Lecture Notes for

Machine Learning in Python

Revisiting CV

A "Way too Early" History of Deep Learning

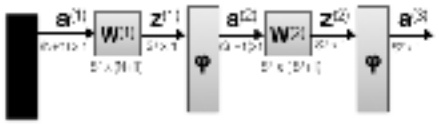
Logistics and Agenda

- Logistics
  - None!
- Agenda
  - Cross Validation
  - Town Hall
  - “Deep Learning” History
  - Remaining Course Topics
    - TensorFlow from 10,000 feet
    - Wide and Deep Networks
    - Convolution Neural Networks
    - Recurrent Neural Networks

Last time:

Back propagation summary

$$J(\mathbf{W}) = \sum_{k}^{M} (\mathbf{y}^{(k)} - \mathbf{a}^{(L)})^2$$



$$w_{ij}^{(l)} \leftarrow w_{ij}^{(l)} - \eta \frac{\partial J(\mathbf{W})}{\partial \mathbf{z}^{(l)}} a_j^{(l)}$$

1. Forward propagate to get $\mathbf{z}$, $\mathbf{a}$ for all layers
2. Get final layer gradient
3. Update back propagation variables
4. Update each $\mathbf{W}^{(l)}$

for each $\mathbf{y}^{(k)}$

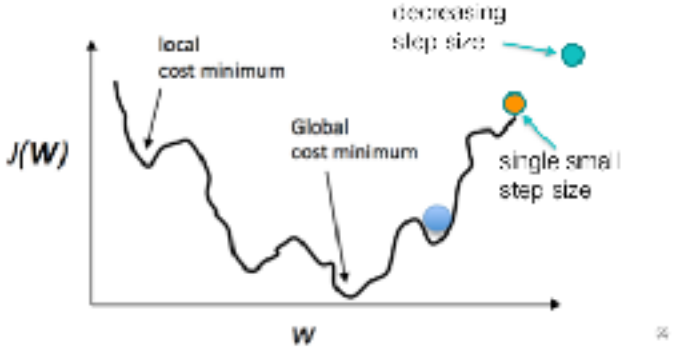
$$\frac{\partial J(\mathbf{W})}{\partial \mathbf{z}^{(2)}} = -2(\mathbf{y}^{(k)} - \mathbf{a}^{(3)}) * \mathbf{a}^{(3)} * (1 - \mathbf{a}^{(3)})$$

$$\frac{\partial J(\mathbf{W})}{\partial \mathbf{z}^{(l)}} = \text{diag}[\mathbf{a}^{(l+1)} * (1 - \mathbf{a}^{(l+1)})] \cdot \mathbf{W}^{(l+1)} \frac{\partial J(\mathbf{W})}{\partial \mathbf{z}^{(l+1)}}$$

$$\mathbf{W}^{(l)} \leftarrow \mathbf{W}^{(l)} - \eta \frac{\partial J(\mathbf{W}^{(l)})}{\partial \mathbf{z}^{(l)}} \cdot \mathbf{a}^{(l)}$$

Problems with Advanced Architectures

- Space is no longer convex
  - **One solution:**
    - start with large step size
    - "cool down" by decreasing step size for higher iterations



Practical Implementation of Architectures

- A new cost function: **Cross entropy**

$$J(\mathbf{W}) = -[\mathbf{y}^{(i)} \ln \mathbf{a}^{(L)} + (1 - \mathbf{y}^{(i)}) \ln(1 - \mathbf{a}^{(L)})]$$

speeds up initial training

$$\frac{\partial J(\mathbf{W})}{\mathbf{z}^{(L)}} = (_i\mathbf{a}^{(L+1)} - \mathbf{y}^{(i)})$$

$$\frac{\partial J(\mathbf{W})}{\mathbf{z}^{(2)}} = (_i\mathbf{a}^{(3)} - \mathbf{y}^{(i)})$$

```
# vectorized backpropagation
sigma3 = (A3-Y_enc) # <- this is only line
sigma2 = (W2.T @ sigma3)*A2*(1-A2)

grad1 = sigma2[1:,:] @ A1
grad2 = sigma3 @ A2.T
```

new update

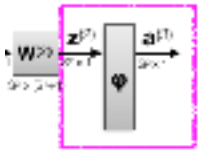
```
# vectorized backpropagation
sigma3 = -2*(Y_enc-A3)*A3*(1-A3)
sigma2 = (W2.T @ sigma3)*A2*(1-A2)

grad1 = sigma2[1:,:] @ A1
grad2 = sigma3 @ A2.T
```

$$\frac{\partial J(\mathbf{W})}{\partial \mathbf{z}^{(2)}} = -2(\mathbf{y}^{(i)} - \mathbf{a}^{(3)}) * \mathbf{a}^{(3)} * (1 - \mathbf{a}^{(3)})$$ old update

bp-5

Practical Implementation of Architectures

- A new nonlinearity: **recitifed linear units**

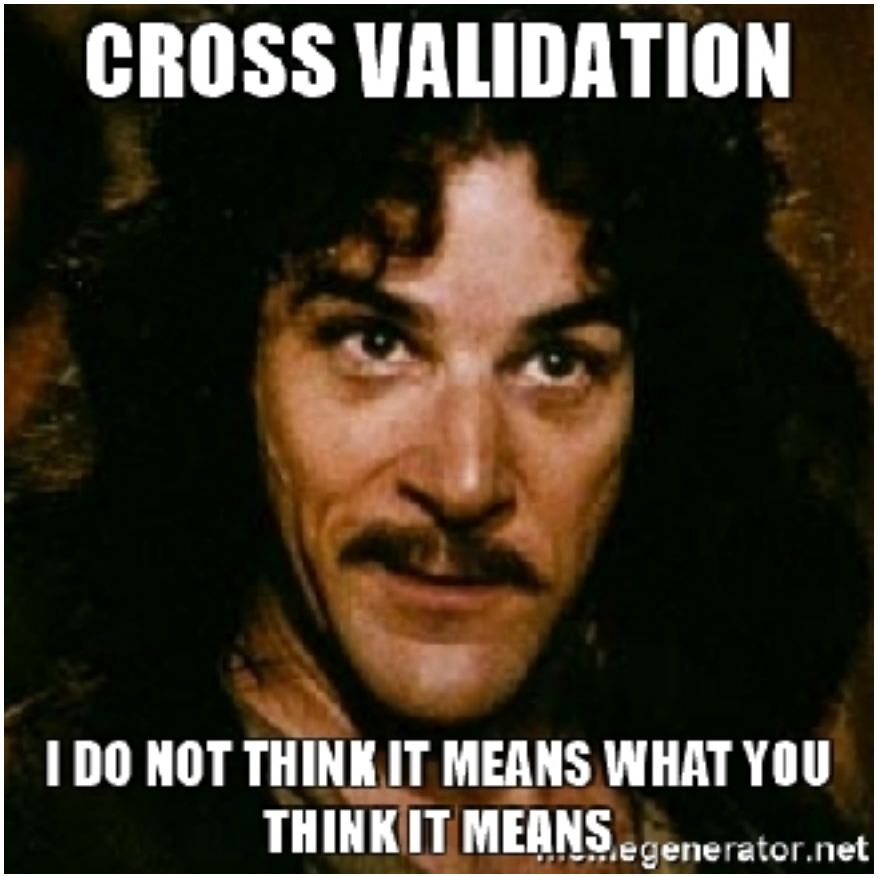


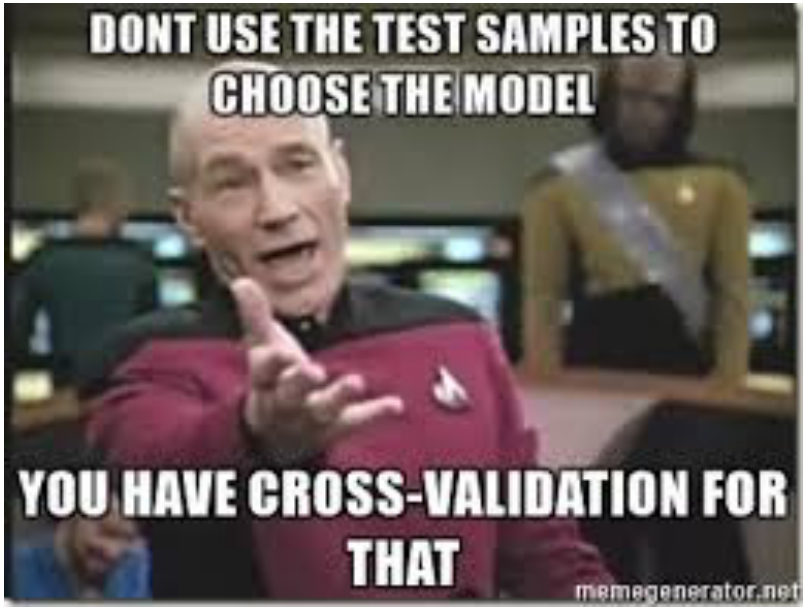
$$\phi(\mathbf{z}^{(i)}) = \begin{cases} \mathbf{z}^{(i)}, \text{ if } \mathbf{z}^{(i)} > 0 \\ 0, \text{ else} \end{cases}$$

it has the advantage of **large gradients** and **extremely simple** derivative

$$\frac{\partial \phi(\mathbf{z}^{(i)})}{\partial \mathbf{z}^{(i)}} = \begin{cases} 1, \text{ if } \mathbf{z}^{(i)} > 0 \\ 0, \text{ else} \end{cases}$$

