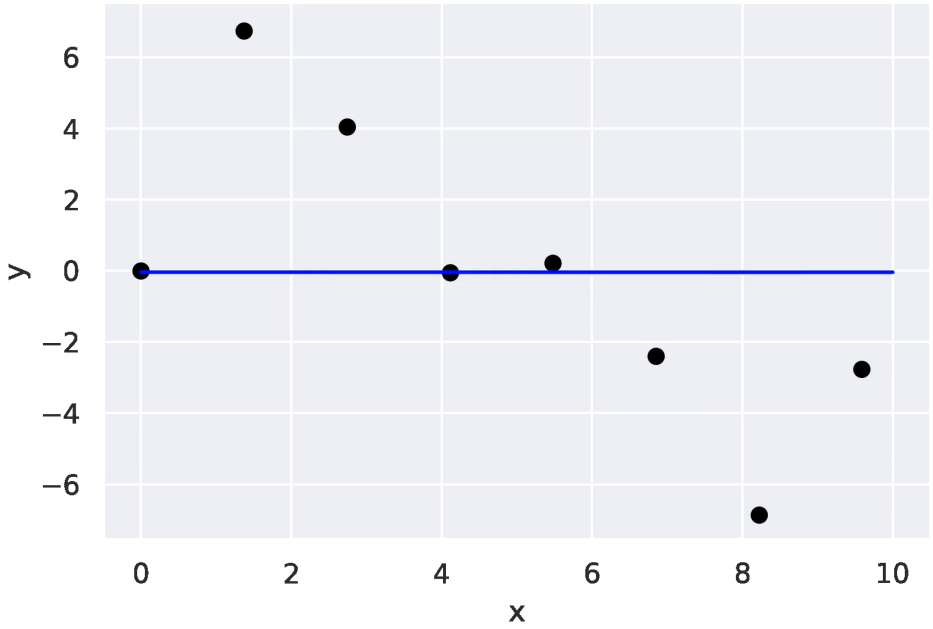
In general,

$$B \geq \text{area under curve}$$

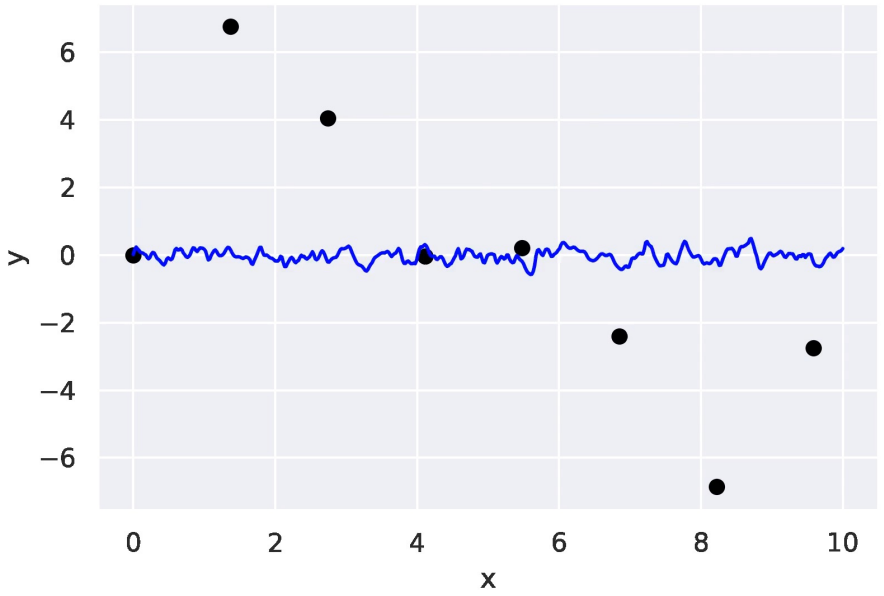

sinc
sine + cosine
B = bandwidth
x

Memorization vs. Generalization

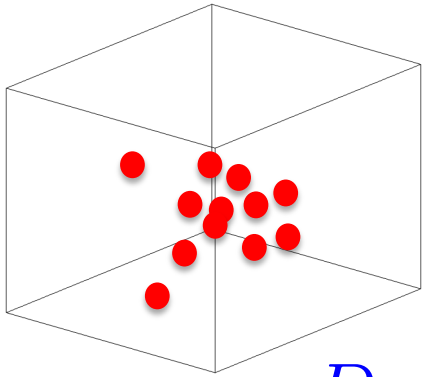
Balanced B



Overfitting B



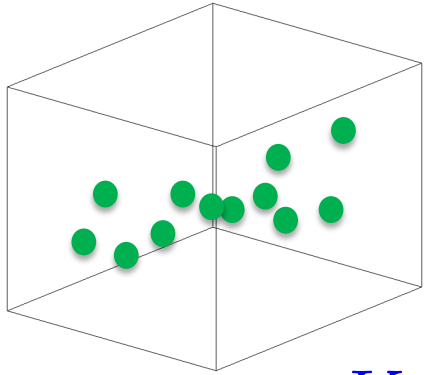
Multi-dimensional Sampling?



$\mathbf{x} \in \mathbb{R}^D$

$S(\mathbf{X})$

$\mathbb{R}^D : \mathbb{R}^K$

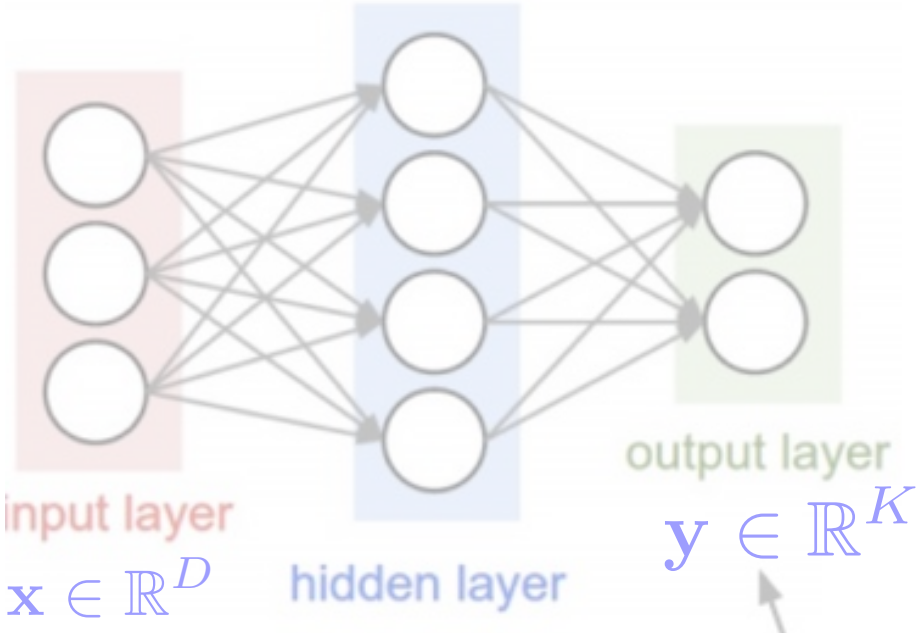


$\mathbf{y} \in \mathbb{R}^K$

"learned"

$$s(\mathbf{x}) \approx \boldsymbol{\theta}\eta(\mathbf{W}\mathbf{x} + \boldsymbol{\delta})$$

"hidden layer"



$$s(\mathbf{x}) \approx \underset{(K \times M)}{\boldsymbol{\theta}} \eta\Big(\underset{(M \times D)}{\mathbf{W}}\underset{(D \times 1)}{\mathbf{x}} + \underset{(M \times 1)}{\boldsymbol{\delta}}\Big)$$

