



deeplearning.ai

# Introduction to ML strategy

---

## Why ML Strategy?

# Motivating example



96%.

Ideas:

- Collect more data ←
- Collect more diverse training set
- Train algorithm longer with gradient descent
- Try Adam instead of gradient descent
- Try bigger network
- Try smaller network
- Try dropout
- Add  $L_2$  regularization
- Network architecture
  - Activation functions
  - # hidden units
  - ...



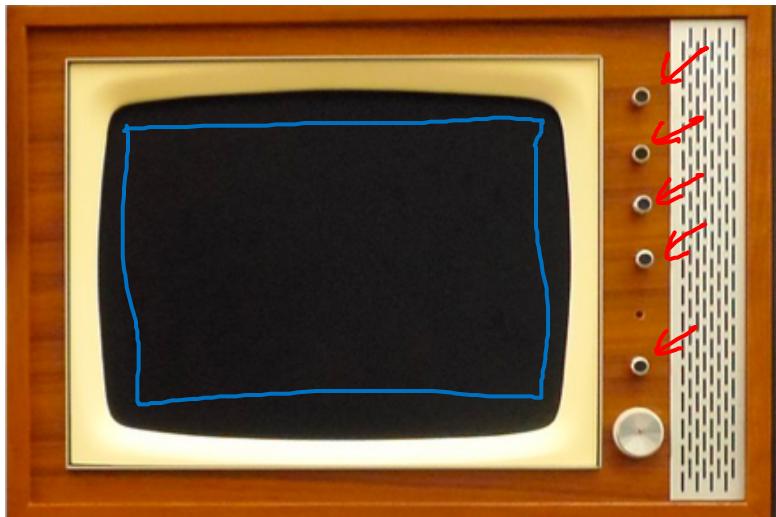
deeplearning.ai

# Introduction to ML strategy

---

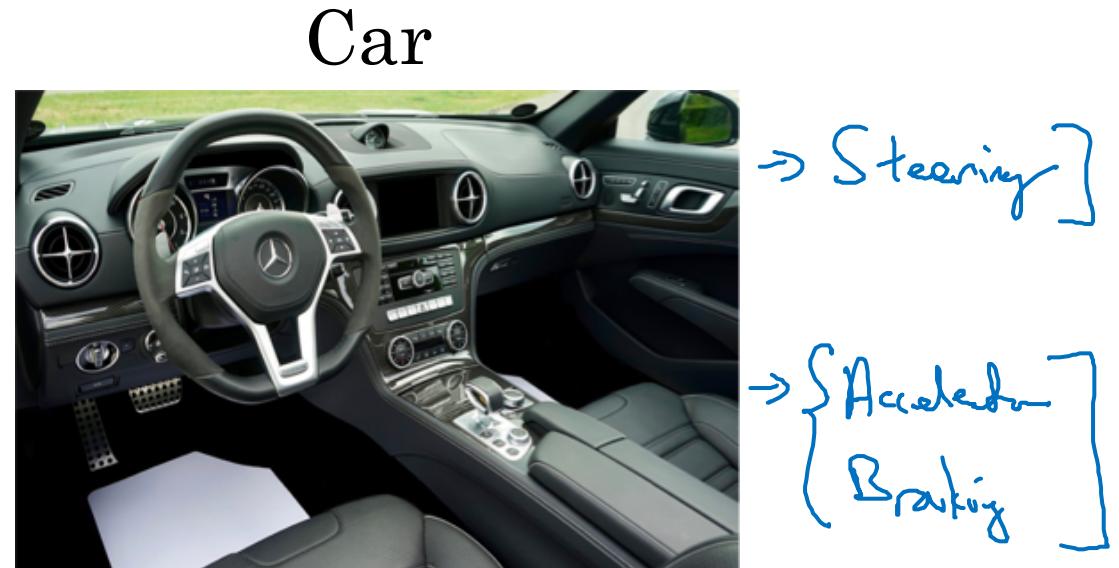
## Orthogonalization

# TV tuning example



Orthogonalization

$$\begin{aligned}
 & 0.1 \times \begin{array}{c} \uparrow \\ \downarrow \end{array} \\
 + & 0.3 \times \begin{array}{c} \leftarrow \\ \rightarrow \end{array} \\
 - & 1.7 \times \begin{array}{c} \diagup \\ \diagdown \end{array} \\
 + & 0.8 \times \begin{array}{c} \leftarrow \\ \rightarrow \end{array} \\
 + \dots & \begin{array}{c} \diagup \\ \diagdown \end{array}
 \end{aligned}$$



$$\rightarrow \underline{0.3 \times \text{angle}} - 0.8 \times \text{speed}$$

$$\rightarrow 2 \times \text{angle} + 0.9 \times \text{speed}.$$

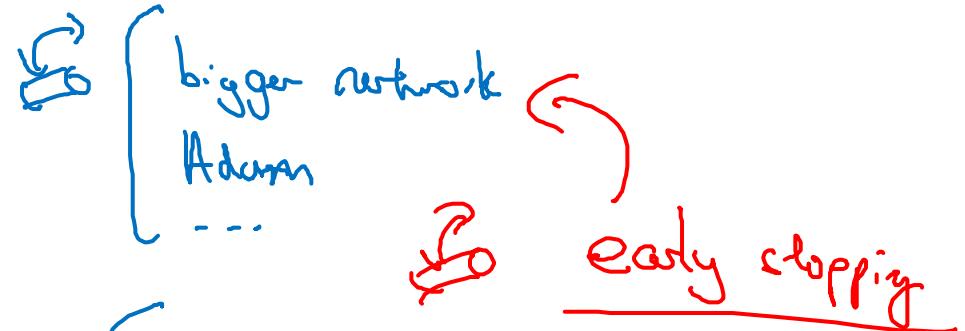


# Chain of assumptions in ML

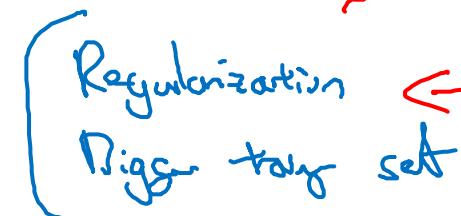
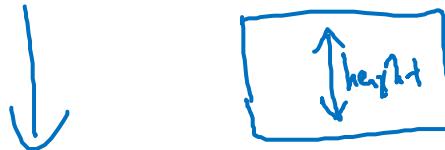
→ Fit training set well on cost function



(≈ human-level performance)



→ Fit dev set well on cost function



→ Fit test set well on cost function

Bigger dev set

→ Performs well in real world

(Happy cat pic off users.)

