



deeplearning.ai

# Introduction to ML strategy

---

## Why ML Strategy?

# Motivating example



90%

## 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
  - ...



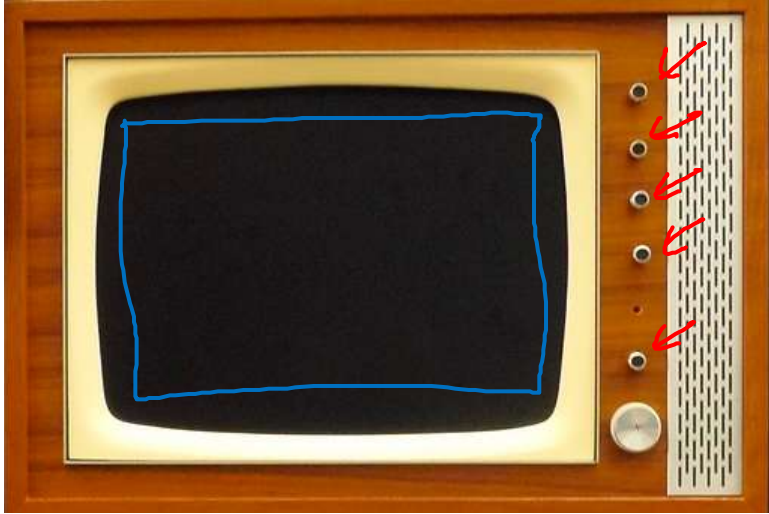
deeplearning.ai

# Introduction to ML strategy

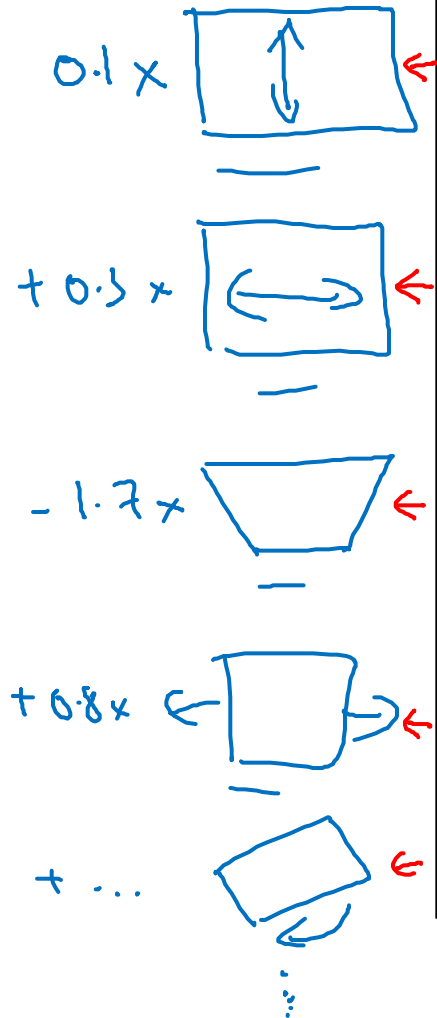
---

## Orthogonalization

# TV tuning example



Orthogonalization



## Car

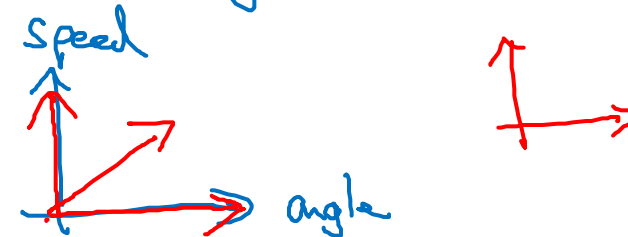


$\rightarrow$  Steering]