Revisiting Cross Validation

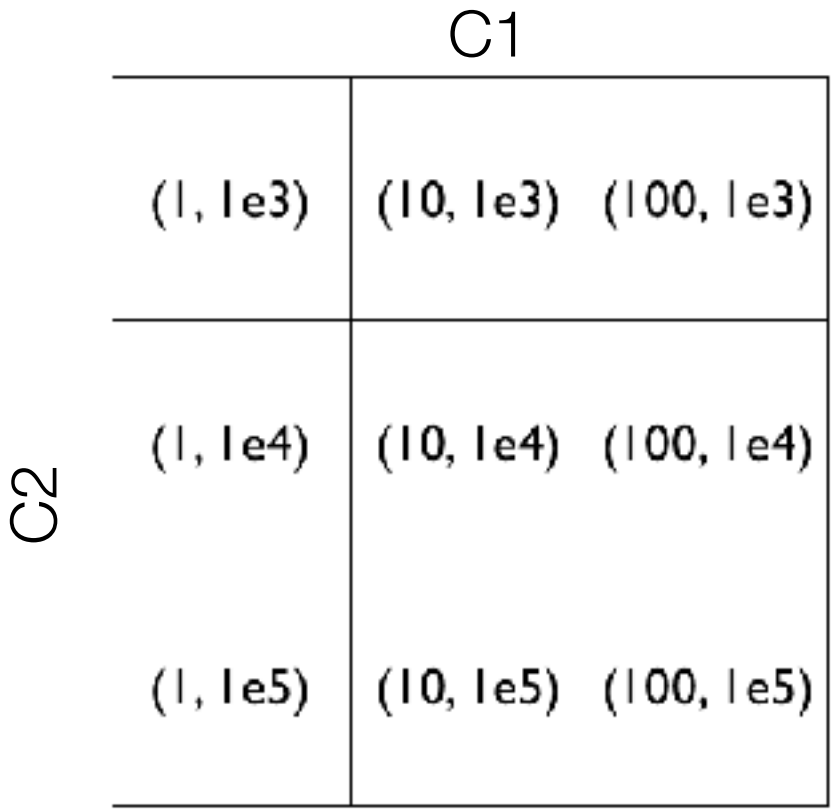




Grid Searching Review

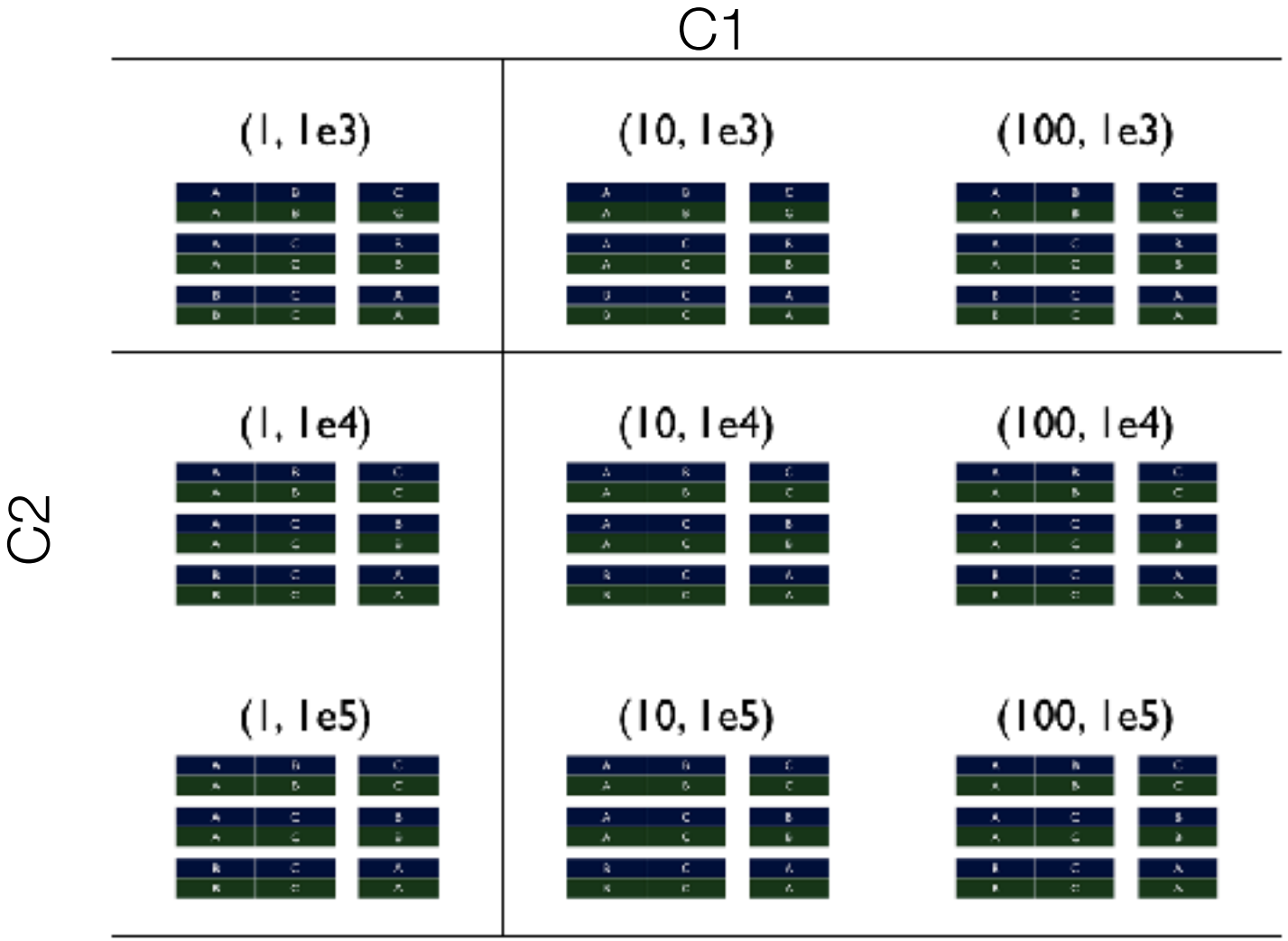
Trying to find the best parameters

```
LR: C1=[1, 10, 100] C2=[1e3,1e4,1e5]
```