Change dev set or  
cost function



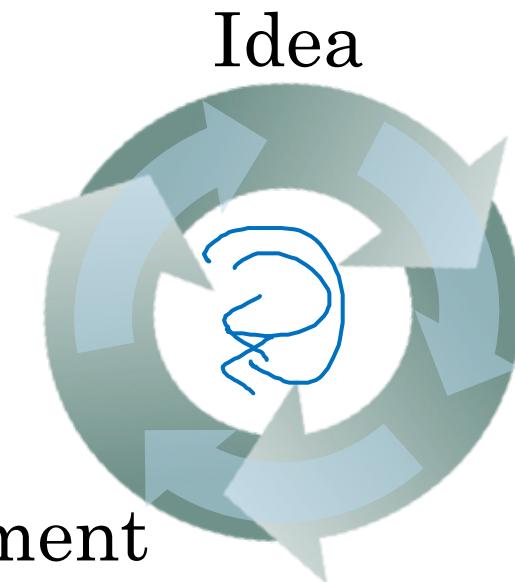
deeplearning.ai

Setting up  
your goal

---

Single number  
evaluation metric

# Using a single number evaluation metric



# Code

→ Of examples recognized as cert.,  
what % actually are certs?

→ what % of actual certs  
are correctly recognized

Classifier	Precision	Recall
A	<u>95%</u>	90%
B	98%	85%

F<sub>1</sub> Score = "Average" of P and R.

$$\left( \frac{2}{\frac{1}{P} + \frac{1}{R}} \cdot \text{"Harmonic Mean"} \right)$$

Dev set + Single number evaluation metric

# Another example

Algorithm	US	China	India	Other
A	<u>3%</u>	7%	5%	9%
B	5%	6%	5%	10%
C	2%	3%	4%	5%
D	5%	8%	7%	2%
E	4%	5%	2%	4%
F	7%	11%	8%	12%



deeplearning.ai

Setting up  
your goal

---

Satisficing and  
optimizing metrics

# Another cat classification example

Classifier	Accuracy	Running time
A	90%	80ms
B	92%	95ms
C	95%	1,500ms

optimizing



satisficing



$$\text{Cost} = \underline{\text{accuracy}} - 0.5 \times \underline{\text{running Time}}$$

Maximize accuracy

Subject to running Time  $\leq \underline{100 \text{ ms.}}$

N metrics : 1 optimizing

N-1 satisficing

Wakewords / trigger words

Alexa, OK Google,

Hey Siri, nihao baidu

你好 百度

accuracy.

#false positive

Maximize accuracy.

s.t.  $\leq 1$  false positive  
every 24 hours.



deeplearning.ai

Setting up  
your goal

---

Train/dev/test  
distributions

# Cat classification dev/test sets

↳ development set, hold out cross validation set

Regions:

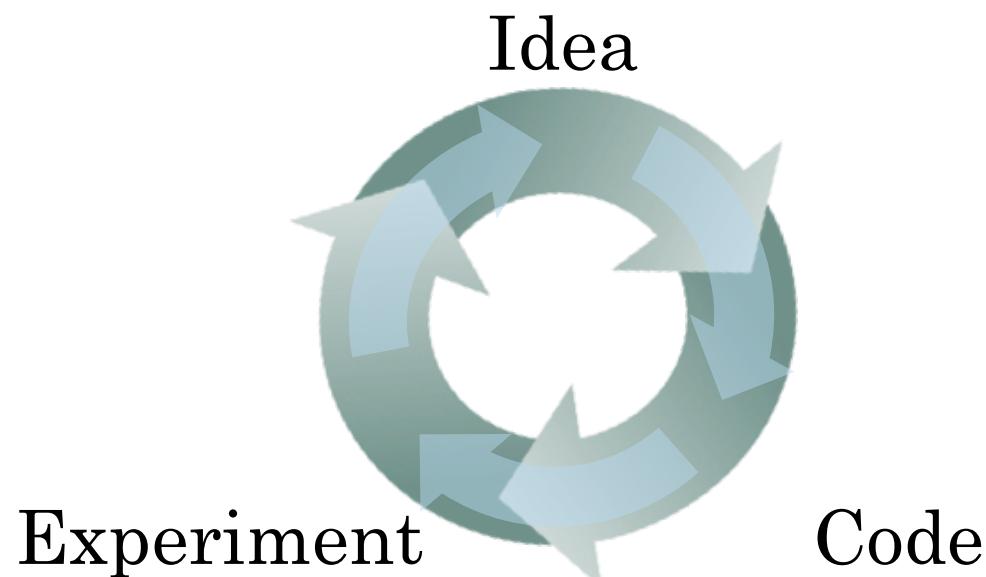
- US
- UK
- Other Europe
- South America
- India
- China
- Other Asia
- Australia



Randomly shuffle into dev/test



dev set  
+  
metric

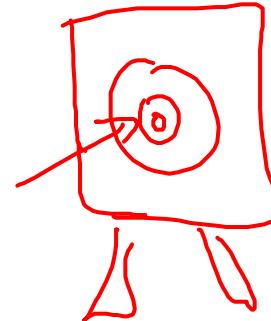


# True story (details changed)

[ Optimizing on dev set on loan approvals for  
medium income zip codes

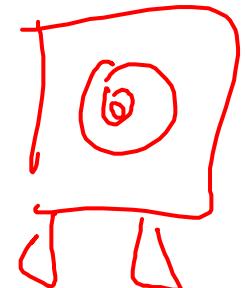


$x \rightarrow y$  (repay loan?)