Today

- Shallow Deep Visual Learning
- **Deep Visual Learning**
- ConvNets and AlexNet

Shallow vs Deep Network

- A shallow learner has **only one** hidden unit,

$$\boldsymbol{\theta}\eta(\mathbf{W}\mathbf{x} + \boldsymbol{\delta})$$

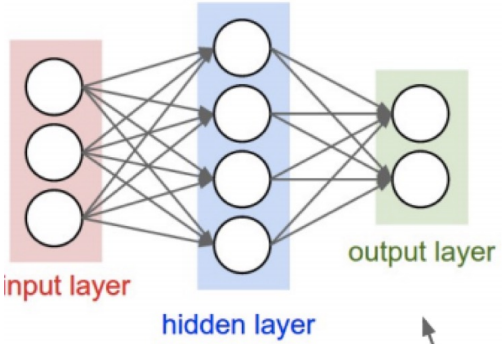
- A deep learner **has more than one** hidden unit,

$$\boldsymbol{\theta}\eta(\mathbf{W}^{(1)}\eta(\mathbf{W}^{(0)}\mathbf{x} + \boldsymbol{\delta}^{(0)}) + \boldsymbol{\delta}^{(1)})$$

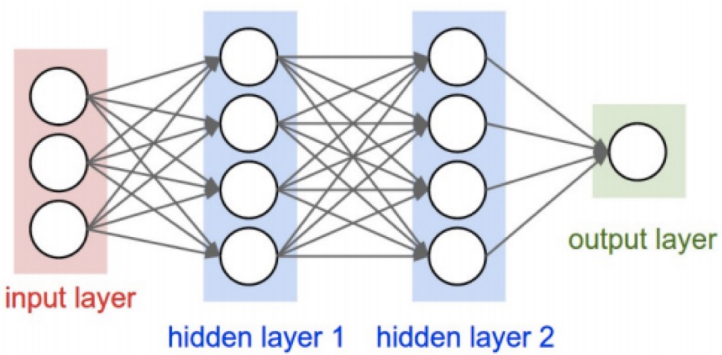
"Geoffrey Hinton"

Shallow vs. Deep Layers

“shallow”



“deep”



“2-layer neural net,” or
“1-hidden-layer neural net”

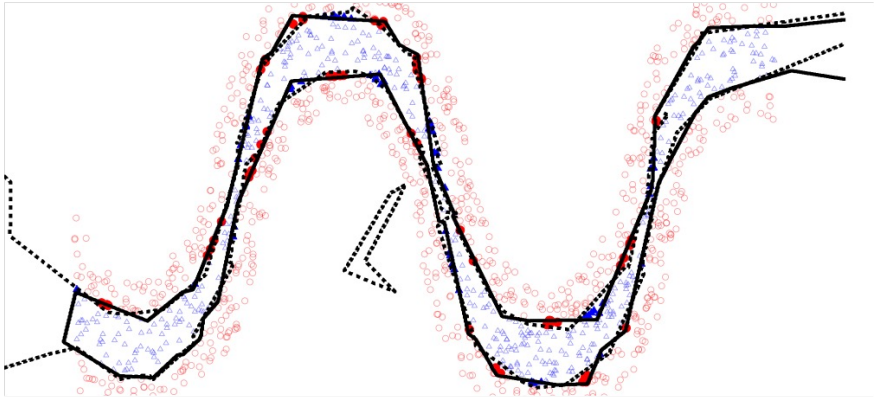
“Fully-connected” layers

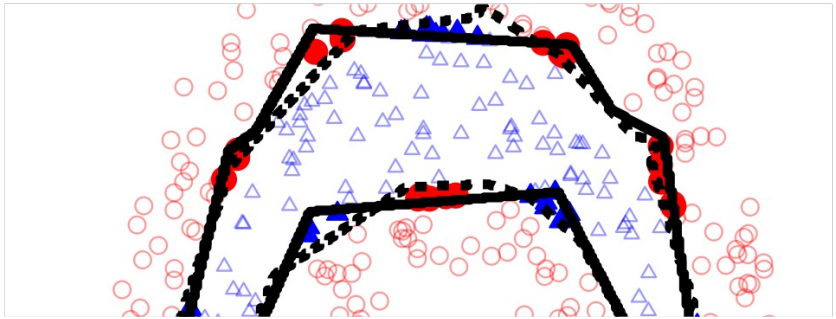
“3-layer neural net,” or
“2-hidden-layer neural net”

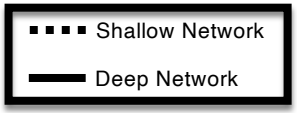
Why deepness?

Why Deep?

- Deeper nets are exponentially more expressive than shallow ones.