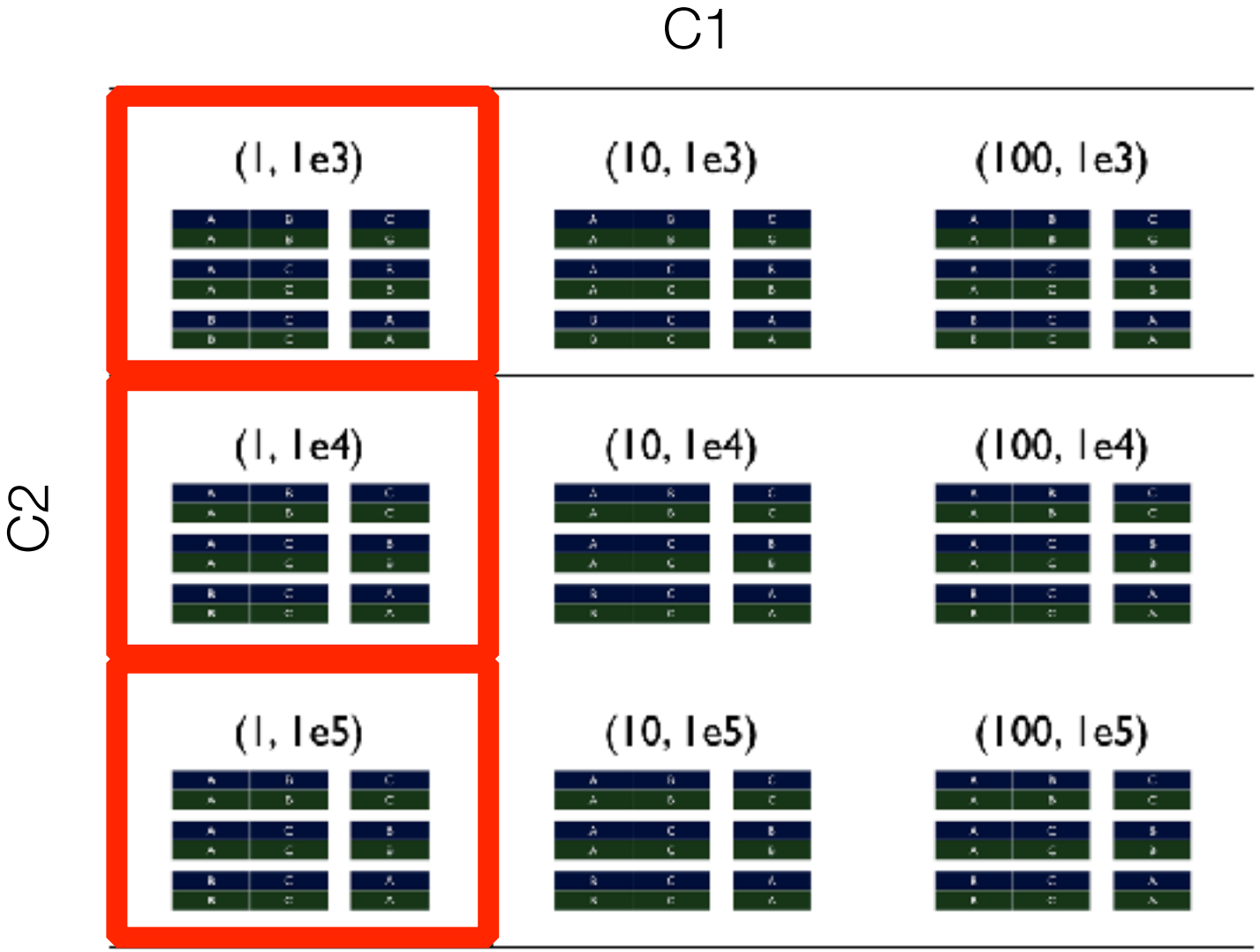
Grid Searching Review

For each value, want to run cross validation…



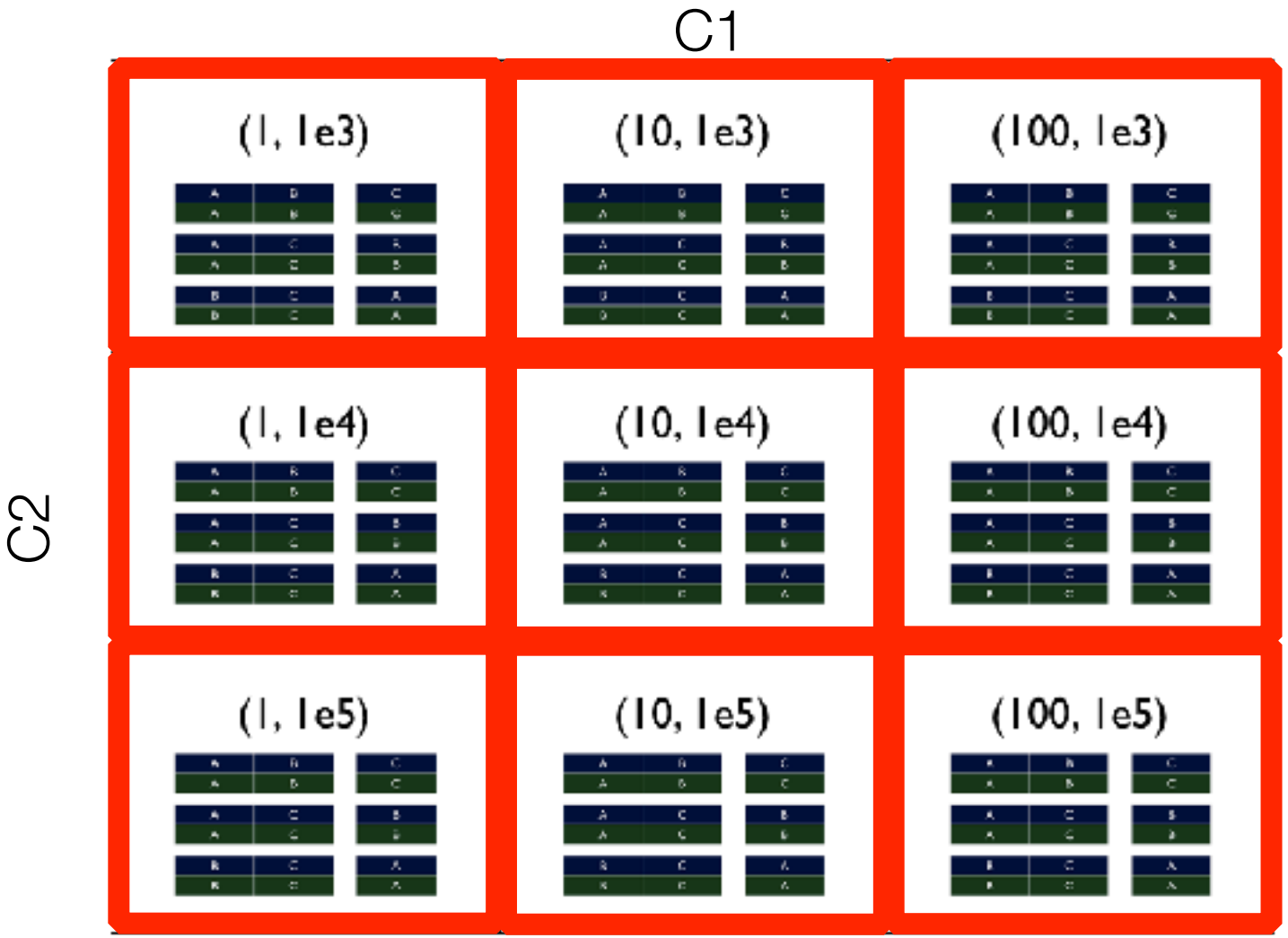
Grid Searching Review

Could perform iteratively

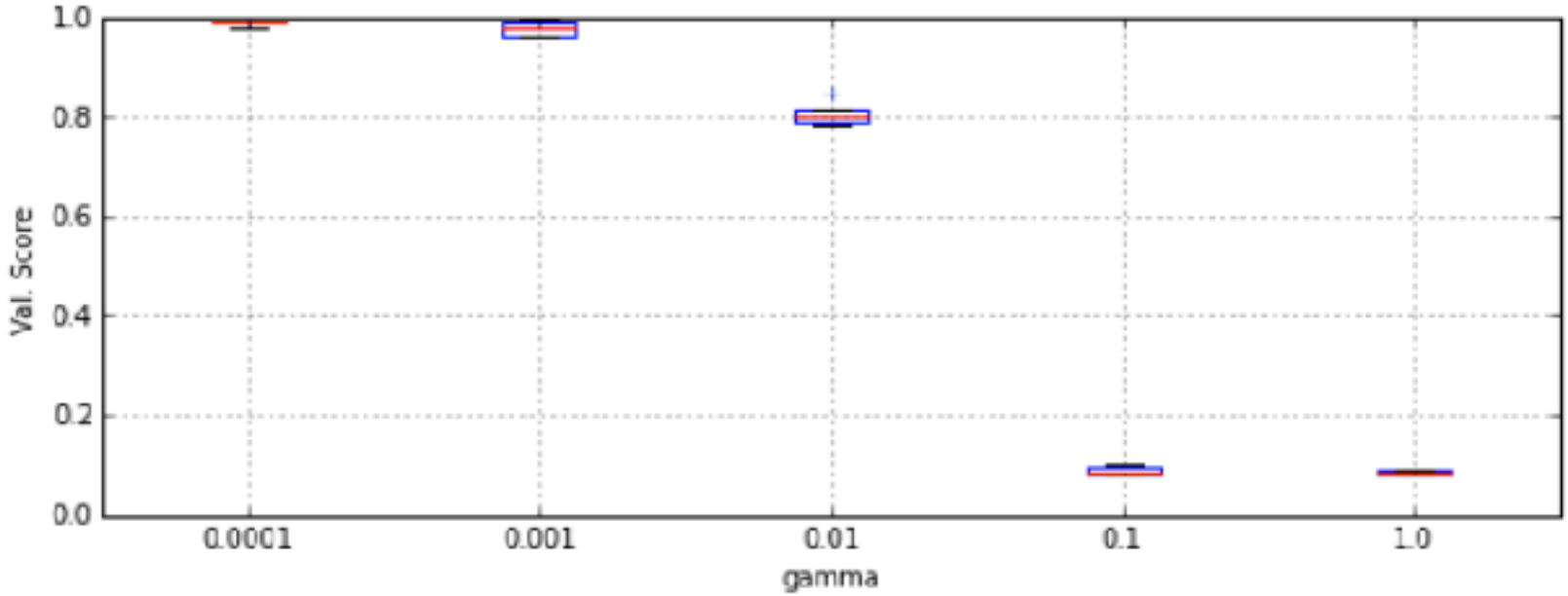


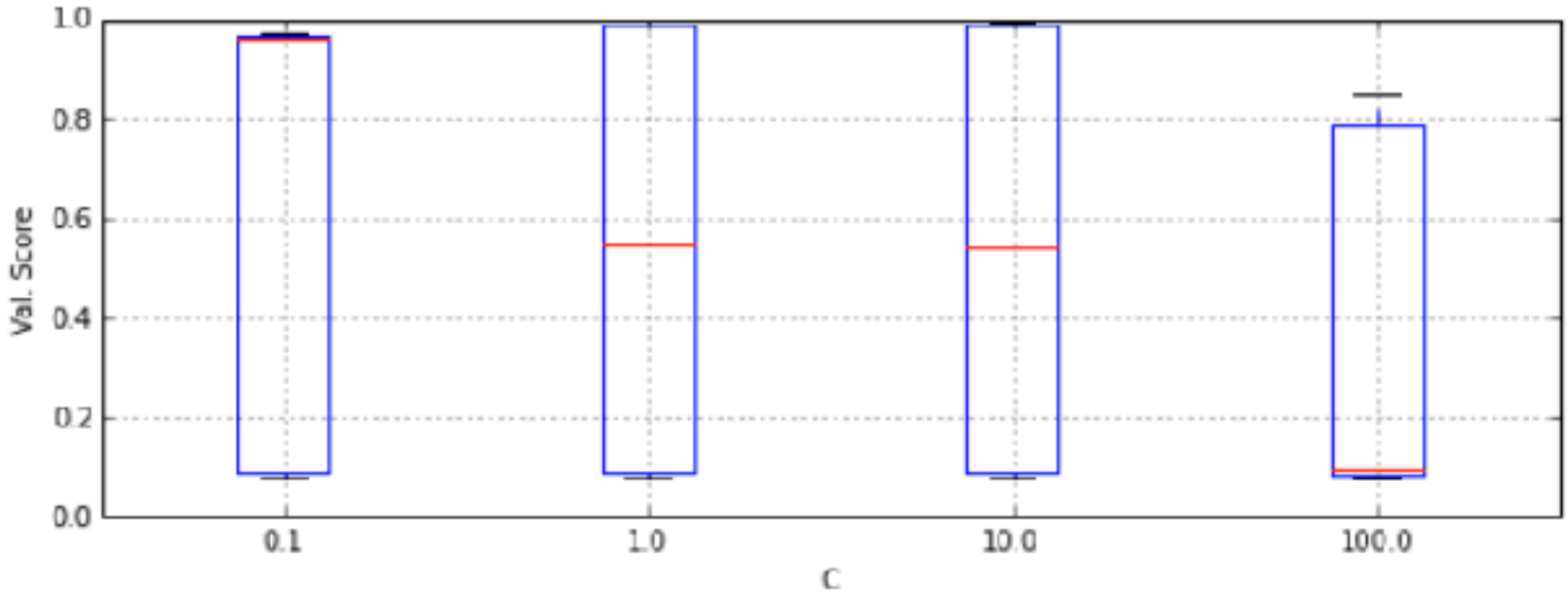
Grid Searching Review

or at random…



```
print(search.report())
search.boxplot_parameters(display_train=False)
```