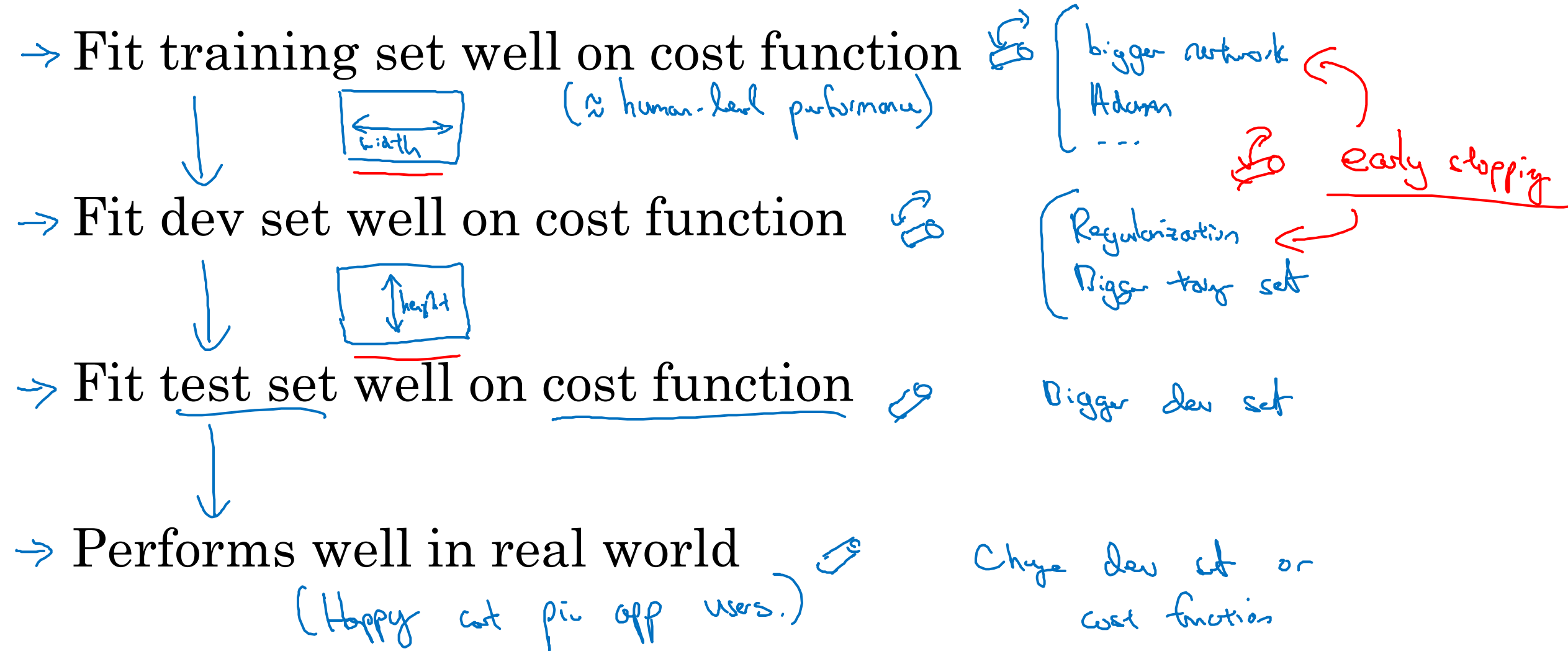
$\rightarrow$  { Accelerate  
Braking }

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

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



# Chain of assumptions in ML





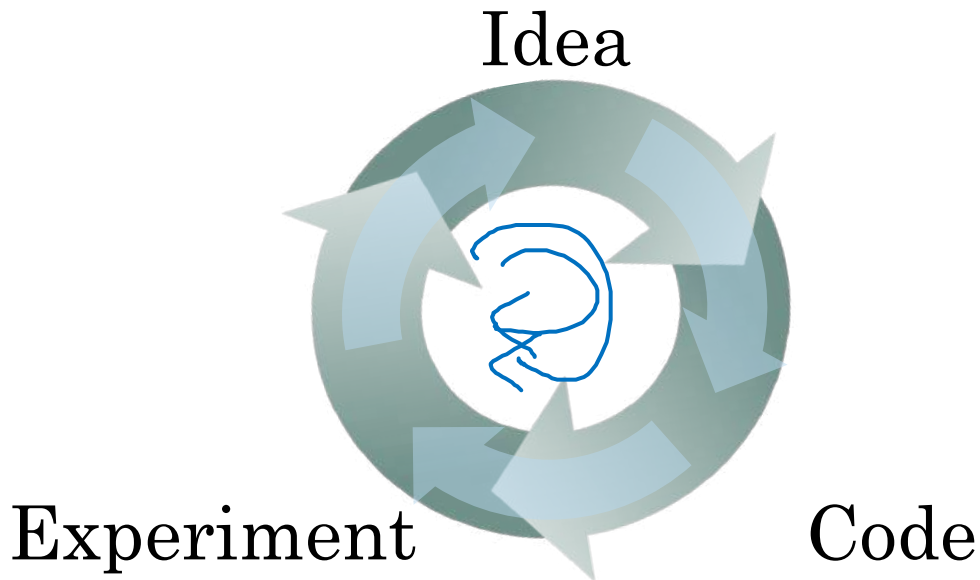
deeplearning.ai

Setting up  
your goal

---

Single number  
evaluation metric

# Using a single number evaluation metric



→ Of examples recognized as cost, what % actually are costs?

→ what % of actual costs are correctly recognized

Classifier	Precision	Recall
A	95%	90%
B	98%	85%

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

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

Dev set + Single number evaluation metric  
real speed up iterating

# 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

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

maximize accuracy

subject to Running Time  $\leq$  100 ms.