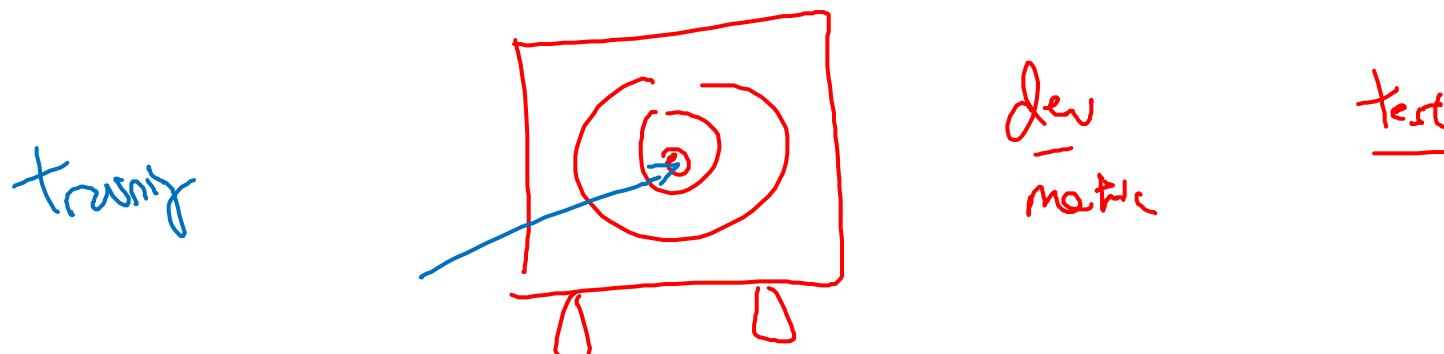
[ Tested on low income zip codes

$\sim 3$  month



# Guideline

Choose a dev set and test set to reflect data you expect to get in the future and consider important to do well on.





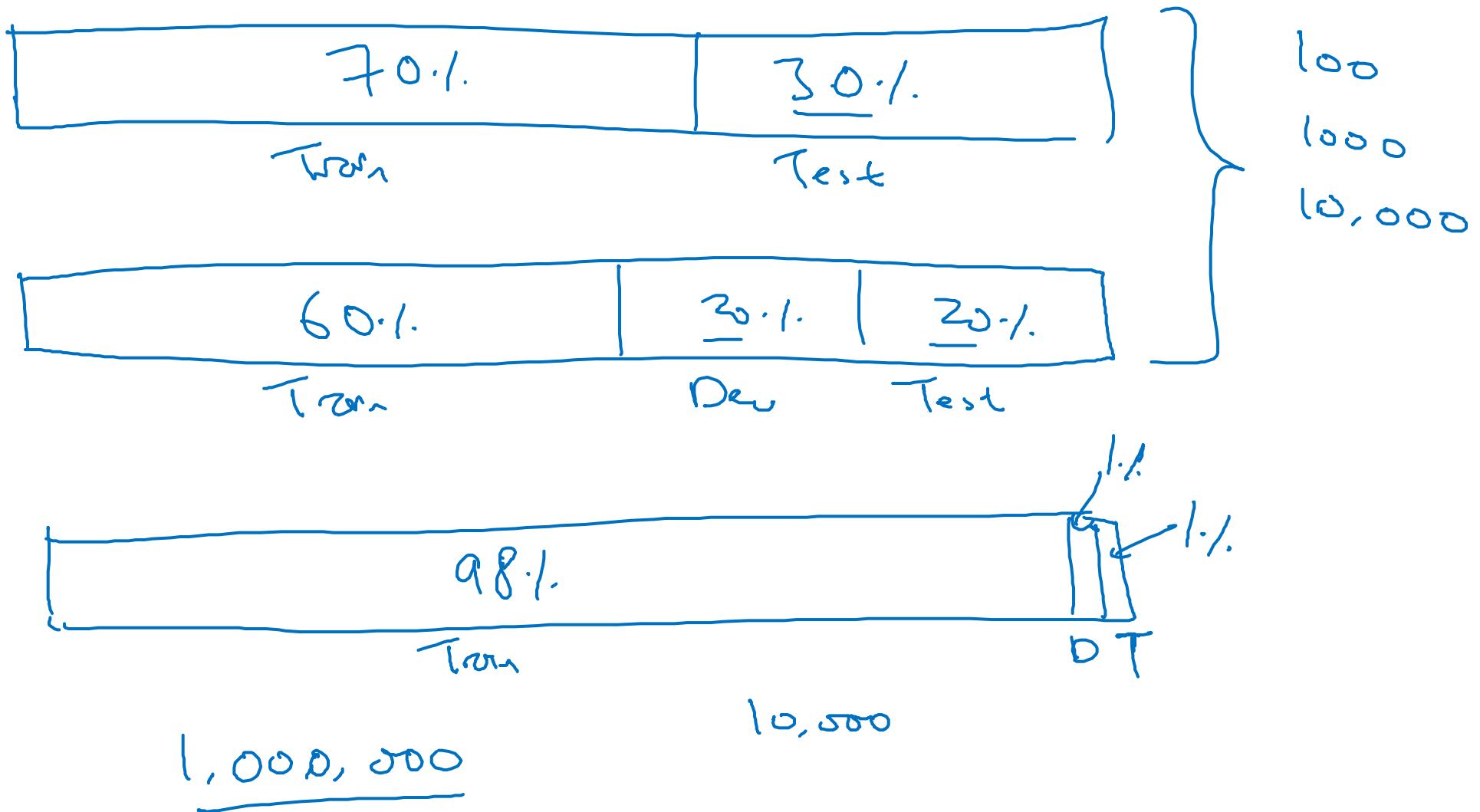
deeplearning.ai

Setting up  
your goal

---

Size of dev  
and test sets

# Old way of splitting data



# Size of dev set

A    B

Set your dev set to be big enough to detect differences in  
algorithm/models you're trying out.