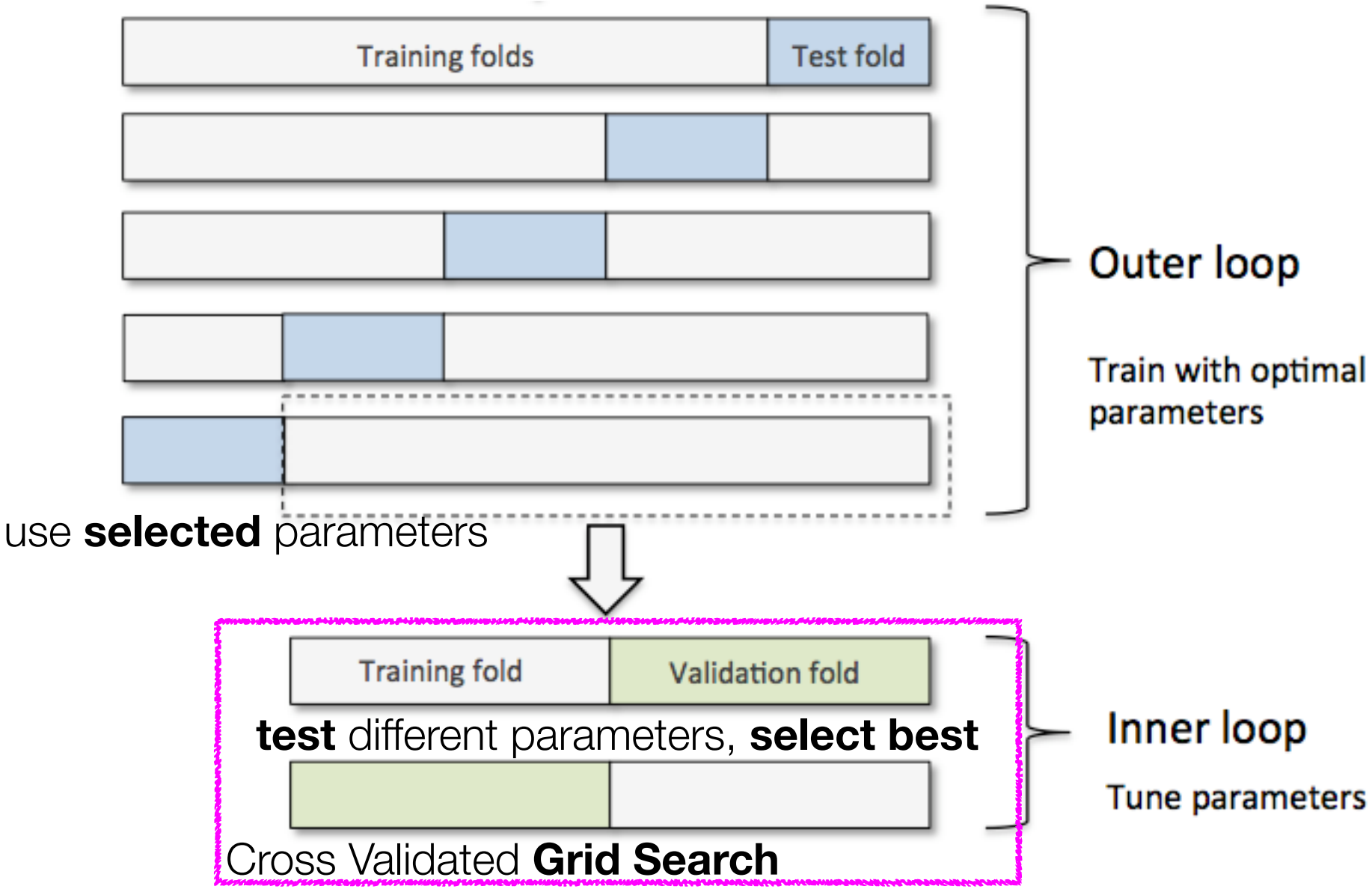




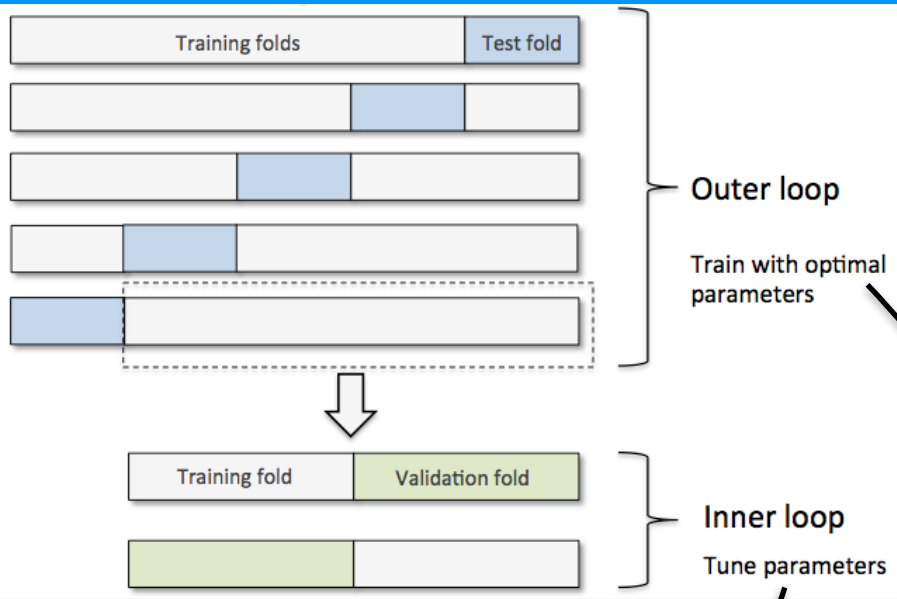
The Problem

- Using the grid search parameters and testing on the same set…
  - the **performance on the dataset** could now be **biased**
  - cannot determine the **expected performance** on **new data**
  - this is **data snooping**

Solution: Nested Cross Validation



Nested Cross Validation: Hyper-parameters




```
gs = GridSearchCV(estimator=pipe_svc,
                  param_grid=param_grid,
                  scoring='accuracy',
                  cv=2)

# Note: Optionally, you could use cv=2
# in the GridSearchCV above to produce
# the 5 x 2 nested CV that is shown in the figure.

scores = cross_val_score(gs, X_train, y_train, scoring='accuracy', cv=5)
print('CV accuracy: %.3f +/- %.3f' % (np.mean(scores), np.std(scores)))
```

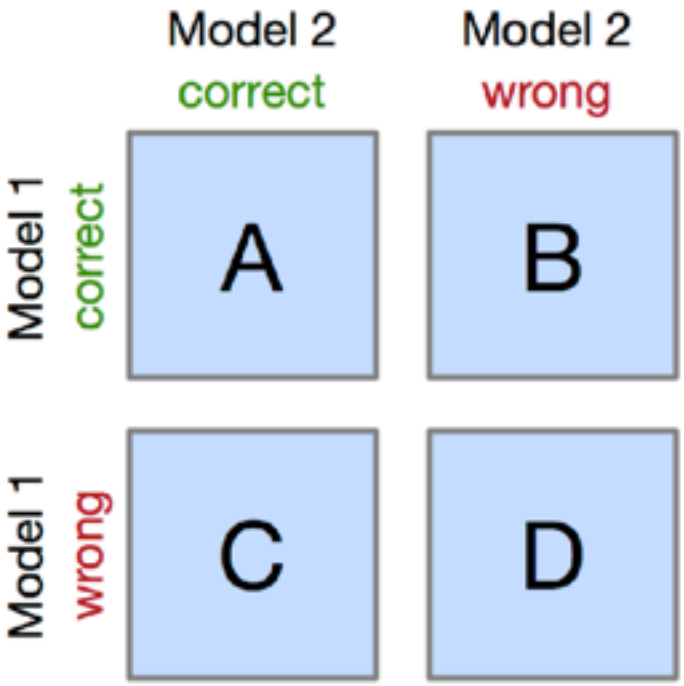
https://github.com/rasbt/python-machine-learning-book/

Self Test

- **What is the end goal of nested cross-validation?**
  - A. To determine hyper parameters
  - B. To estimate generalization performance
  - C. To estimate generalization performance when performing hyper parameter tuning
  - D. To estimate the variation in tuned hyper parameters

McNemar Testing

Few assumptions, **Null hypothesis**: these models are the same!



One caveat: Statistical power depends upon B+C, which might be small, even with lots of test data.

McNemar and Edwards, 1948

$$\chi^2 = \frac{\left(|B - C| - 1\right)^2}{B + C}$$

χ squared statistic, one DOF

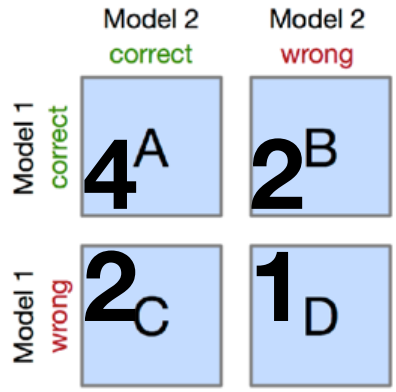
Steps:

1. Compare each model's performance on the same test data (2x2 matrix)
2. Calculate χ statistic
3. Look up *critical value* associated with χ statistic for given confidence
4. Can you reject the null hypothesis that the models are the same (p<0.05)?

https://sebastianraschka.com/blog/2018/model-evaluation-selection-part4.html

McNemar Example

Model 1	Model 2	Label	Matrix
T-shirt	T-shirt	T-shirt	A
Sneaker	T-shirt	Sneaker	B
T-shirt	Pullover	Pullover	C
Sneaker	Sneaker	Sneaker	A
T-shirt	Sneaker	Sneaker	C
Pullover	Pullover	T-shirt	D
Pullover	T-shirt	Pullover	B
Sneaker	Sneaker	Sneaker	A
Sneaker	Sneaker	Sneaker	A



McNemar and Edwards, 1948

$$\chi^2 = \frac{\left(|B - C| - 1\right)^2}{B + C}$$

$$\chi^2 = \frac{(|2 - 2| - 1)^2}{2 + 2} = 0.25$$

Confidence	0.90	0.95	0.99
1 DOF, Critical Value	2.706	3.841	6.635

https://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm

Since 0.25 < 3.841, we cannot reject the null hypothesis. This means **we cannot say the models are different** based on the evidence.

Some History of Deep Learning

Neural Networks: Where we left it

- Before 1986: AI Winter
- 1986: *Rumelhart*, *Hinton*, and *Williams* popularize gradient calculation for multi-layer network
  - *technically* introduced by Werbos in 1982
- **difference**: Rumelhart *et al.* validated ideas with a computer
- until this point no one could train a multiple layer network consistently
- algorithm is popularly called **Back-Propagation**
- wins pattern recognition prize in 1993, becomes de-facto machine learning algorithm in the 90's

David Rumelhart

1942-2011

Geoffrey Hinton

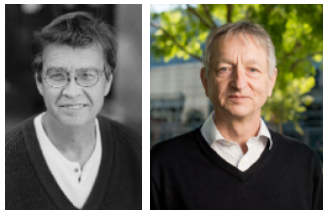
Machine Learning Timeline (Neural Nets)

- Up to this point: back propagation saved AI winter
- 80's, 90's, 2000's: neural networks for image processing start to get deeper
  - but back propagation no longer efficient for training
  - Back propagation gradient **stagnates** research—can't train **deeper** networks

- Second AI winter begins, research in NN plummets
- Funding for and accepted papers with Neural Networks asymptotically approaches zero

1949, Hebb's Law
Close neuron fire together

1960, Widrow & Hoff
Adaline Network



1986, Rumelhart & Hinton
Back-propagation

2003, Vapnik
Kernel SVMs

1940 — 1960 — 1980 — 2000 — 2020

Period of Discovery | First AI Winter | Golden Age of NN | 2nd AI Winter | Age of Deep Learning

1943, McCulloch & Pitts
Logic Gates of The Mind

1957, Rosenblatt
Perceptron

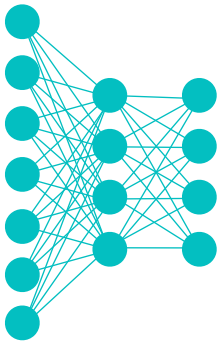
1969, Minsky & Papert
Linear Models are Doomed