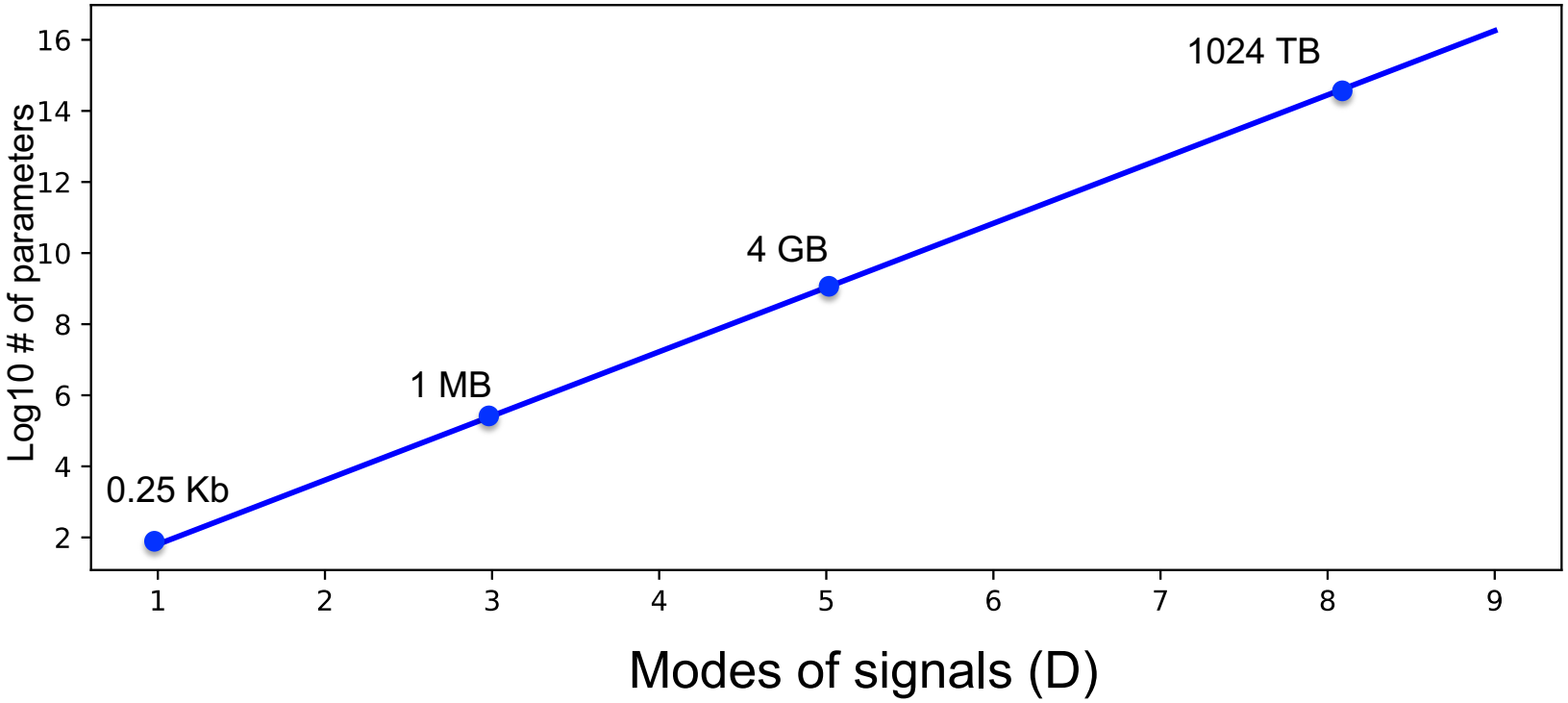






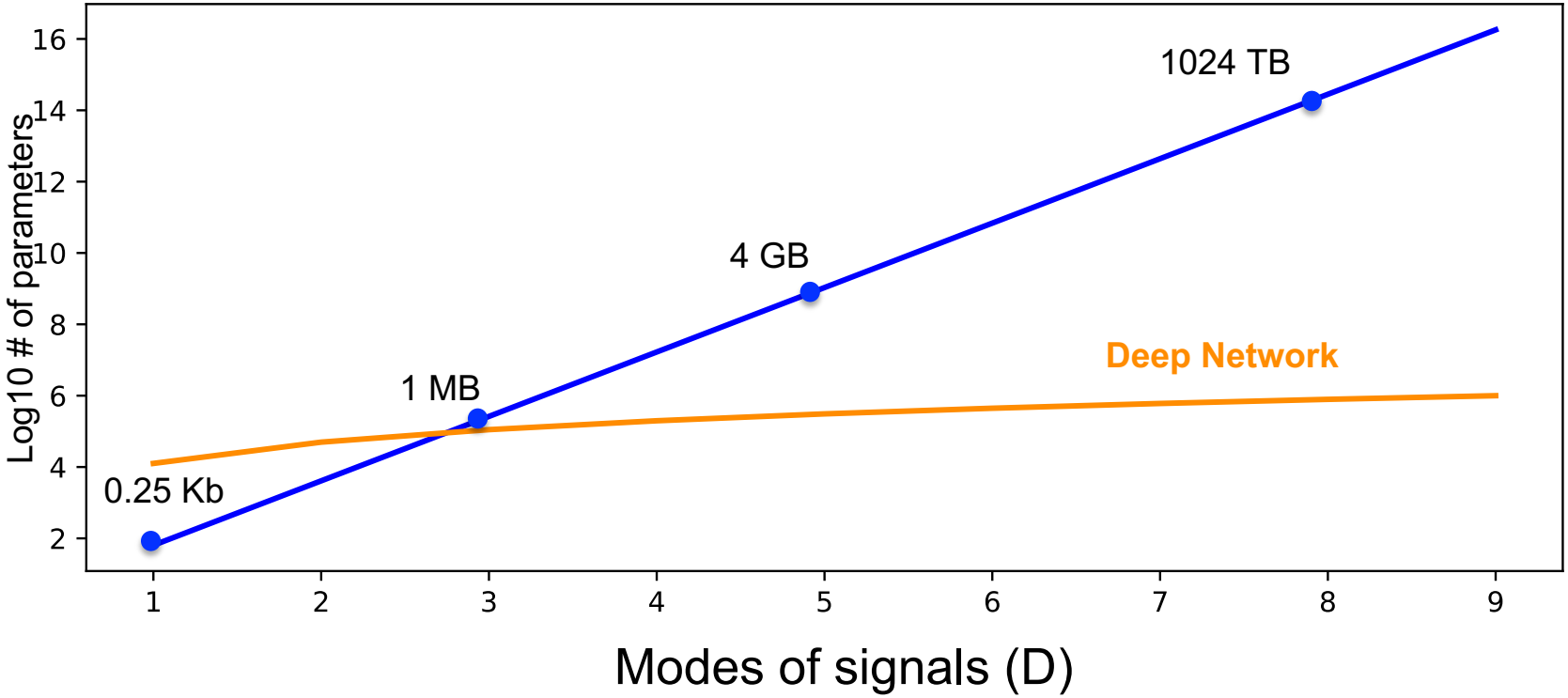
Montufar, Guido F., et al. "On the number of linear regions of deep neural networks." NeurIPS 2014.

Curse of Dimensionality (B = 32)



$$\text{\# of parameters} = (2B)^D$$

Curse of Dimensionality (B = 32)



Does this benefit apply to all signals?

Let's have a play!!!

https://colab.research.google.com/github/slucey-cs-cmu-edu/RVSS24/blob/main/MLP_Example.ipynb

Some things to try!!!

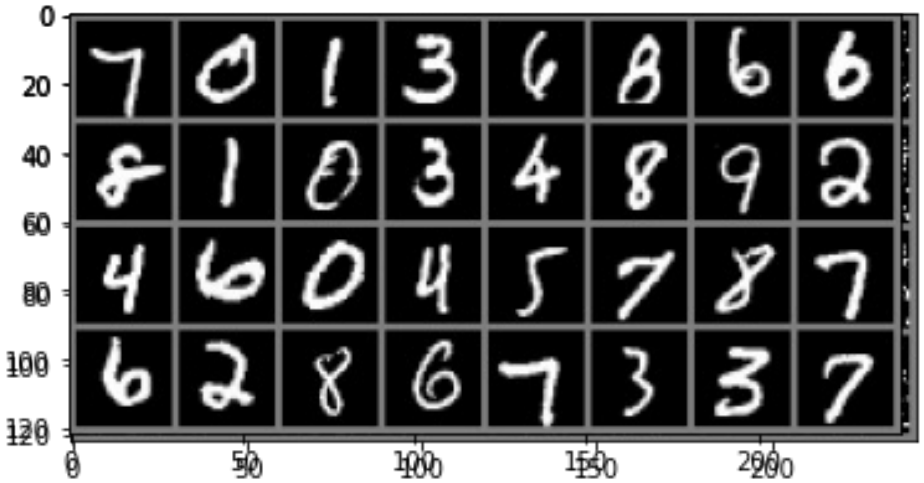
- What happens to performance if you use a linear function?

```
class MLPNet(nn.Module):
    def __init__(self):
        super(MLPNet, self).__init__()

        # First fully connected layers input image is 28x28 = 784 dim.
        self.fc0 = nn.Linear(784, 10) # nparam = 784*256 = 38400
        # Two more fully connected layers
        #self.fc1 = nn.Linear(256, 84)
        #self.fc2 = nn.Linear(84, 10)

    def forward(self, x):
        # Flattens the image like structure into vectors
        x = torch.flatten(x, start_dim=1)

        # fully connected layers with activations
        x = self.fc0(x)
        #x = F.relu(x)
        #x = self.fc1(x)
        #x = F.relu(x)
        #x = self.fc2(x)
        # Outputs are log(p) so softmax followed by log.
        #return(x)
        output = F.log_softmax(x, dim=1)
        return output
```

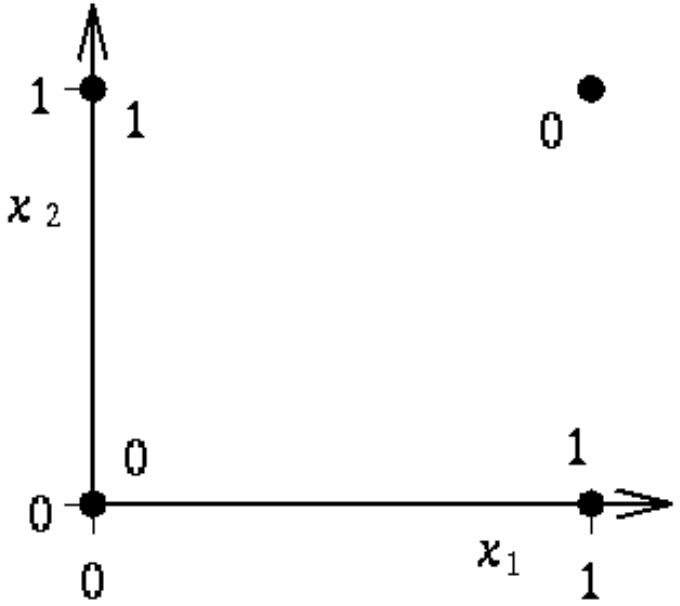
Some things to try!!!