N metrics : 1 optimizing  
N-1 satisfying

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

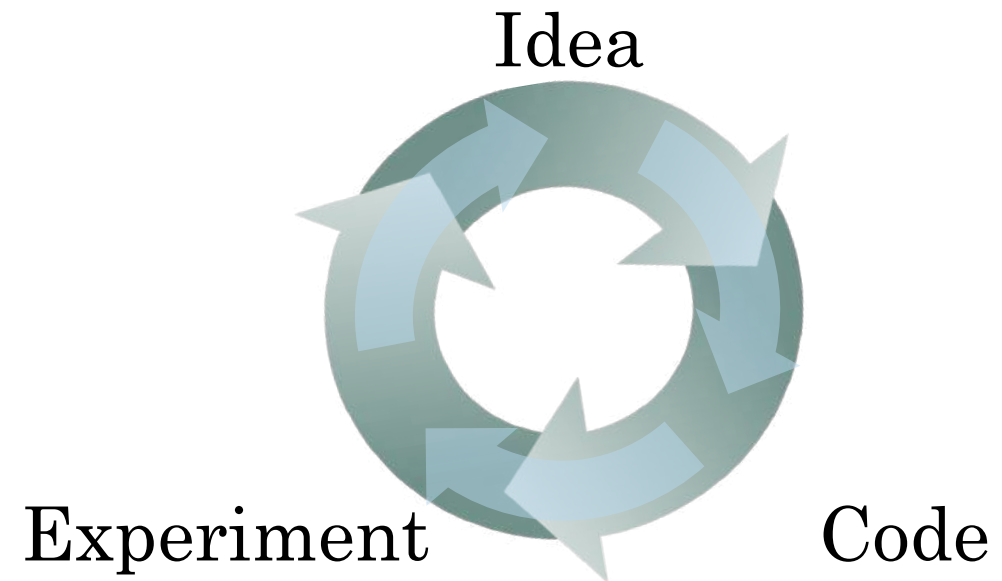
Dev

Test

→ 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

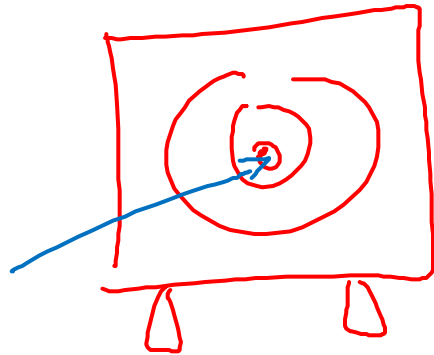
~ 3 month



# 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.

training



dev  
metric

test



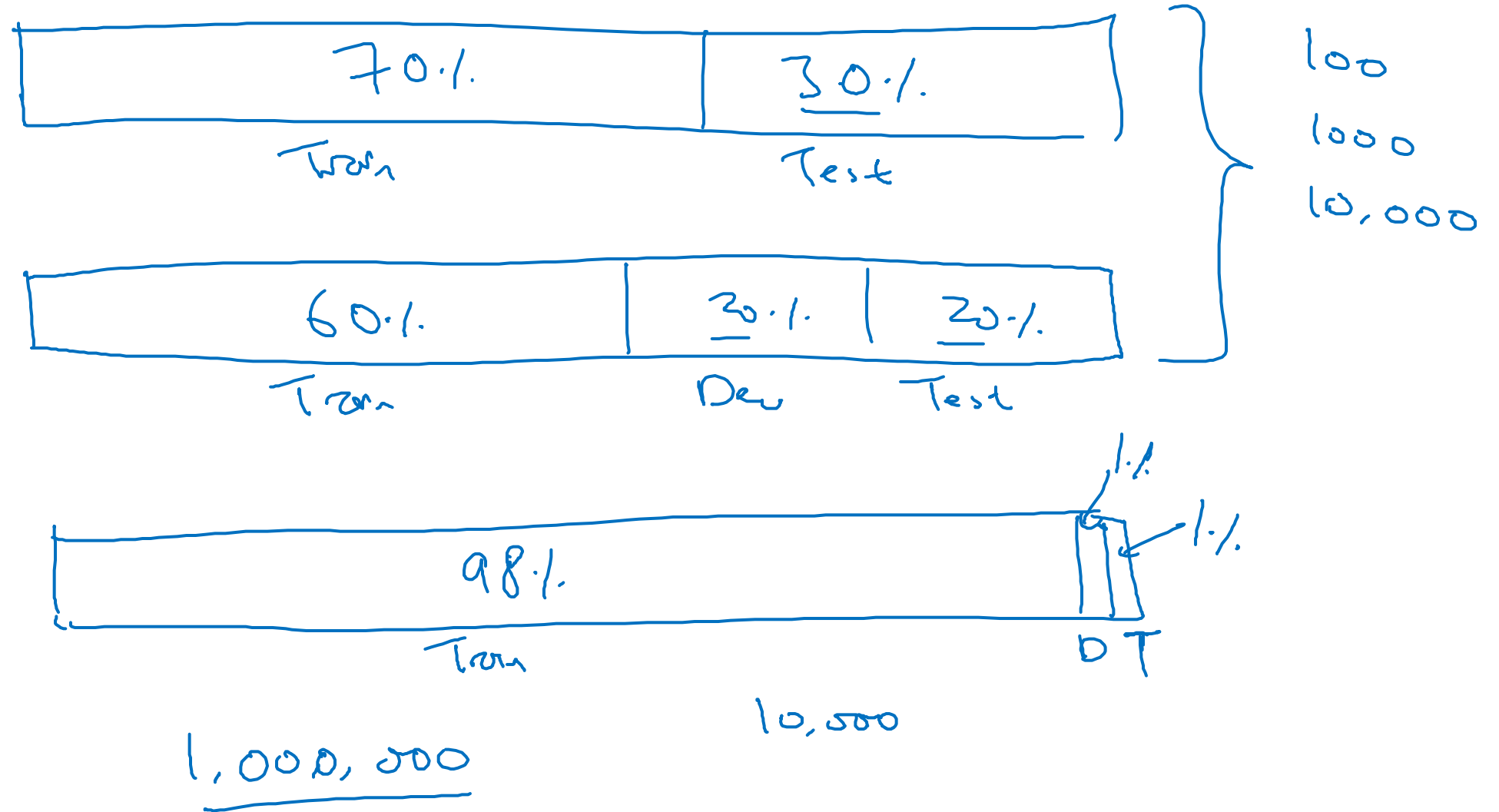
deeplearning.ai

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  