2001, Breiman
Random Forests

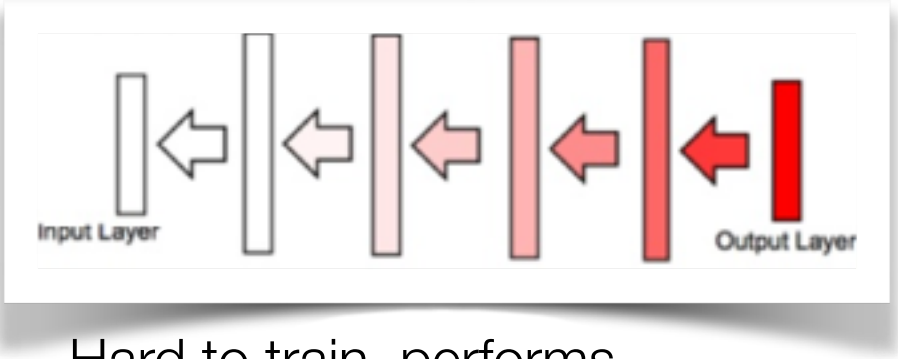
Read this: http://www.andreykurenkov.com/writing/a-brief-history-of-neural-nets-and-deep-learning/

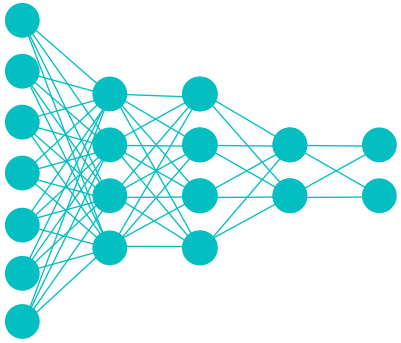
History of Deep Learning: Winter

- AI Winter is coming:



Easy to train, performs on par with other methods





Hard to train, performs worse than other methods

Researcher have difficulty reconciling expressiveness with performance

~chance (untrainable)

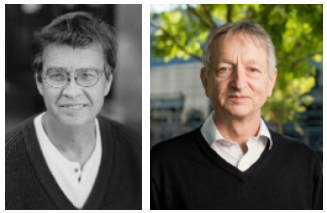
Read this: http://www.andreykurenkov.com/writing/a-brief-history-of-neural-nets-and-deep-learning/

Machine Learning Timeline (Neural Nets)

- 2004: Hinton secures funding from CIFAR based on his reputation
  - *eventually*: Canada would be savior for NN
  - Hinton rebrands: **Deep Learning**
- 2006: Hinton publishes paper on using pre-training and Restricted Boltzmann Machines
- 2007: Another paper: Deep networks are more efficient when pre-trained
  - RBMs not really the important part

1949, Hebb's Law
Close neuron fire together

1960, Widrow & Hoff
Adaline Network



1986, Rumelhart & Hinton
Back-propagation

2003, Vapnik
Kernel SVMs

1940 | 1960 | 1980 | 2000 | 2020

Period of Discovery | First AI Winter | Golden Age of NN | 2nd AI Winter | Age of Deep Learning

1943, McCulloch & Pitts
Logic Gates of The Mind

1957, Rosenblatt
Perceptron

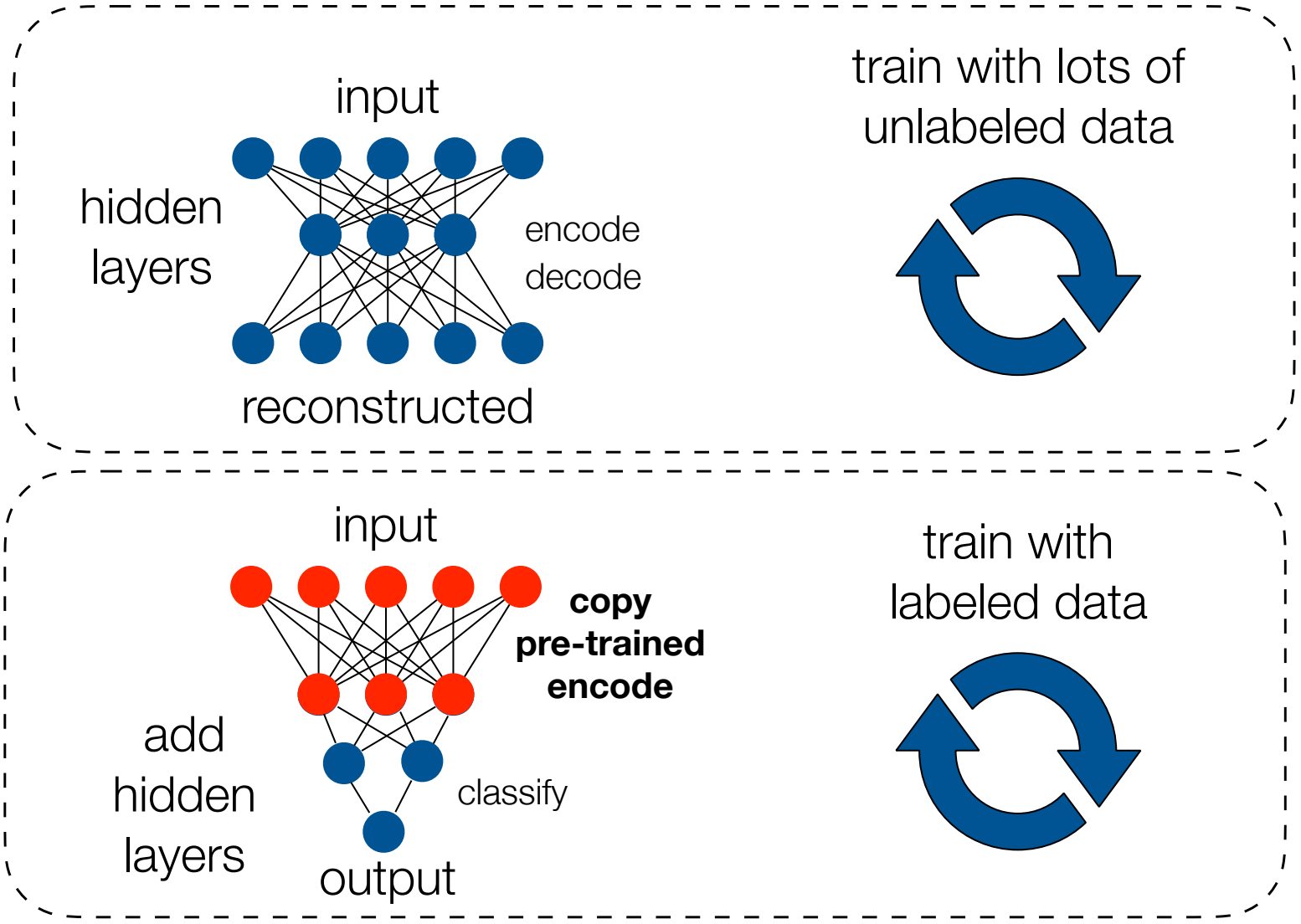
1969, Minsky & Papert
Linear Models are Doomed

2001, Breiman
Random Forests

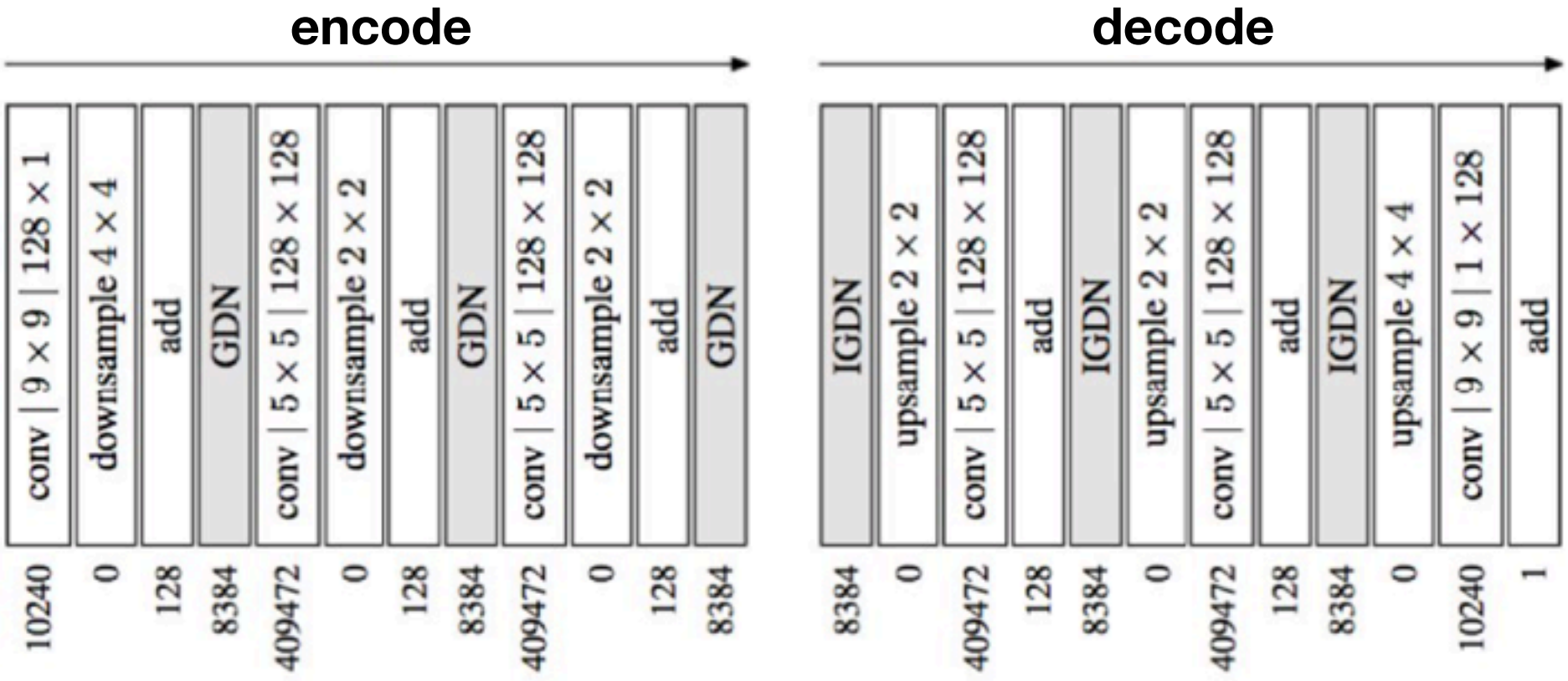
Read this: http://www.andreykurenkov.com/writing/a-brief-history-of-neural-nets-and-deep-learning/

Pre-training: still in the long winter

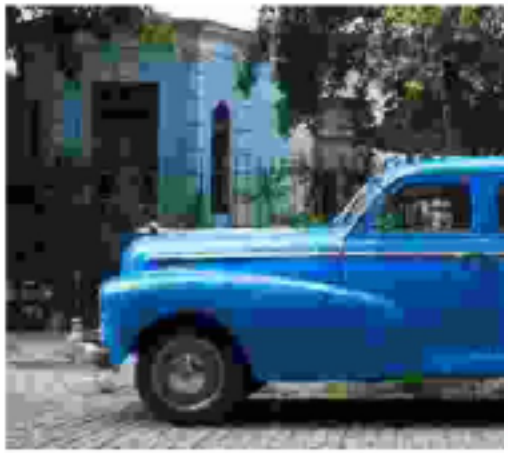
- auto-encoding (a form of pre-training)