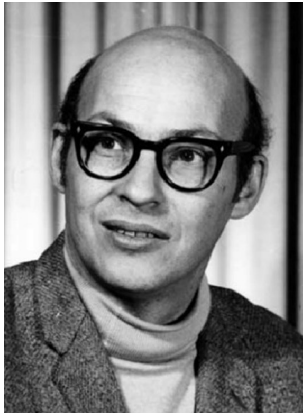
- What happens to performance if you use a linear function?
- What happens when you permute the pixels?

```
from numpy.random import permutation
idx_permute = torch.from_numpy(permutation(784))
transform = transforms.Compose([transforms.ToTensor(),
                                transforms.Lambda(lambda x: x.view(-1)[idx_permute].view(1, 28, 28) ),
                                transforms.Normalize((0.5,), (0.5,)),
                                ])
```

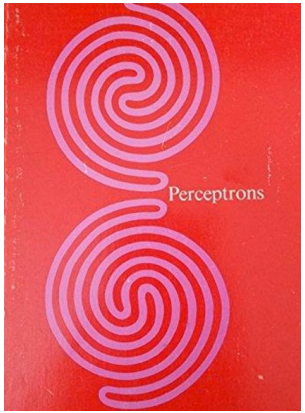


What about XORs?

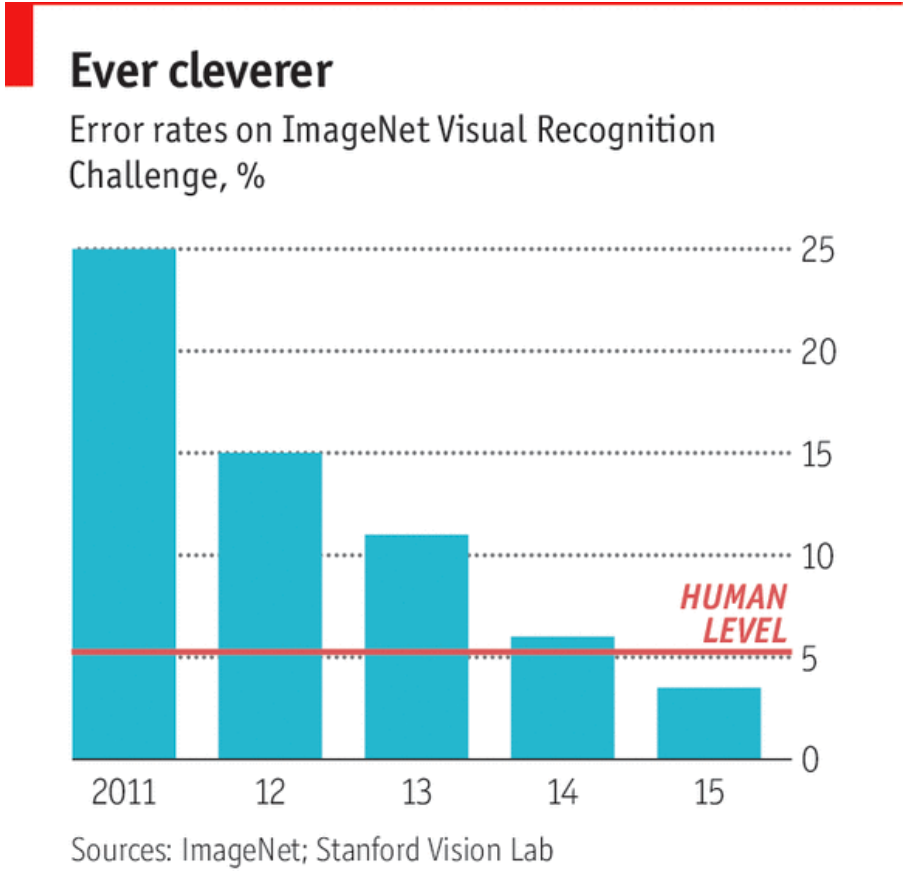




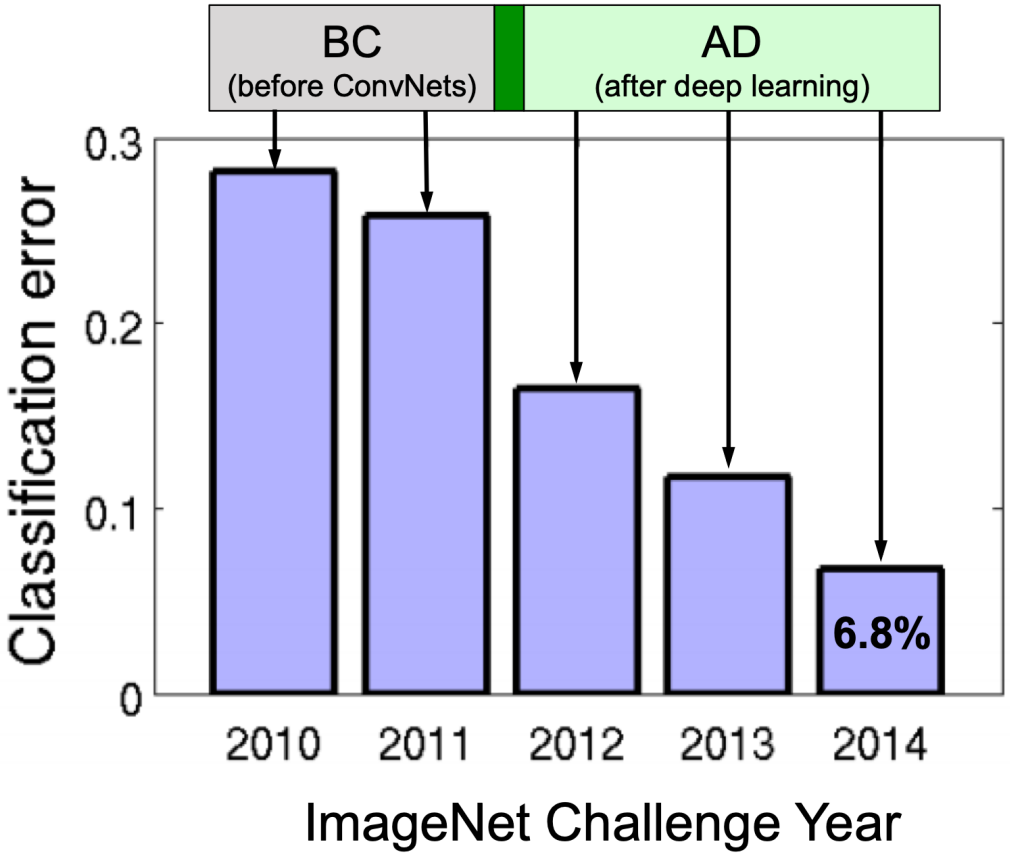
“Marvin Minsky”



Deep Learning – A Breakthrough!!



Impact on Object Recognition



ImageNet

- Over 15M labeled high resolution images.
- Roughly 22K categories.
- Collected from web and labeled by Amazon Mechanical Turk.