100: small  
10%

A                          B  
97% → 97.1 %  
0.1%  
1%

1,000

10,000

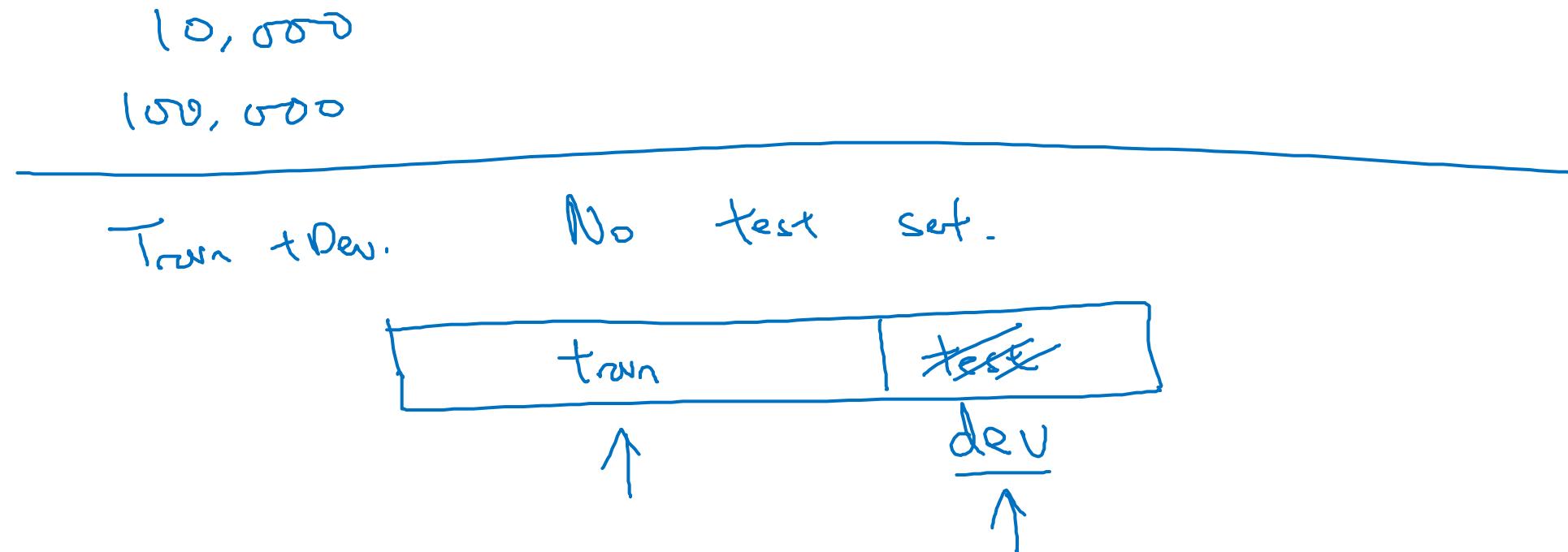
100,000

0.01%  
0.001%

Online advertising

# Size of test set

- Set your test set to be big enough to give high confidence in the overall performance of your system.





deeplearning.ai

Setting up  
your goal

---

When to change  
dev/test sets and  
metrics

# Cat dataset examples

Metric + Dev : Prefer A  
You/users : Prefer B.

→ Metric: classification error

Algorithm A: 3% error → Pornographic

✓ Algorithm B: 5% error

Error:  $\frac{1}{\sum_i w^{(i)}} \times m_{dev}$

$$\sum_{i=1}^{m_{dev}} \frac{w^{(i)}}{\{y_{pred}^{(i)} + y^{(i)}\}}$$

predval value (0/1)

$$w^{(i)} = \begin{cases} 1 & \text{if } x^{(i)} \text{ is non-porn} \\ 10 & \text{if } x^{(i)} \text{ is porn} \end{cases}$$

# Orthogonalization for cat pictures: anti-porn

- 1. So far we've only discussed how to define a metric to evaluate classifiers. ← Place target 
- 2. Worry separately about how to do well on this metric. 

An (shot at target)

$$\rightarrow J = \frac{1}{\sum w^{(i)}} \sum_{i=1}^m w^{(i)} \ell(\hat{y}^{(i)}, y^{(i)})$$

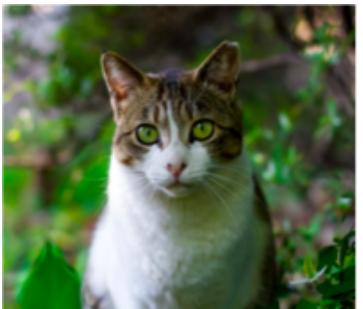


# Another example

Algorithm A: 3% error

✓ Algorithm B: 5% error ↙

→ Dev/test ↘



→ User images ↗



If doing well on your metric + dev/test set does not correspond to doing well on your application, change your metric and/or dev/test set.