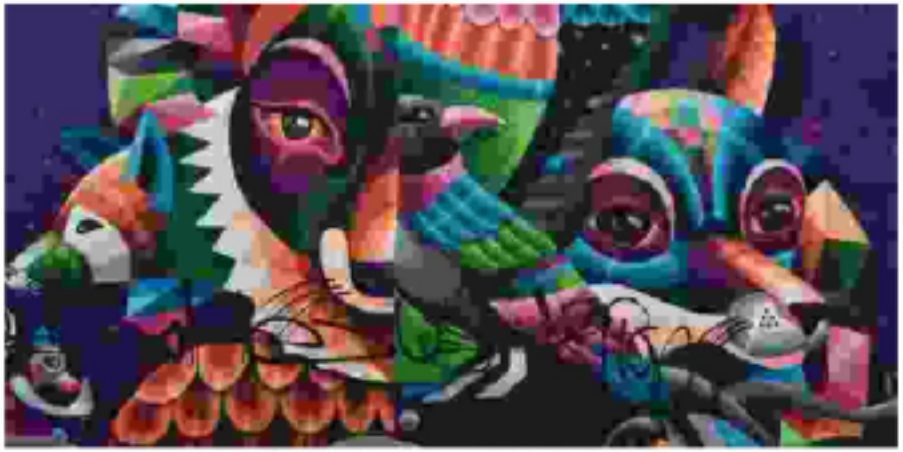
Pre-training: modern example



https://arxiv.org/abs/1611.01704



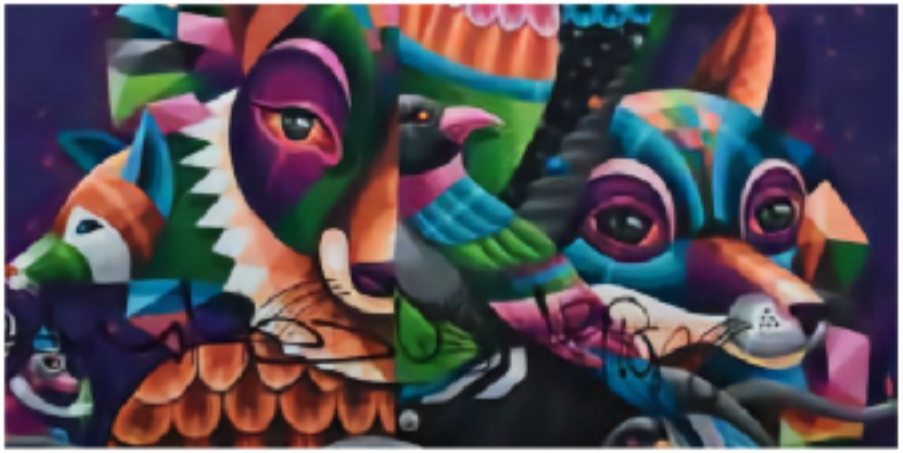
JPEG, 6006 bytes (0.170 bit/px), RMSE: 19.75



JPEG, 5928 bytes (0.168 bit/px), RMSE: 15.44/12.40, PSNR: 24.36 dB/26.26 dB

RMSE: 11.07/10.60, PSNR: 27.25 dB/27.63 dB

Proposed method, 5910 bytes (0.167 bit/px), RMSE



Proposed method, 5685 bytes (0.161 bit/px), RMSE: 10.41/5.98, PSNR: 27.78 dB/32.60 dB

it/px), RMSE: 6.10/5.09, PSNR: 32.43 dB/34.00 dB

JPEG 2000, 5918 bytes (0.167 bit/px), RMSE: 1

JPEG 2000, 5724 bytes (0.162 bit/px), RMSE: 13.75/7.02, PSNR: 25.36 dB/31.20 dB

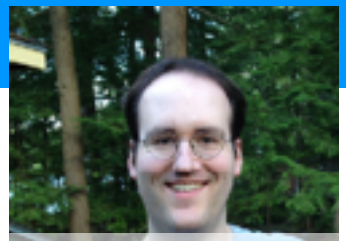
x), RMSE: 8.56/5.71, PSNR: 29.49 dB/32.99 dB

Still in the Long Winter

- 2009: Hinton's lab starts using GPUs, Also Andrew Ng
  - GPUs decrease training time by 70 fold…
- 2010: Hinton's and Ng's students go to internships with Microsoft, Google, IBM, and Facebook

Research at Google

Navdeep Jaitly
Google Brain Team

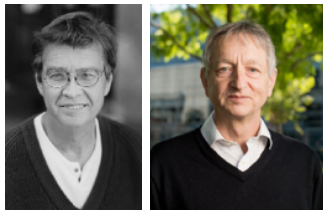


George Dahl
Google Brain Team

Abdel-rahman Mohamed
Microsoft Research
Redmond, Washington | Computer Software
Current Microsoft
Previous University of Toronto, IBM, Microsoft
Education University of Toronto

1949, Hebb's Law
Close neuron fire together

1960, Widrow & Hoff
Adaline Network



1986, Rumelhart & Hinton
Back-propagation

2003, Vapnik
Kernel SVMs

- Xbox Voice
- Android Speech Recognition
- IBM Watson
- DeepFace
- All of Baidu

1940 | 1960 | 1980 | 2000

Period of Discovery | First AI Winter | Golden Age of NN | 2nd AI Winter

1943, McCulloch & Pitts
Logic Gates of The Mind

1957, Rosenblatt
Perceptron

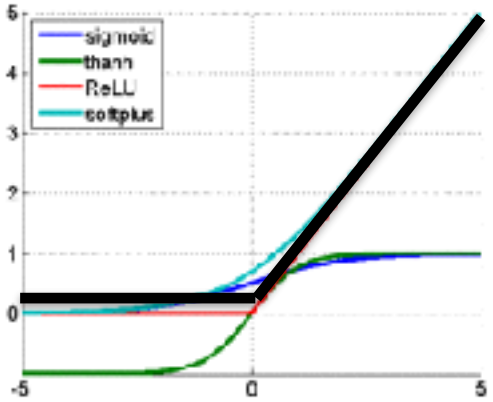
1969, Minsky & Papert
Linear Models are Doomed

2001, Breiman
Random Forests

Read this: http://www.andreykurenkov.com/writing/a-brief-history-of-neural-nets-and-deep-learning/

Still in the long Winter

- 2011: Glorot and Bengio investigate more systematic methods for why past deep architectures did not work
  - **discover some interesting, simple fixes**: the type of neurons chosen and the selection of initial weights
  - do not require pre-training to get deep networks properly trained, just sparser representations and less complicated derivatives



ReLU: $f(x) = \max(0,x)$
$f'(x) = 1$ if $x>0$ else 0

1949, Hebb's Law
Close neuron fire together


1960, Widrow & Hoff
Adaline Network


1986, Rumelhart & Hinton
Back-propagation

2003, Vapnik
Kernel SVMs

1940 — 1960 — 1980 — 2000 — 2020

Period of Discovery | First AI Winter | Golden Age of NN | 2nd AI Winter | Age of Deep Learning

1943, McCulloch & Pitts
Logic Gates of The Mind

1957, Rosenblatt
Perceptron

1969, Minsky & Papert
Linear Models are Doomed

2001, Breiman
Random Forests

2011, Bengio
Init and ReLU

Read this: http://www.andreykurenkov.com/writing/a-brief-history-of-neural-nets-and-deep-learning/

Still in the long Winter

- ReLU not the only way to do it!
  - Sigmoid Weighted Linear Unit
    SiLU
    - also called Swish
  - Mixing of sigmoid, σ, and ReLU

Ramachandran P, Zoph B, Le QV. Swish: a Self-Gated Activation Function. arXiv preprint arXiv:1710.05941. 2017 Oct 16

Elfwing, Stefan, Eiji Uchibe, and Kenji Doya. "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning." Neural Networks (2018).

$$\frac{\partial \varphi(z)}{\partial z} = \varphi(z)+\sigma(z)[1-\varphi(z)]$$

$$= a^{(l+1)} + \sigma(z^{(l)})\cdot[1-a^{(l+1)}]$$

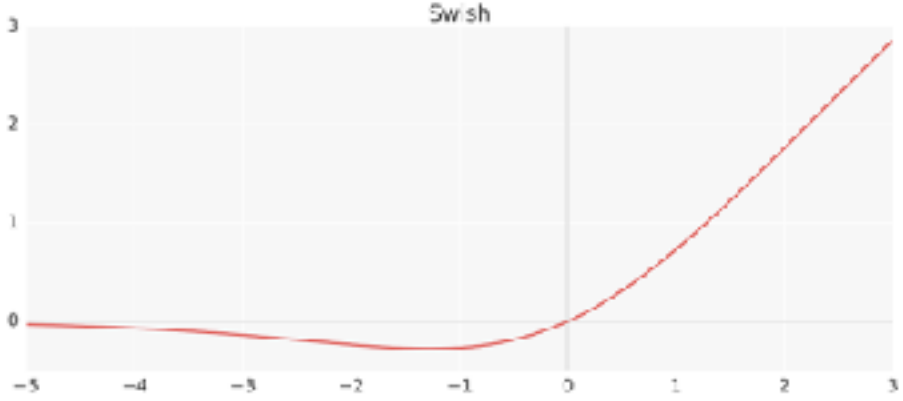


Figure 1: The Swish activation function.