Example in Code

```
class MLPNet(nn.Module):
    def __init__(self):
        super(MLPNet, self).__init__()

        # First fully connected layers input image is 28x28 = 784 dim.
        self.fc0 = nn.Linear(784, 256) # nparam = 784*256 = 38400
        # Two more fully connected layers
        self.fc1 = nn.Linear(256, 84)
        self.fc2 = nn.Linear(84, 10)

    def forward(self, x):
        # Flattens the image like structure into vectors
        x = torch.flatten(x, start_dim=1)

        # fully connected layers with activations
        x = self.fc0(x)
        x = F.relu(x)
        x = self.fc1(x)
        x = F.relu(x)
        x = self.fc2(x)
        # Outputs are log(p) so softmax followed by log.
```

$$\eta(\mathbf{W}^{(1)}\eta(\mathbf{W}^{(0)}\mathbf{x}+\boldsymbol{\delta}^{(0)})+\boldsymbol{\delta}^{(1)})$$

Example in Code

```
class MLPNet(nn.Module):
    def __init__(self):
        super(MLPNet, self).__init__()

        # First fully connected layers input image is 28x28 = 784 dim.
        self.fc0 = nn.Linear(784, 256) # nparam = 784*256 = 38400
        # Two more fully connected layers
        self.fc1 = nn.Linear(256, 84)
        self.fc2 = nn.Linear(84, 10)

    def forward(self, x):
        # Flattens the image like structure into vectors
        x = torch.flatten(x, start_dim=1)

        # fully connected layers with activations
        x = self.fc0(x)
        x = F.relu(x)
        x = self.fc1(x)
        x = F.relu(x)
        x = self.fc2(x)
        # Outputs are log(p) so softmax followed by log.
```

$$\eta(\mathbf{W}^{(1)}\eta(\mathbf{W}^{(0)}\mathbf{x}+\boldsymbol{\delta}^{(0)})+\boldsymbol{\delta}^{(1)})$$

Why the Non-Linearity?

$$\eta(\mathbf{W}^{(2)}\eta(\mathbf{W}^{(1)}\mathbf{x} + \boldsymbol{\delta}^{(1)}) + \boldsymbol{\delta}^{(2)})$$

Mysteries still remain?

DOI:10.1145/3446776

Understanding Deep Learning (Still) Requires Rethinking Generalization

By Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals

Abstract

Despite their massive size, successful deep artificial neural networks can exhibit a remarkably small gap between training and test performance. Conventional wisdom attributes small generalization error either to properties of the model family or to the regularization techniques used during training.

Through extensive systematic experiments, we show how these traditional approaches fail to explain why large neural networks generalize well in practice. Specifically, our experiments establish that state-of-the-art convolutional networks for image classification trained with stochastic gradient methods easily fit a random labeling of the training data. This phenomenon is qualitatively unaffected by explicit regularization and occurs even if we replace the true images by completely unstructured random noise. We corroborate these experimental findings with a theoretical construction showing that simple depth two neural networks already have perfect finite sample expressivity as soon as the number of parameters exceeds the number of data points as it usually does in practice.

We interpret our experimental findings by comparison with traditional models.

We supplement this republication with a new section at the end summarizing recent progresses in the field since the original version of this paper.

underwrites the generalization ability of a model has occupied the machine learning research community for decades.

There are a variety of theories proposed to explain generalization.

Uniform convergence, margin theory, and algorithmic stability are but a few of the important conceptual tools to reason about generalization. Central to much theory are different notions of *model complexity*. Corresponding generalization bounds quantify how much data is needed as a function of a particular complexity measure. Despite much significant theoretical work, the prescriptive and descriptive value of these theories remains debated.

This work takes a step back. We do not offer any new theory of generalization. Rather, we offer a few simple experiments to interrogate the empirical import of different purported theories of generalization. With these experiments at hand, we broadly investigate what practices do and do not promote generalization, what does and does not measure generalization?

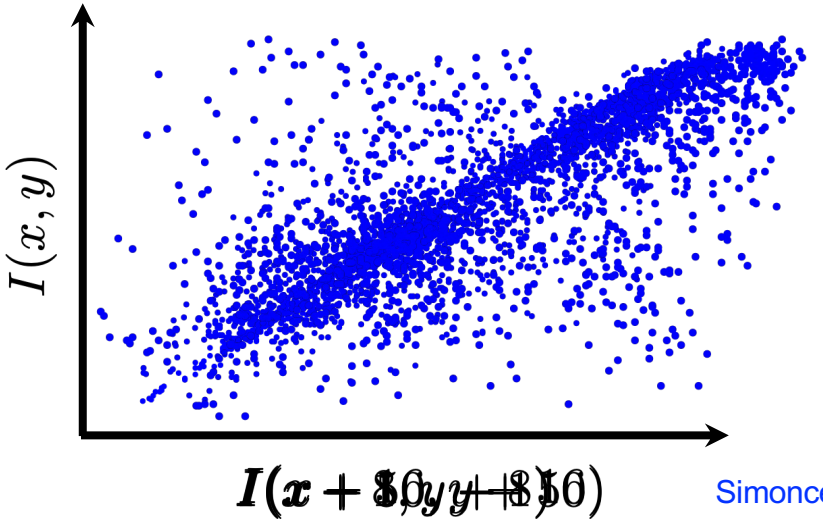
1.1. The randomization test

In our primary experiment, we create a copy of the training data where we replace each label independently by a random label chosen from the set of valid labels. A dog picture labeled "dog" might thus become a dog picture labeled "air-

Today

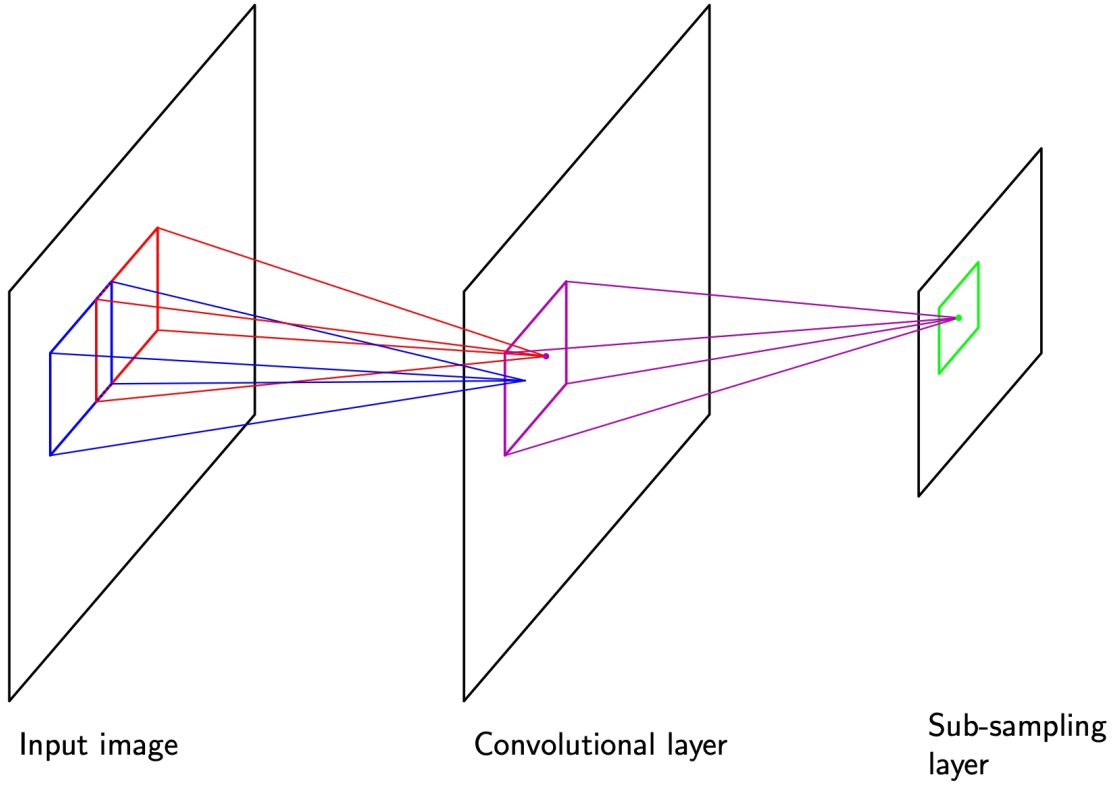
- Shallow Deep Visual Learning
- Deep Visual Learning
- **ConvNets and AlexNet**