└ 1%

1,000

10,000

100,000

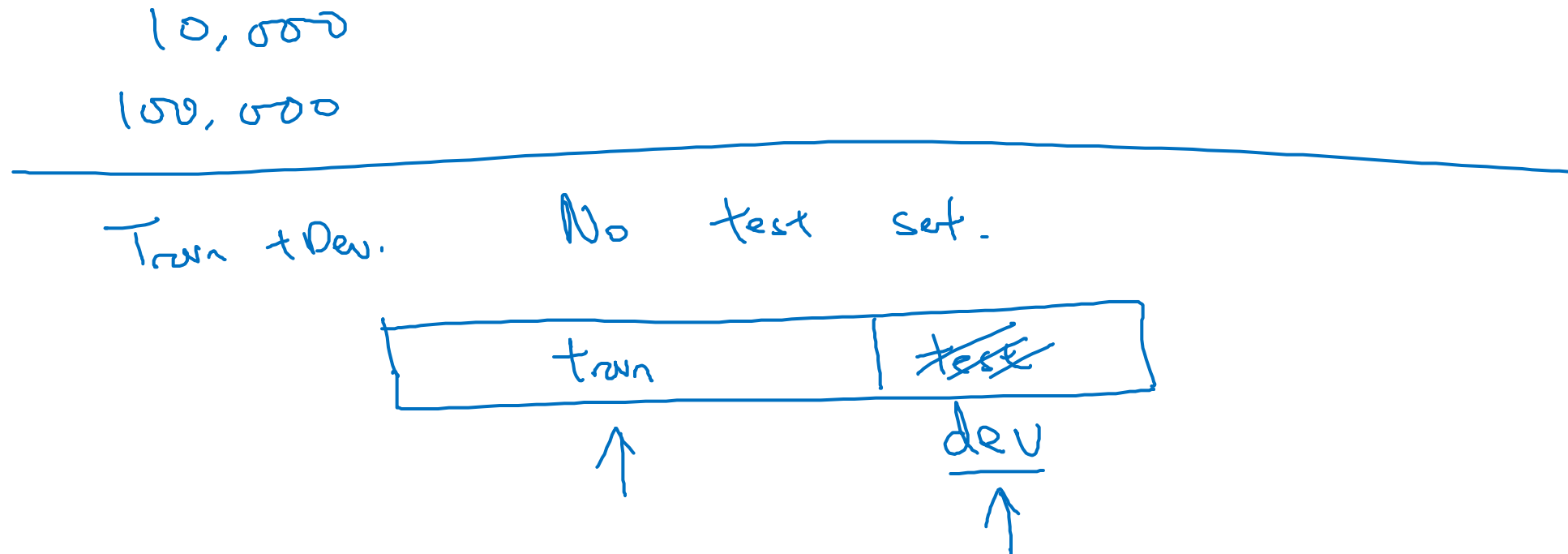
<sup>A</sup> 97% → <sup>B</sup> 97.1%  
0.1%  
└

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:  $\frac{1}{\sum_i w^{(i)}} \cdot \frac{1}{m_{dev}} \sum_{i=1}^{m_{dev}} w^{(i)} \mathbb{I}\{y_{pred}^{(i)} \neq y^{(i)}\}$

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

$\mathbb{I}\{y_{pred}^{(i)} \neq y^{(i)}\}$   
predicted value (0/1)