Derivative Calculation:

$$= \sigma(x) + x \cdot \sigma(x)(1 - \sigma(x))$$
$$= \sigma(x) + x \cdot \sigma(x) - x \cdot \sigma(x)^2$$
$$= x \cdot \sigma(x) + \sigma(x)(1 - x \cdot \sigma(x))$$

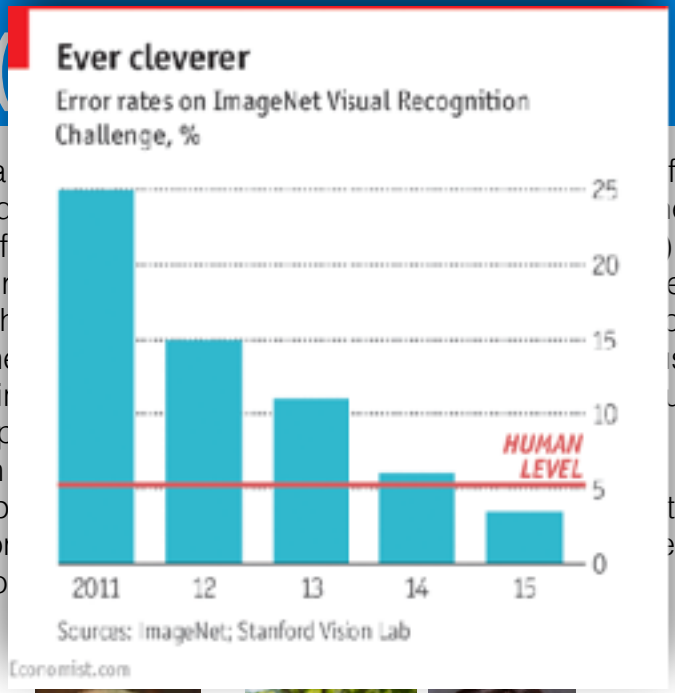
Machine Learning Timeline (

- **ImageNet competition occurs**
- **Second place**: 26.2% error rate
- **First place**:
  - From Hinton's lab, uses convolutional network with ReLU and dropout
  - 15.2% error rate
- Computer vision adopts deep learning with convolutional neural networks en mass

"I ha
hanc
last f
so I
as sh
come
pacin
happ
from
skep
a cor
Visio

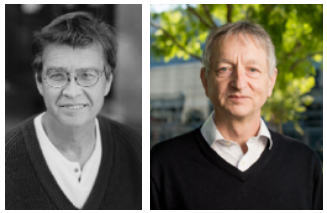
first
e
),
ef
on
s
ust

t as
er



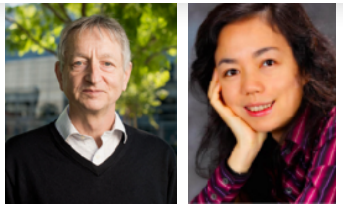
1949, Hebb's Law
Close neuron fire together

1960, Widrow & Hoff
Adaline Network



1986, Rumelhart & Hinton
Back-propagation

2003, Vapnik
Kernel SVMs



2012, Hinton, Fei-Fei Li
CNNs win ImageNet

1940 — 1960 — 1980 — 2000 — 2020

Period of Discovery | First AI Winter | Golden Age of NN | 2nd AI Winter | Age of Deep Learning

1943, McCulloch & Pitts
Logic Gates of The Mind

1957, Rosenblatt
Perceptron

1969, Minsky & Papert
Linear Models are Doomed

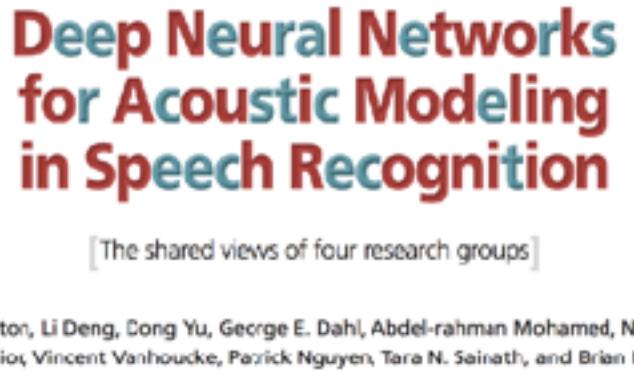
2001, Breiman
Random Forests

2011, Bengio
Init and ReLU

Read this: http://www.andreykurenkov.com/writing/a-brief-history-of-neural-nets-and-deep-learning/

Machine Learning Timeline (Neural Nets)

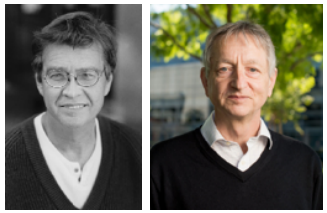
- 2012: Hinton Lab, Google, IBM, and Microsoft jointly publish paper, popularity for deep learning methods increases



https://www.cs.toronto.edu/~gdahl/papers/deepSpeechReviewSPM2012.pdf

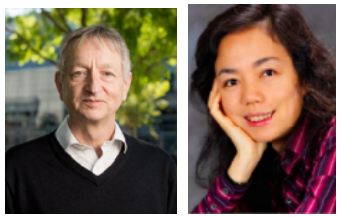
1949, Hebb's Law
Close neuron fire together

1960, Widrow & Hoff
Adaline Network



1986, Rumelhart & Hinton
Back-propagation

2003, Vapnik
Kernel SVMs



2012, Hinton, Fei-Fei Li
CNNs win ImageNet

1940 | 1960 | 1980 | 2000 | 2020

Period of Discovery | First AI Winter | Golden Age of NN | 2nd AI Winter | Age of Deep Learning

1943, McCulloch & Pitts
Logic Gates of The Mind

1957, Rosenblatt
Perceptron

1969, Minsky & Papert
Linear Models are Doomed

2001, Breiman
Random Forests

2011, Bengio
Init and ReLU

Read this: http://www.andreykurenkov.com/writing/a-brief-history-of-neural-nets-and-deep-learning/

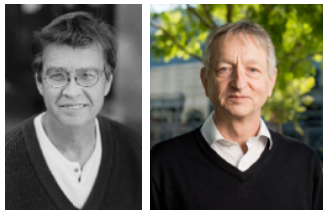
Machine Learning Timeline (Neural Nets)

- 2013: Andrew Ng and Google (BrainTeam)
  - run unsupervised feature creation on YouTube videos (becomes computer vision benchmark)

The work resulted in unsupervised neural net learning of an unprecedented scale - 16,000 CPU cores powering the learning of a whopping 1 billion weights. The neural net was trained on Youtube videos, entirely without labels, and learned to recognize the most common objects in those videos.

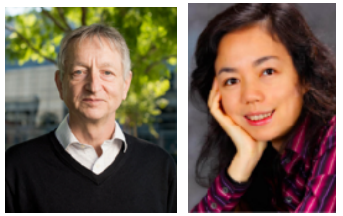
1949, Hebb's Law
Close neuron fire together

1960, Widrow & Hoff
Adaline Network



1986, Rumelhart & Hinton
Back-propagation

2003, Vapnik
Kernel SVMs



2012, Hinton, Fei-Fei Li
CNNs win ImageNet

1940 | 1960 | 1980 | 2000 | 2020

Period of Discovery | First AI Winter | Golden Age of NN | 2nd AI Winter | Age of Deep Learning

1943, McCulloch & Pitts
Logic Gates of The Mind

1957, Rosenblatt
Perceptron

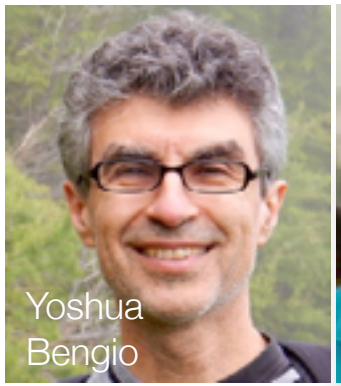
1969, Minsky & Papert
Linear Models are Doomed

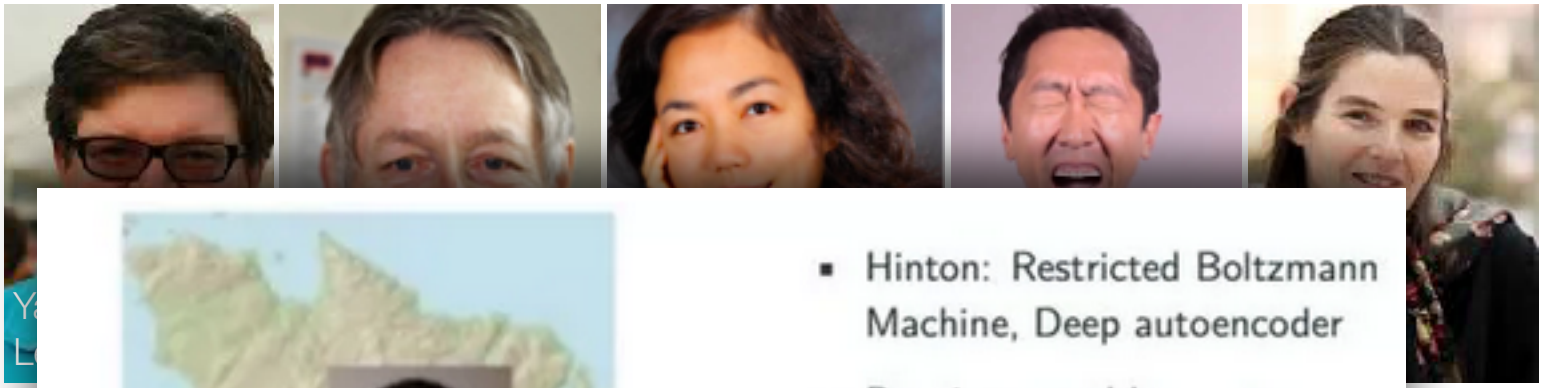
2001, Breiman
Random Forests

2011, Bengio
Init and ReLU

Read this: http://www.andreykurenkov.com/writing/a-brief-history-of-neural-nets-and-deep-learning/

A summary of the Deep Learning people:





Stayed at Univ. Montreal
Advises IBM

- Hinton: Restricted Boltzmann Machine, Deep autoencoder
- Bengio: neural language modeling.
- LeCun: Convolutional Neural Network
- NIPS, ICML, CVPR, ACL
- Google Brain, Deep Mind.
- FaceBook AI.