I



Simoncelli & Olshausen 2001

Convolutional Neural Network

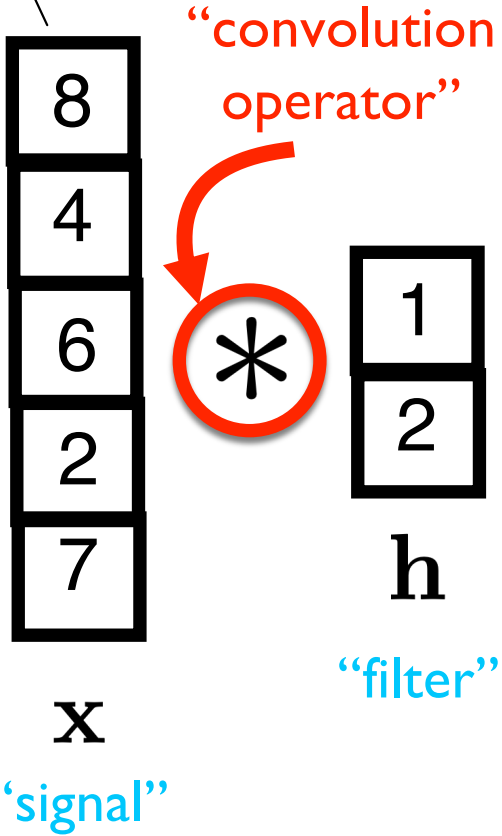


LeCun 1980

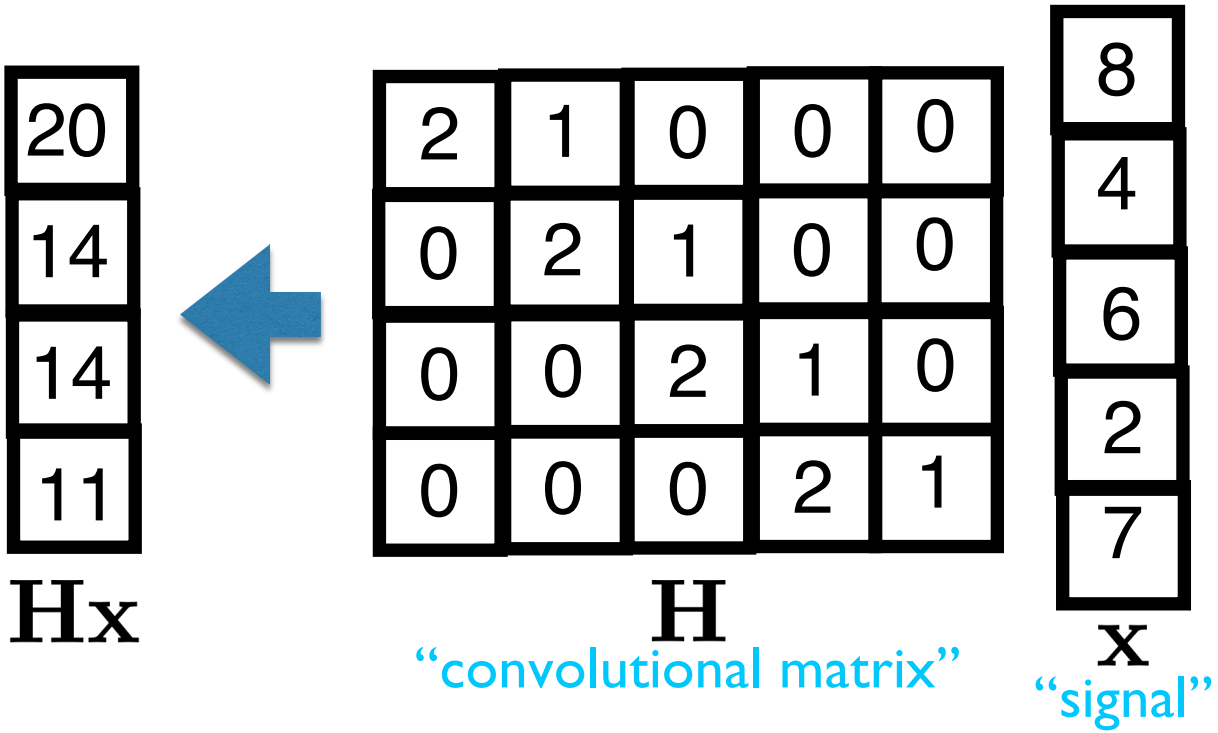
Reminder: Convolution

```
>>> from scipy.signal import \
    convolve as conv
>>> conv(x,h,'valid')
array([20, 14, 14, 11])
```

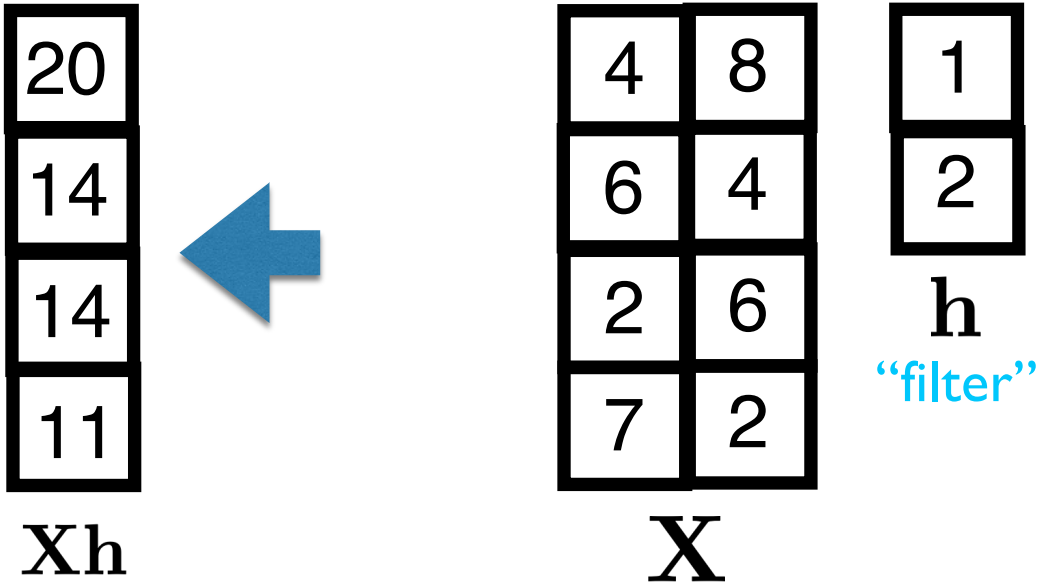
$$\frac{\partial[\mathbf{x} * \mathbf{h}]}{\mathbf{h}^T} = ???$$



Reminder: Convolution



Reminder: Convolution



$$\frac{\partial[\mathbf{x} * \mathbf{h}]}{\mathbf{h}^T} = \frac{\partial[\mathbf{X}\mathbf{h}]}{\mathbf{h}^T} = \mathbf{X}^T$$

Efficiency of Convolution

Input size: 320 by 280
Filter size: 2 by 1
Output size: 319 by 280

	2	319*280*320*280 > 8e9	2*319*280 = 178,640
	319*280*3 = 267,960	> 16e9	Same as convolution (267,960)