# 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. ↺
- ↗ Aim (shoot at target)

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



# 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.



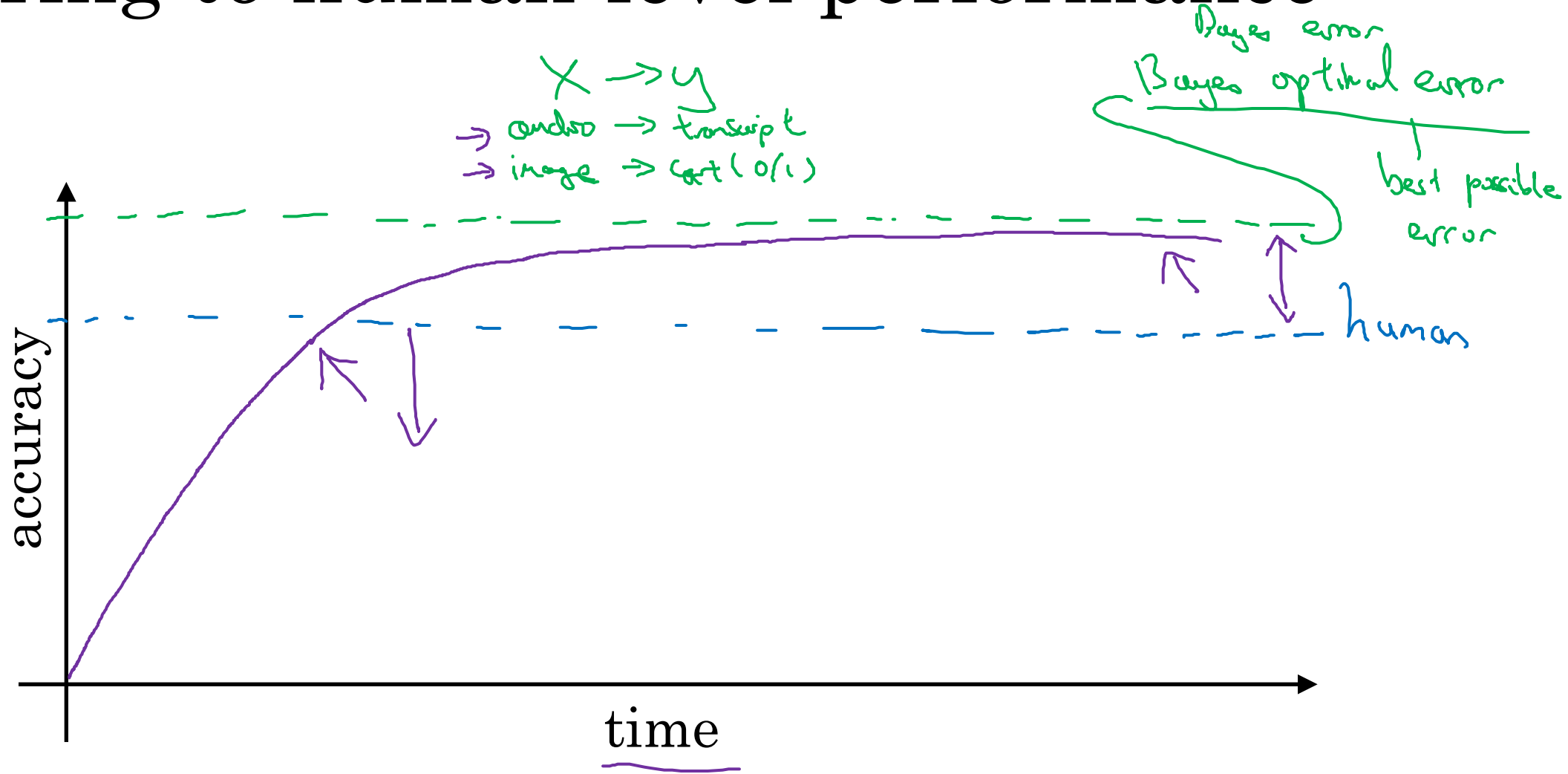
deeplearning.ai

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.