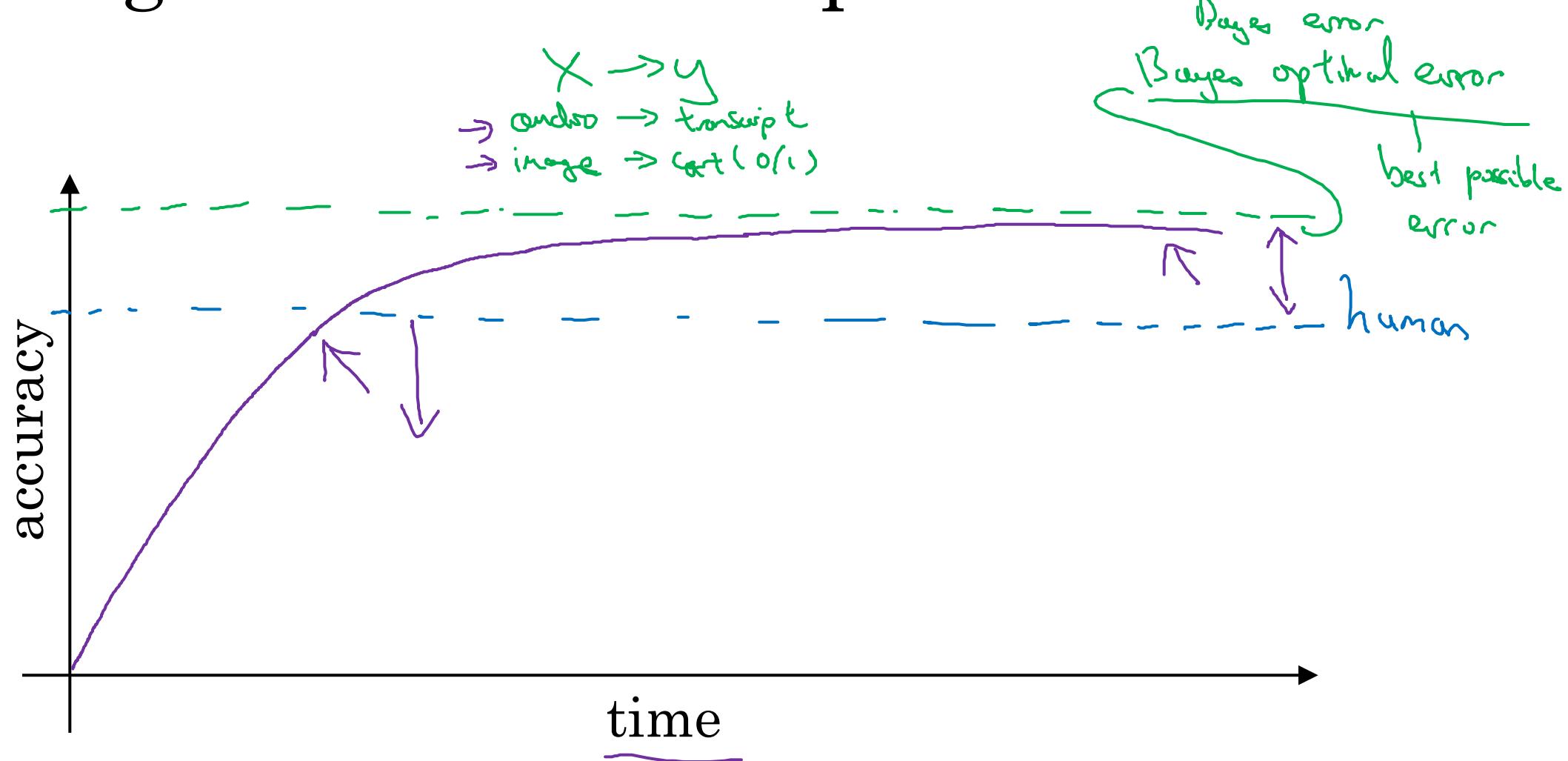
deeplearning.ai

Comparing to human-level performance

---

Why human-level performance?

# Comparing to human-level performance



# Why compare to human-level performance

Humans are quite good at a lot of tasks. So long as ML is worse than humans, you can:

- - Get labeled data from humans.  $(x, y)$
- - Gain insight from manual error analysis:  
Why did a person get this right?
- - Better analysis of bias/variance.



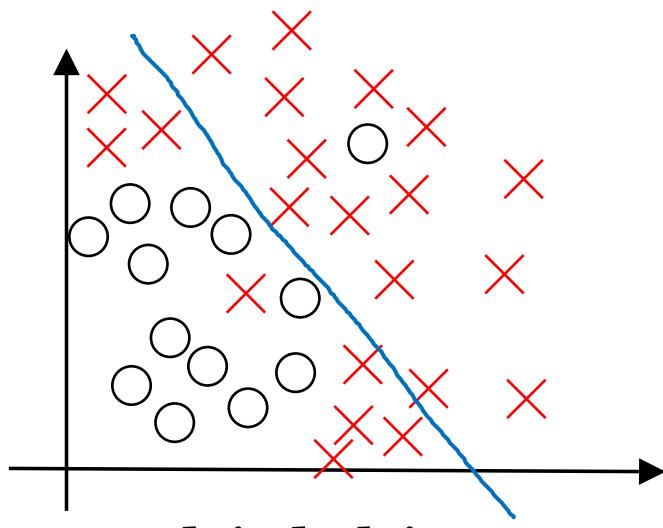
deeplearning.ai

Comparing to human-level performance

---

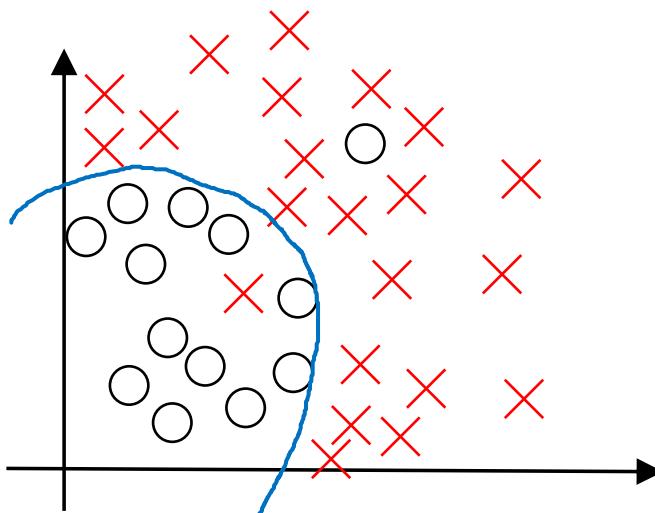
Avoidable bias

# Bias and Variance

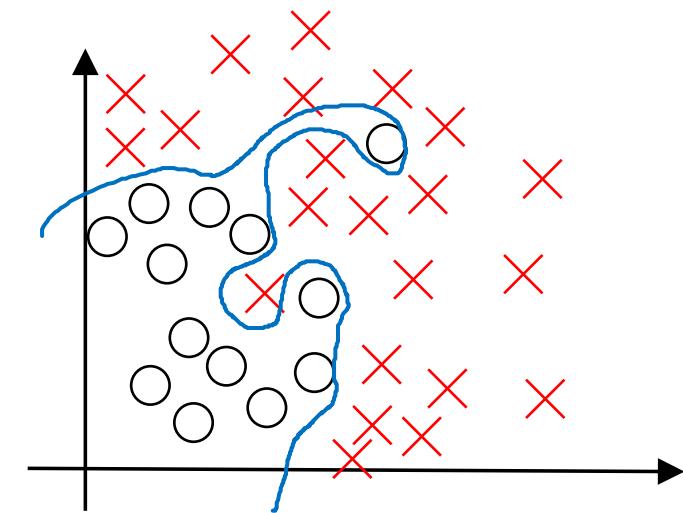


high bias

*Underfitting*



“just right”



high variance

*Overfitting*

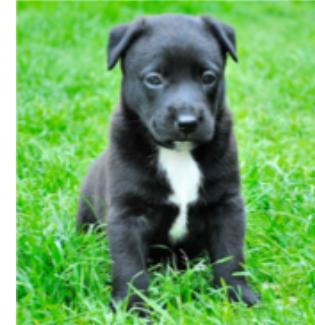
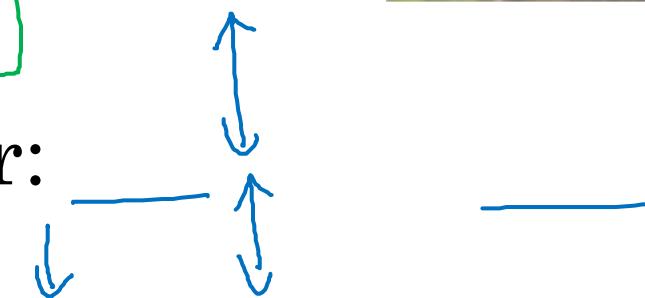
# Bias and Variance