Vectorizing 2D-Convolution

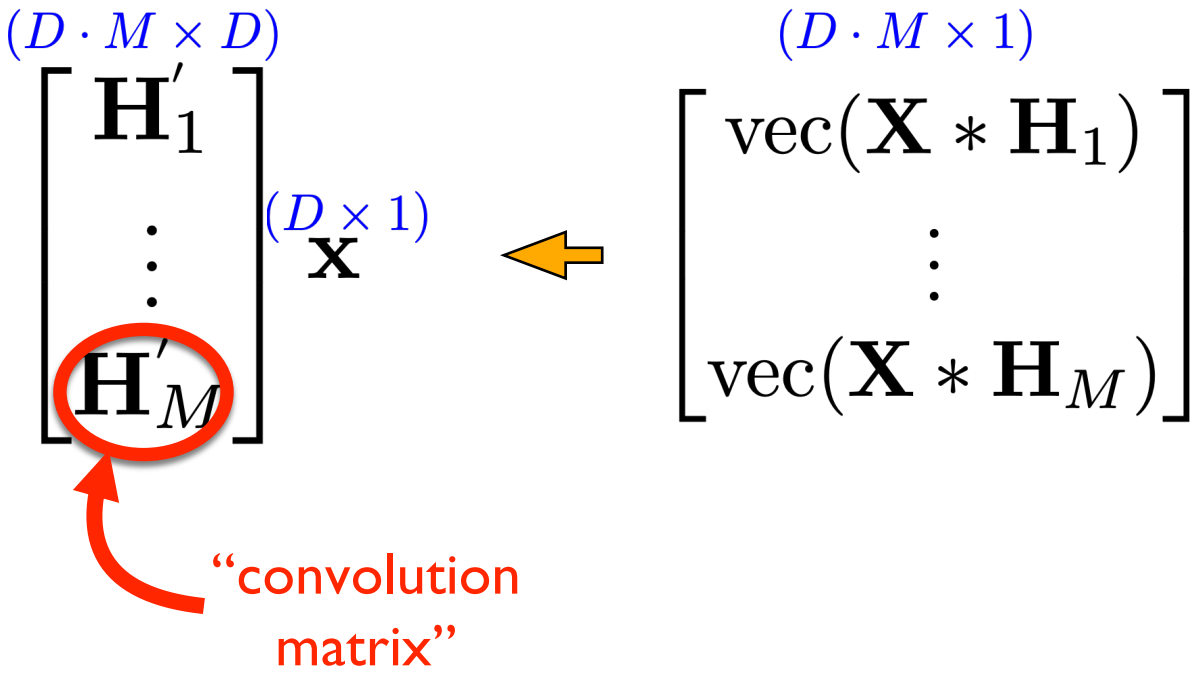
$$\text{vec}(\mathbf{Y}) = \mathbf{X}\,\text{vec}(\mathbf{K} * \mathbf{H})$$

```
def forward(self, x):
    # Flattens the image like structure into vect
    x = torch.flatten(x, start_dim=1)
```

Vectorizing 2D-Convolution

$$\mathrm{vec}(\mathbf{Y}) = \mathbf{H}' \times \mathrm{vec}(\mathbf{X})$$

Multiple Filters



Removing Redundancy - Striding

```
Stride: [4,4]
```

Every 4th element

Every 4th element

207	245	77	21	247	211	240	1
219	41	58	179	161	154	184	98
215	145	187	71	251	249	65	100
192	2	189	247	166	63	232	213
105	94	66	190	156	61	89	145
159	154	87	184	101	105	72	71
192	111	6	94	60	70	65	226
175	120	210	226	80	183	168	184
134	56	36	240	159	178	76	135
239	244	199	9	132	104	188	185
245	210	78	199	0	92	9	246
5	121	187	122	107	47	12	119
230	171	135	36	82	54	65	37
61	140	79	19	161	96	127	187
56	223	46	6	180	186	142	244
28	20	61	2	178	187	98	220

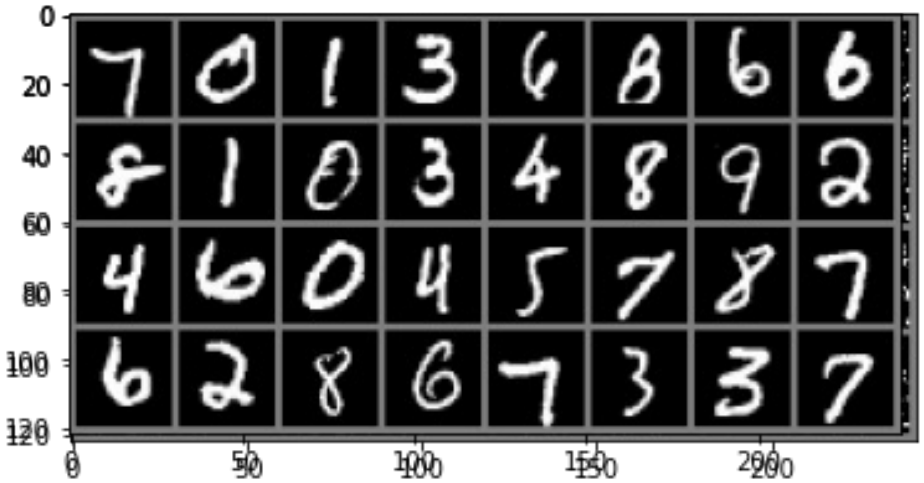
Let’s have another play!!!

https://colab.research.google.com/github/slucey-cs-cmu-edu/RVSS24/blob/main/CNN_Example.ipynb

Some things to try!!!

- What happens to performance if you use a linear function?
- What happens when you permute the pixels?

```
from numpy.random import permutation
idx_permute = torch.from_numpy(permutation(784))
transform = transforms.Compose([transforms.ToTensor(),
                                transforms.Lambda(lambda x: x.view(-1)[idx_permute].view(1, 28, 28) ),
                                transforms.Normalize((0.5,), (0.5,)),
                                ])
```



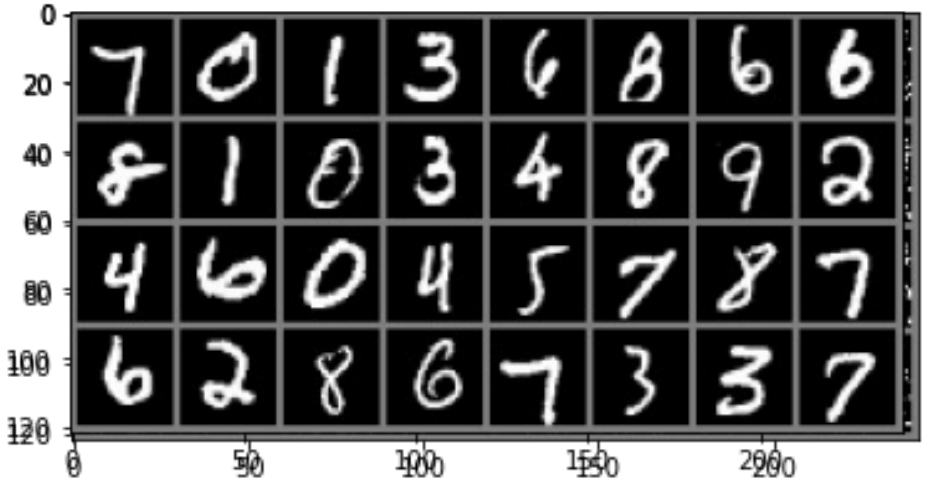
Adding Striding - CNN

```
class ConvNet(nn.Module):
    def __init__(self):
        super(ConvNet, self).__init__()
        self.conv1 = nn.Conv2d(1, 6, 5, 2)
        self.conv2 = nn.Conv2d(6, 16, 5, 2)
        self.fc1 = nn.Linear(256, 84)
        self.fc2 = nn.Linear(84, 10)

    def forward(self, x):
        # Input goes to convolution so no need t
        x = self.conv1(x)
        x = F.relu(x)
        x = self.conv2(x)
        x = F.relu(x)
        x = torch.flatten(x, start_dim=1)
        x = self.fc1(x)
        x = F.relu(x)
        x = self.fc2(x)
```

Permuting Pixels - CNN

- What happens when you permute the pixels in a CNN?

```
from numpy.random import permutation
idx_permute = torch.from_numpy(permutation(784))
transform = transforms.Compose([transforms.ToTensor(),
                                transforms.Lambda(lambda x: x.view(-1)[idx_permute].view(1, 28, 28) ),
                                transforms.Normalize((0.5,), (0.5,)),
                                ])
```



ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

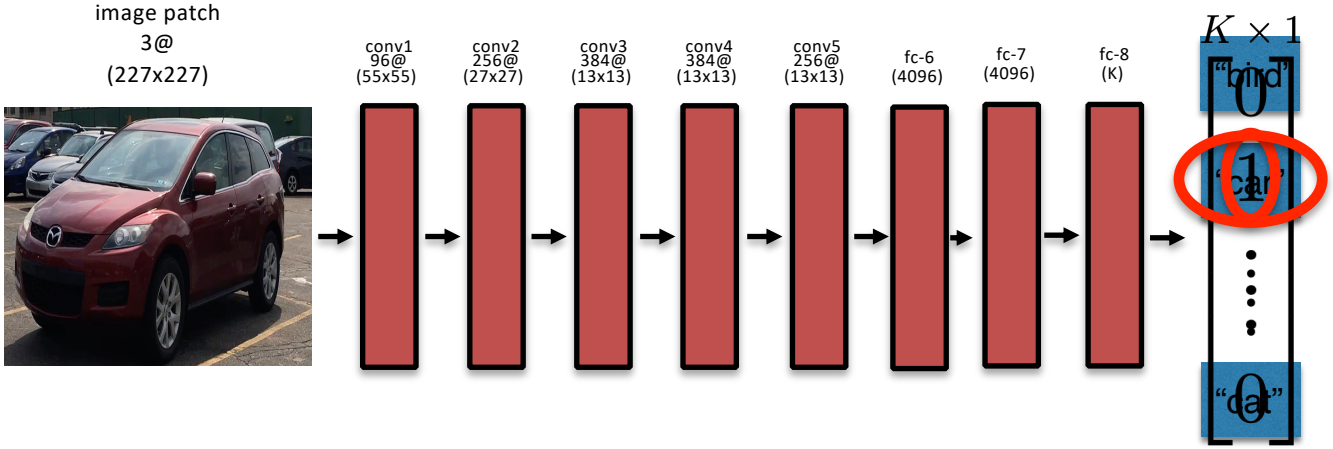
Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

AlexNet

- AlexNet won the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%. (Second best was 26.2 %).
- Network has 25 layers, but only 8 layers with learnable weights.
  - 5 convolutional weights.
  - 3 fully connected weights.



A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. NeurIPS, 2012.

AlexNet in PyTorch

```
>>> from torchvision import models
>>> net = models.alexnet()
```

AlexNet in PyTorch

```
>>> net
AlexNet(
  (features): Sequential(
    (0): Conv2d(3, 64, kernel_size=(11, 11), stride=(4, 4), padding=(2, 2))
    (1): ReLU(inplace)
    (2): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
    (3): Conv2d(64, 192, kernel_size=(5, 5), stride=(1, 1), padding=(2, 2))
    (4): ReLU(inplace)
    (5): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
    (6): Conv2d(192, 384, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (7): ReLU(inplace)
    (8): Conv2d(384, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (9): ReLU(inplace)
    (10): Conv2d(256, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (11): ReLU(inplace)
    (12): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
  )
  (classifier): Sequential(
    (0): Dropout(p=0.5)
    (1): Linear(in_features=9216, out_features=4096, bias=True)
    (2): ReLU(inplace)
    (3): Dropout(p=0.5)
    (4): Linear(in_features=4096, out_features=4096, bias=True)
    (5): ReLU(inplace)
    (6): Linear(in_features=4096, out_features=1000, bias=True)
  )
)
```

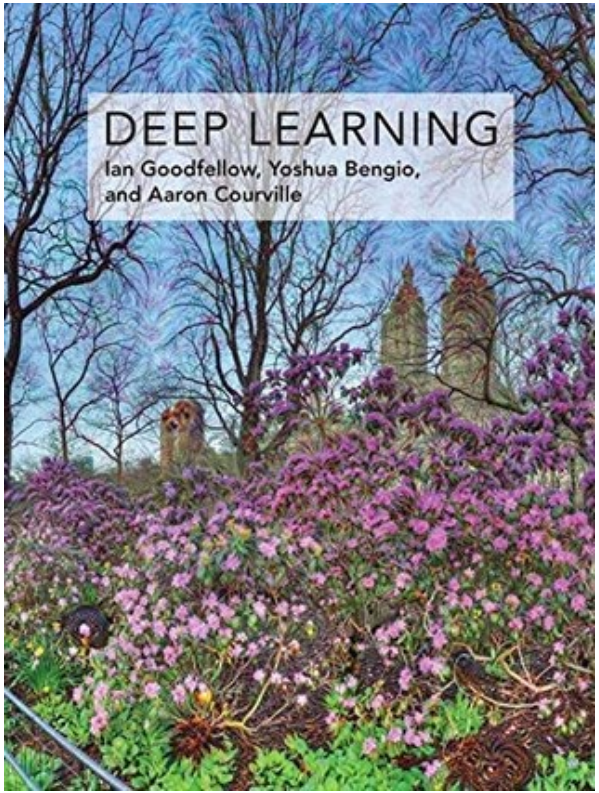
AlexNet in PyTorch

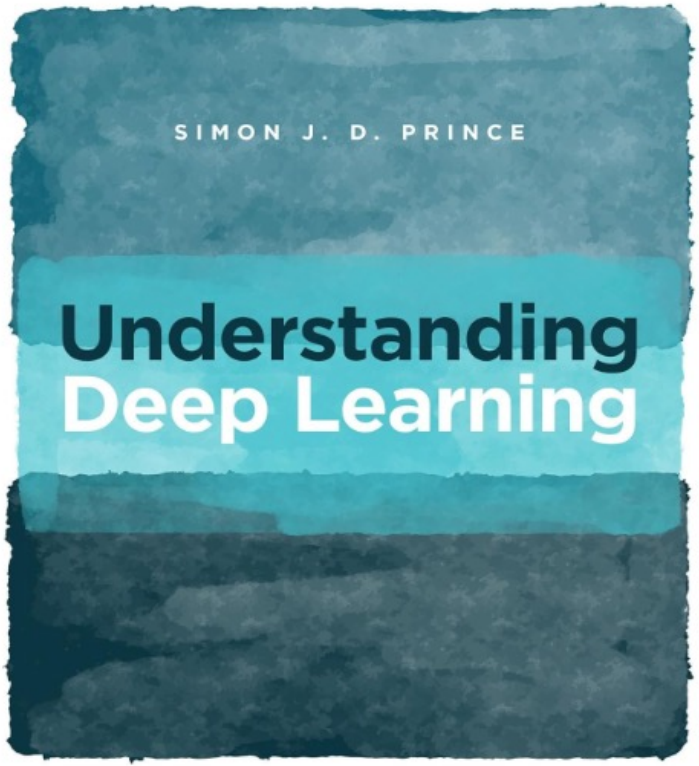
```
>>> from torchvision import models
>>> net = models.alexnet()
```

AlexNet in PyTorch

```
>>> net
AlexNet(
  (features): Sequential(
    (0): Conv2d(3, 64, kernel_size=(11, 11), stride=(4, 4), padding=(2, 2))
    (1): ReLU(inplace)
    (2): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
    (3): Conv2d(64, 192, kernel_size=(5, 5), stride=(1, 1), padding=(2, 2))
    (4): ReLU(inplace)
    (5): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
    (6): Conv2d(192, 384, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (7): ReLU(inplace)
    (8): Conv2d(384, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (9): ReLU(inplace)
    (10): Conv2d(256, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (11): ReLU(inplace)
    (12): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
  )
  (classifier): Sequential(
    (0): Dropout(p=0.5)
    (1): Linear(in_features=9216, out_features=4096, bias=True)
    (2): ReLU(inplace)
    (3): Dropout(p=0.5)
    (4): Linear(in_features=4096, out_features=4096, bias=True)
    (5): ReLU(inplace)
    (6): Linear(in_features=4096, out_features=1000, bias=True)
  )
)
```

More to read…


DEEP LEARNING
Ian Goodfellow, Yoshua Bengio,
and Aaron Courville


SIMON J. D. PRINCE
Understanding
Deep Learning