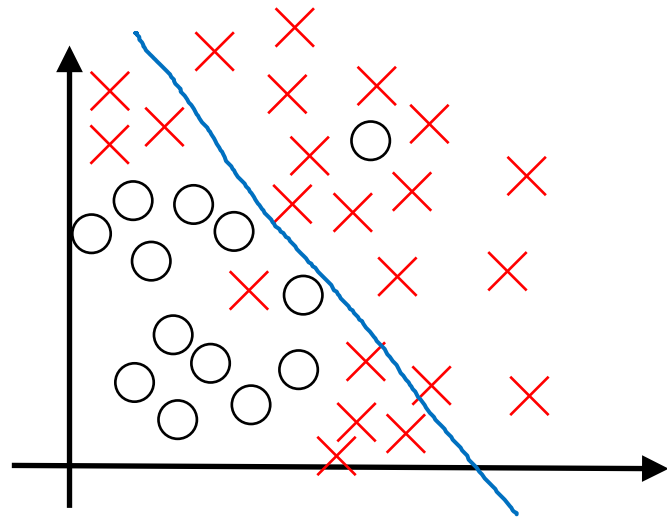
deeplearning.ai

Comparing to human-  
level performance

---

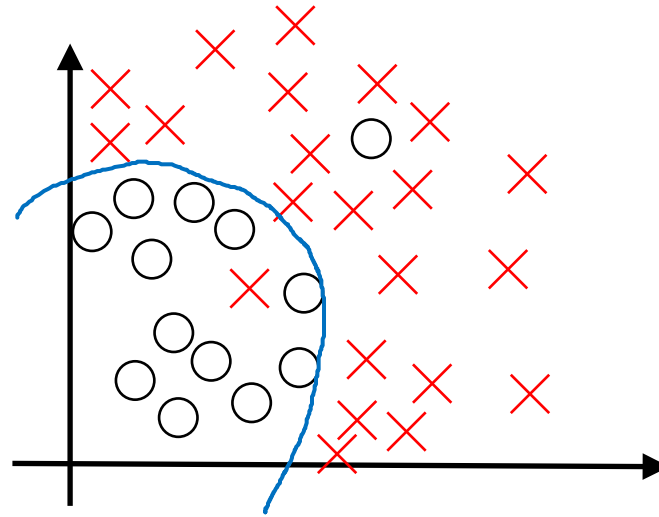
**Avoidable bias**

# Bias and Variance

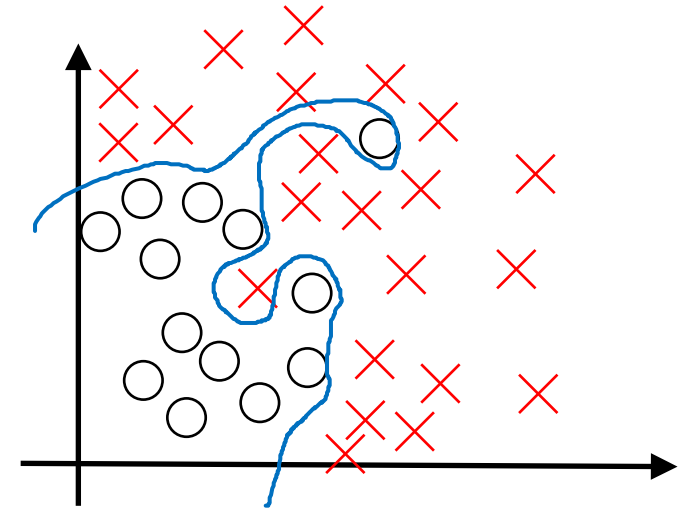


high bias

*underfitting*



“just right”



high variance

*overfitting*

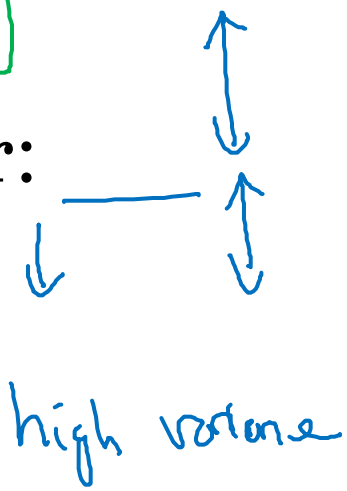
# Bias and Variance

Cat classification

Human-level  $\approx 0\%$  ----

Training set error:

Dev set error:

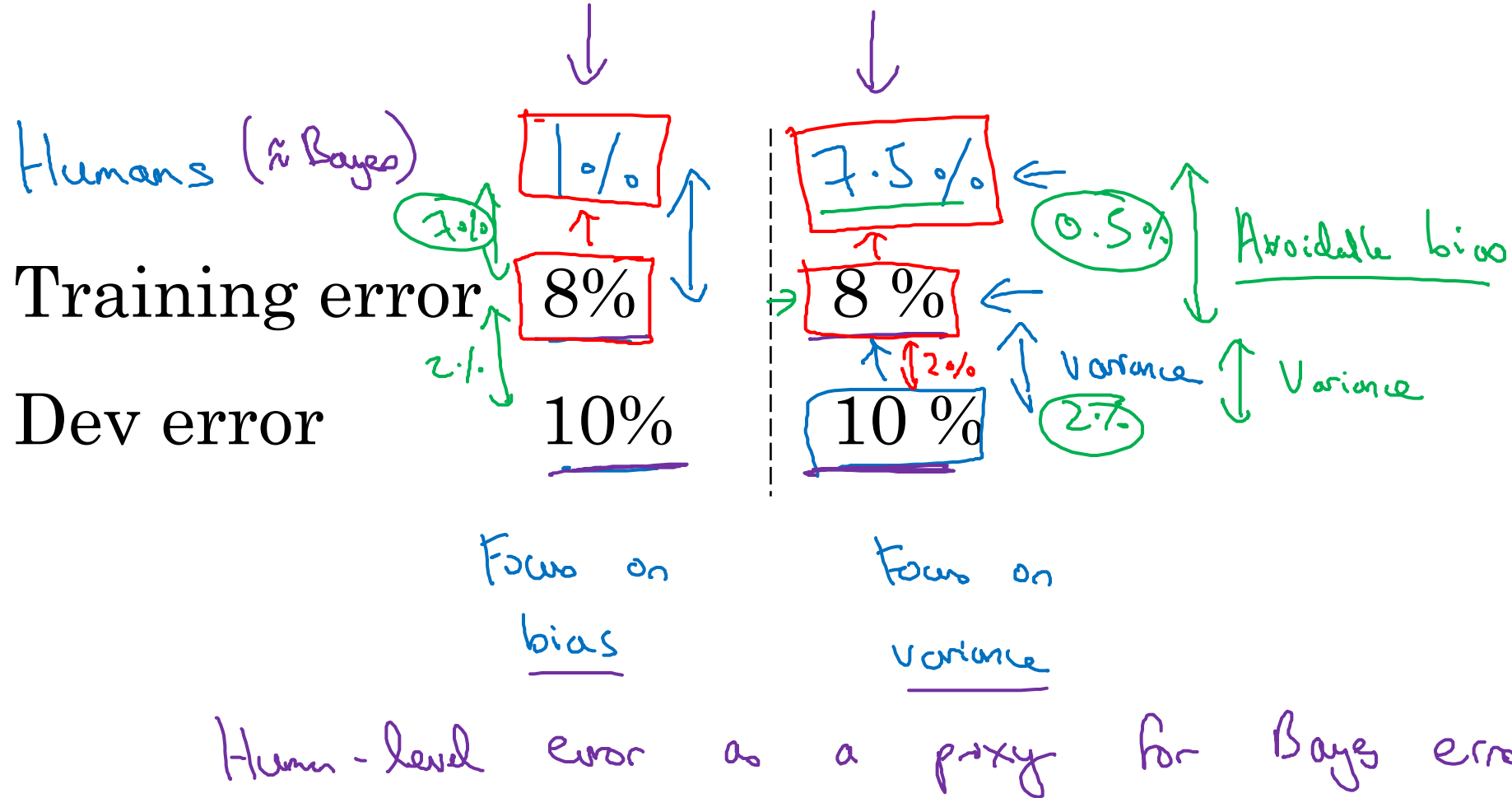


high bias

high bias  
high variance

low bias  
low variance

# 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 ←

Bayes error  $\leq$  0.5 %

What is “human-level” error?



# Error analysis example

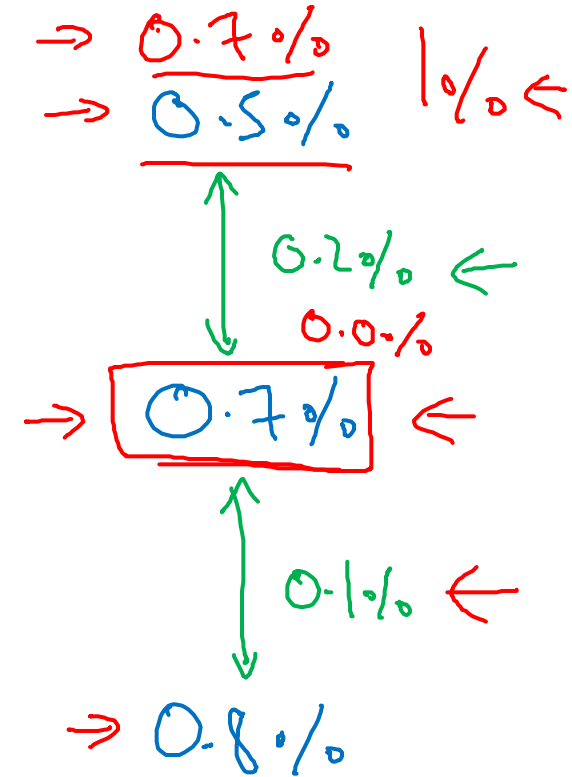
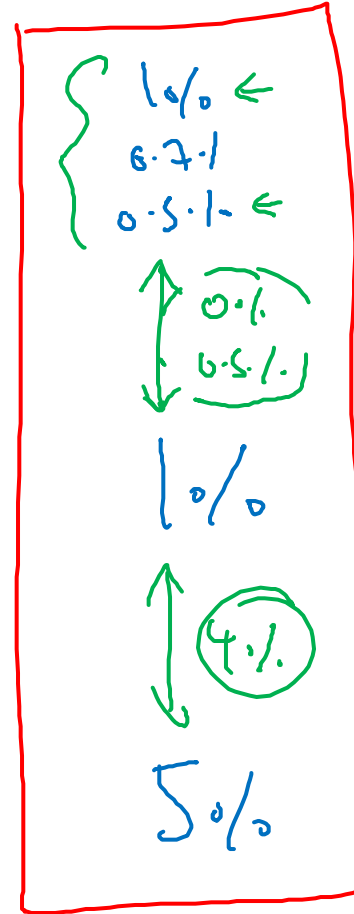
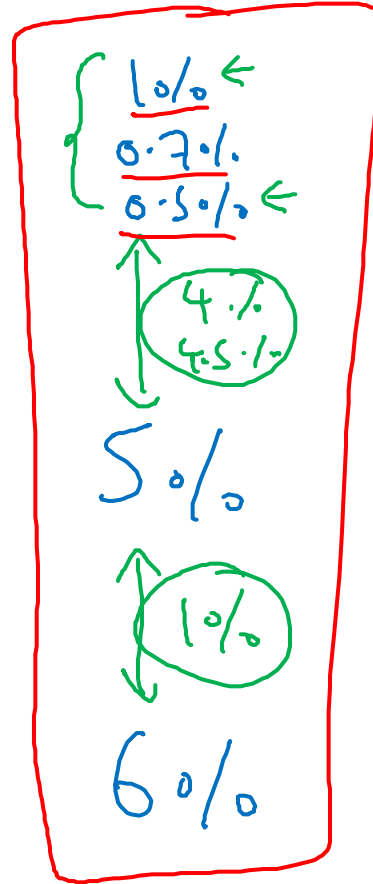
Human (proxy for Bayes error)