Cat classification

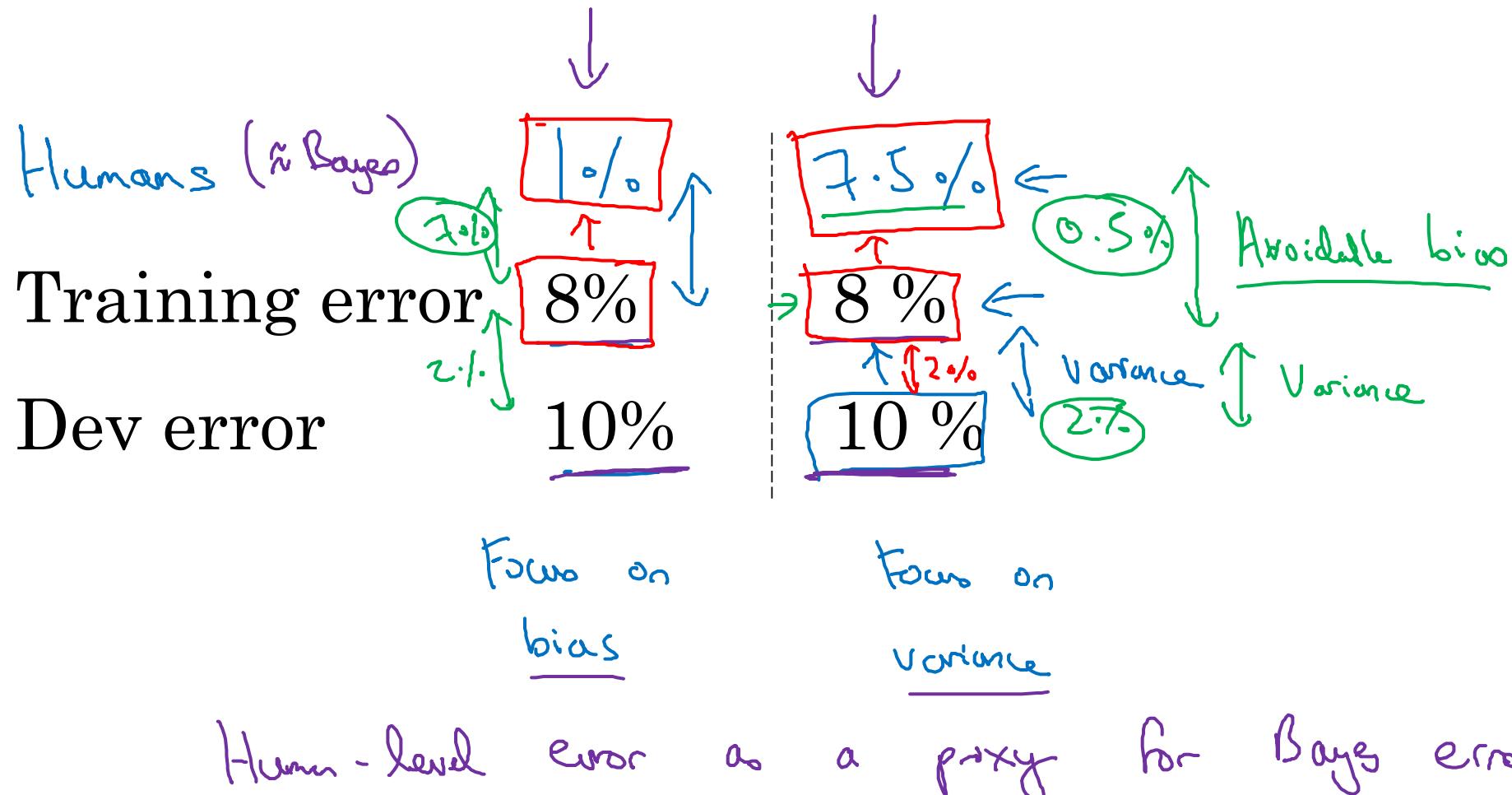
Human-level  $\approx 0\%$

Training set error:

Dev set error:



# Cat classification example





deeplearning.ai

Comparing to human-level performance

---

Understanding  
human-level  
performance

# Human-level error as a proxy for Bayes error

Medical image classification example:

Suppose:

- (a) Typical human ..... 3 % error
- (b) Typical doctor ..... 1 % error
- (c) Experienced doctor ..... 0.7 % error
- (d) Team of experienced doctors .. 0.5 % error



What is “human-level” error?