Read this paper as it sums up ac nicely

And predicts the future of deep learning

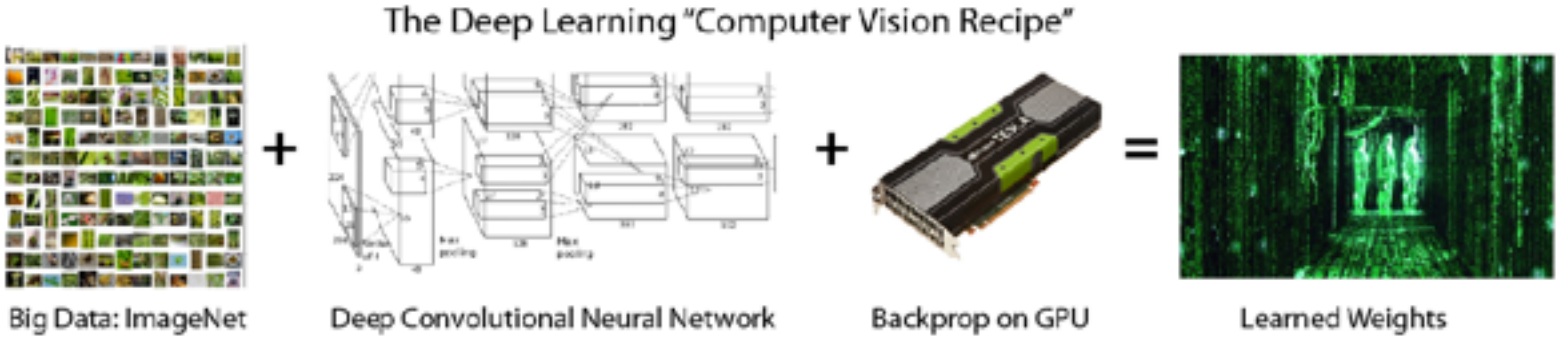
Yann LeCun[1,3], Yoshua Bengio[2] & Geoffrey Hinton[4,5]

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and

Lecture Notes for Machine Learnin

History of Deep Learning

- **Hinton** summarized what we learned in deep learning from the 2006 to present. Where we went wrong before present day:
  - labeled dataset were 1000s of times too small
  - computers were millions of times too slow
  - weights were initialized in stupid ways
  - we used the wrong non-linearities
- Or **Larson's Laws**:
  - use a GPU when possible, init weights for consistent gradient magnitude, ReLU/SiLU where it makes sense (like in early feedforward layers), and lots of dropout in the final layers that tend to learn more quickly!



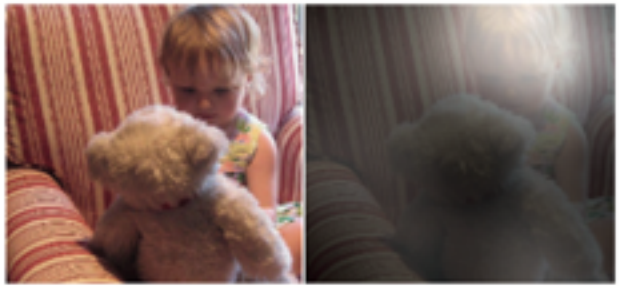
Read this: http://www.andreykurenkov.com/writing/a-brief-history-of-neural-nets-and-deep-learning/

Famous examples:

A woman is throwing a **frisbee** in a park.

A **dog** is standing on a hardwood floor.

A **stop** sign is on a road with a mountain in the background

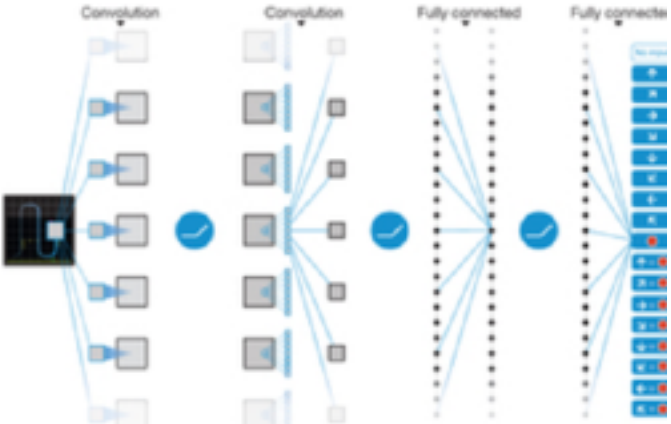


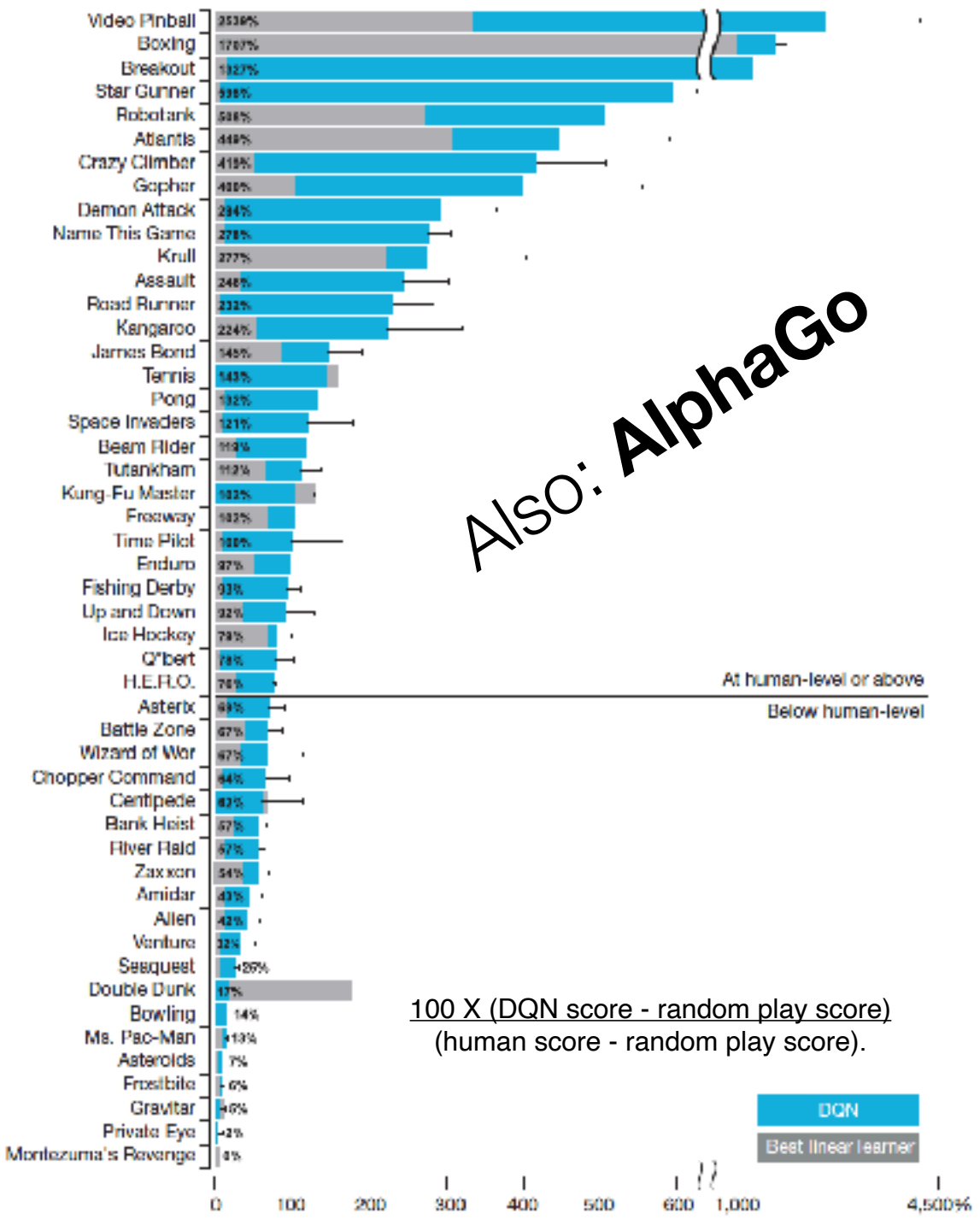
A little **girl** sitting on a bed with a teddy bear.

A group of **people** sitting on a boat in the water.

A giraffe standing in a forest with **trees** in the background.

More Famous





Also: **AlphaGo**

Credit for Deep Learning

Official ACM @TheOfficialACM

Yoshua Bengio, Geoffrey Hinton and Yann LeCun, the fathers of #DeepLearning, receive the 2018 #ACMTuringAward for conceptual and engineering breakthroughs that have made deep neural networks a critical component of computing today. bit.ly/2HVJtdV

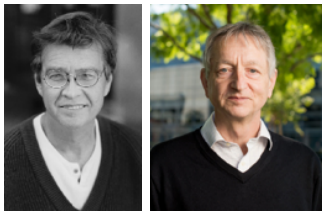
Machine learning is the science of credit assignment. The machine learning community itself profits from proper credit assignment to its members. The inventor of an important method should get credit for inventing it. She may not always be the one who popularizes it. Then the popularizer should get credit for popularizing it (but not for inventing it). Relatively young research areas such

Review of Deep Learning History

- Up to this point: back propagation saved AI winter for NN (Hinton and others!)
- 80's, 90's, 2000's: convolutional networks for image processing start to get deeper
  - but back propagation no longer does great job at training them
- SVMs and Random Forests gain traction…
  - The second AI winter begins, research in NN plummets
- 2004: Hinton secures funding from CIFAR in 2004 Hinton rebrands: Deep Learning
- 2006: Auto-encoding and Restricted Boltzmann Machines
- 2007: Deep networks are more efficient when pre-trained
- 2009: GPUs decrease training time by 70 fold…
- 2010: Hinton's students go to internships with Microsoft, Google, and IBM, making their speech recognition systems faster, more accurate and deployed in only 3 months…
- 2012: Hinton Lab, Google, IBM, and Microsoft jointly publish paper, popularity sky-rockets for deep learning methods
- 2011-2013: Ng and Google run unsupervised feature creation on YouTube videos (becomes computer vision benchmark)
- 2012+: Pre-training is not actually needed, just solutions for vanishing gradients (like ReLU, SiLU, initializations, more data, GPUs)

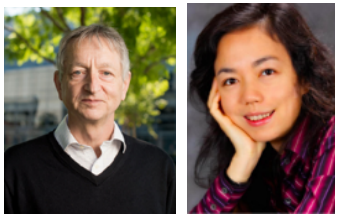
1949, Hebb's Law
Close neuron fire together

1960, Widrow & Hoff
Adaline Network

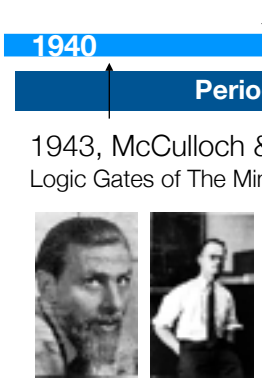


1986, Rumelhart & Hinton
Back-propagation

2003, Vapnik
Kernel SVMs



2012, Hinton, Fei-Fei Li
CNNs win ImageNet



1943, McCulloch & Pitts
Logic Gates of The Mind

1957, Rosenblatt
Perceptron

1969, Minsky & Papert
Linear Models are Doomed

2001, Breiman
Random Forests

2011, Bengio
Init and ReLU

2015, Google
Tensorflow Open Source

1940 | 1960 | 1980 | 2000 | 2020

Period of Discovery | First AI Winter | Golden Age of NN | 2nd AI Winter | Age of Deep Learning

End of Session

- Next Time:
  - Introduction to TensorFlow
  - Wide and Deep Networks