$$\text{Baye error} \leq 0.5\%$$

# Error analysis example

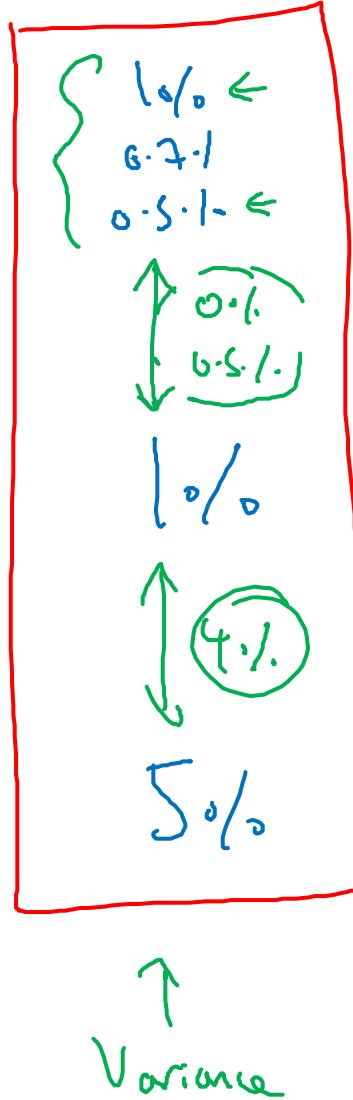
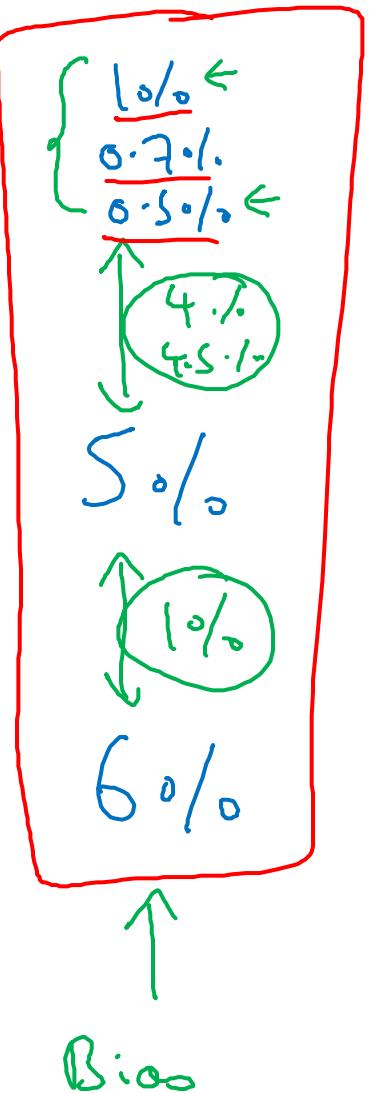
Human (proxy for Bayes error)



Training error

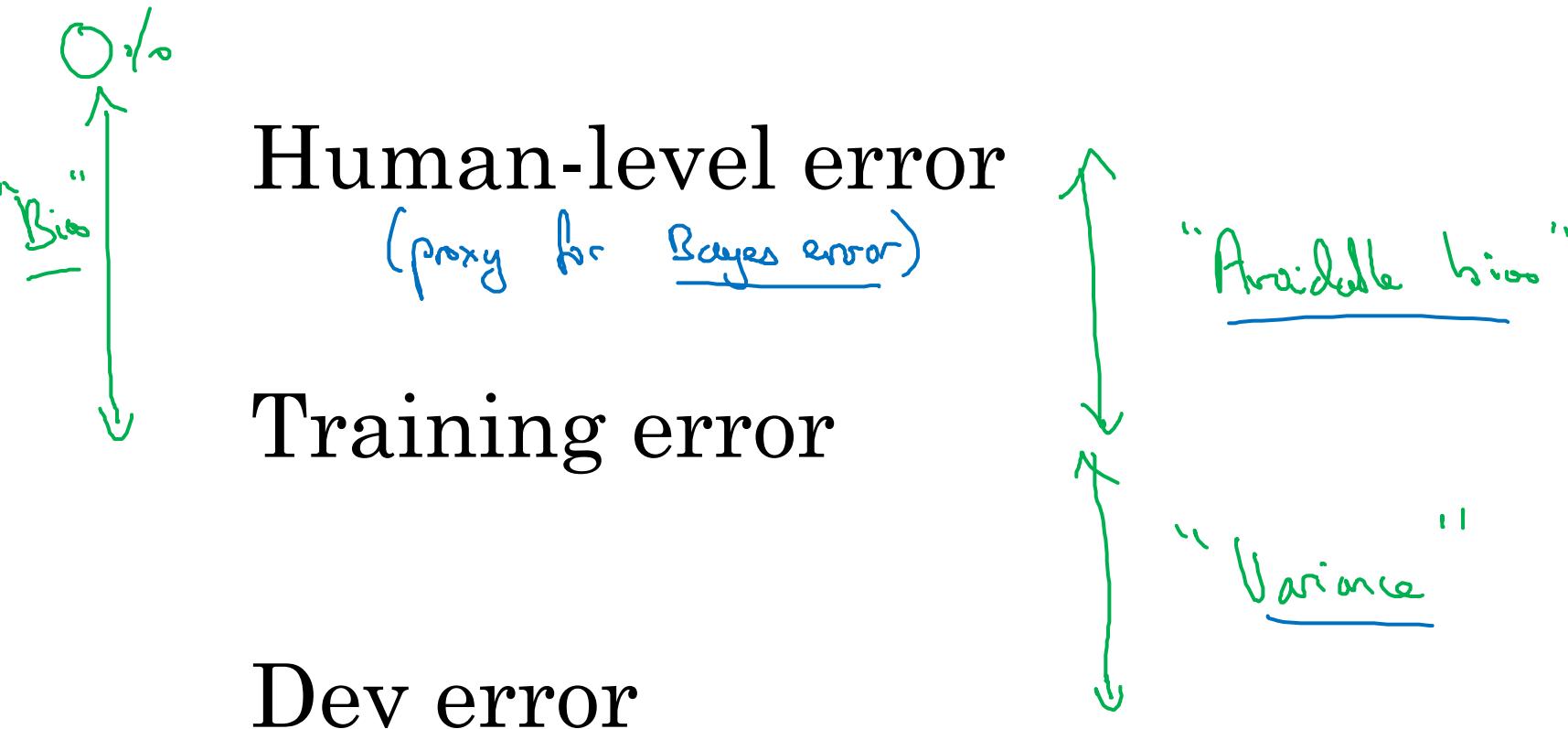


Dev error



$$\begin{aligned} &\rightarrow \frac{0.7\%}{0.5\%} \quad 1\% \leftarrow \\ &\rightarrow \underline{\underline{0.5\%}} \quad 0.2\% \leftarrow \\ &0.0\% \leftarrow \\ &\rightarrow \boxed{0.7\%} \quad 0.0\% \leftarrow \\ &\rightarrow \underline{\underline{0.7\%}} \quad 0.1\% \leftarrow \\ &0.0\% \leftarrow \\ &\rightarrow 0.8\% \end{aligned}$$

# Summary of bias/variance with human-level performance





deeplearning.ai

Comparing to human-level performance

---

Surpassing human-level performance

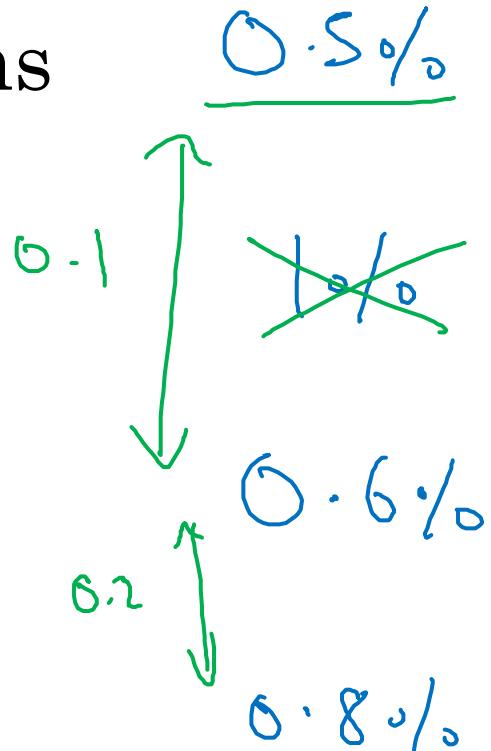
# Surpassing human-level performance

Team of humans

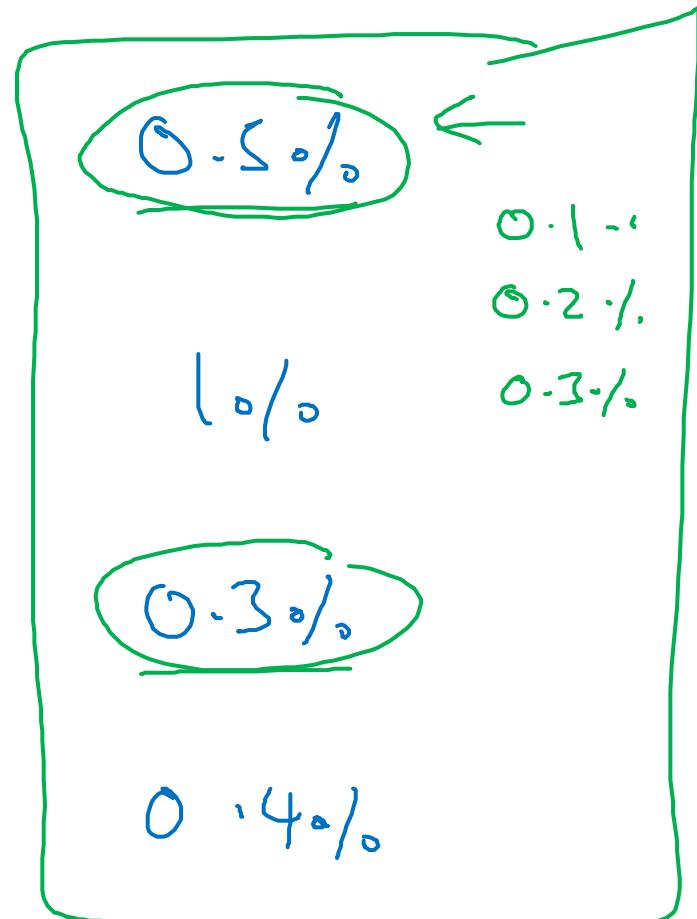
One human

Training error

Dev error

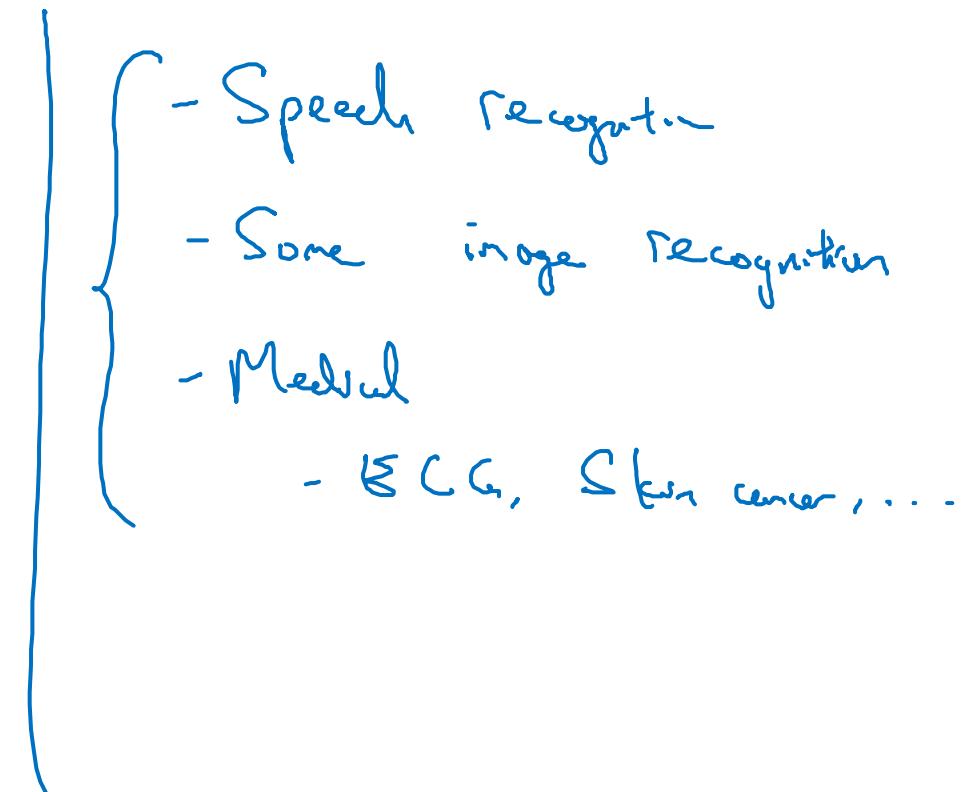


What is avoidable bias?



# Problems where ML significantly surpasses human-level performance

- - Online advertising
- - Product recommendations
- - Logistics (predicting transit time)
- - Loan approvals



Structural data

Not natural perception

Lots of data



deeplearning.ai

Comparing to human-level performance

---

Improving your model performance

# The two fundamental assumptions of supervised learning

1. You can fit the training set pretty well.



$\sim$  Avoidable bias

2. The training set performance generalizes pretty well to the dev/test set.



$\sim$  Variance

# Reducing (avoidable) bias and variance