↑ Avoidable bias  
↓

Training error

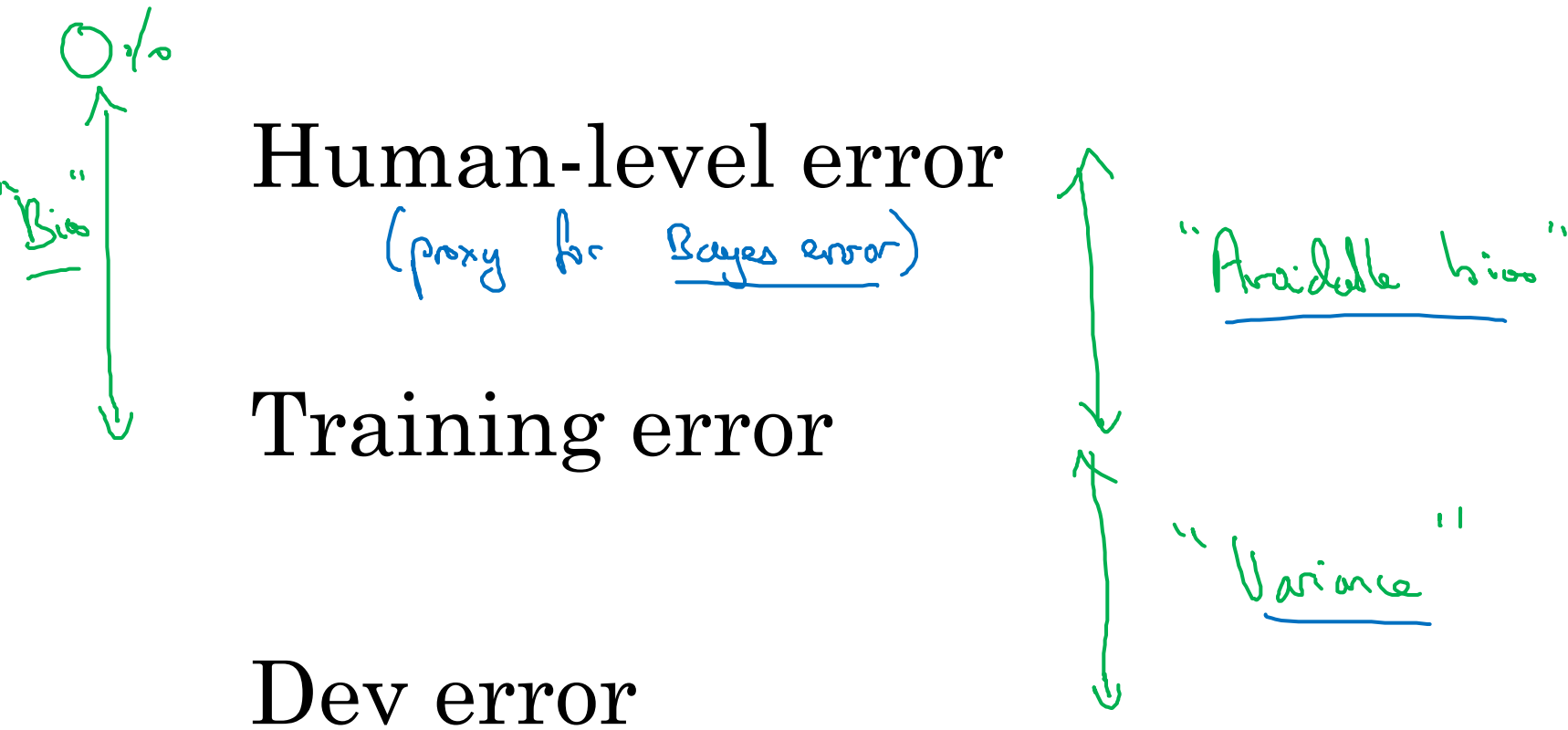
↑ Variance  
↓

Dev error





# 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

0.5%

One human

0.1  $\updownarrow$  ~~1.0%~~

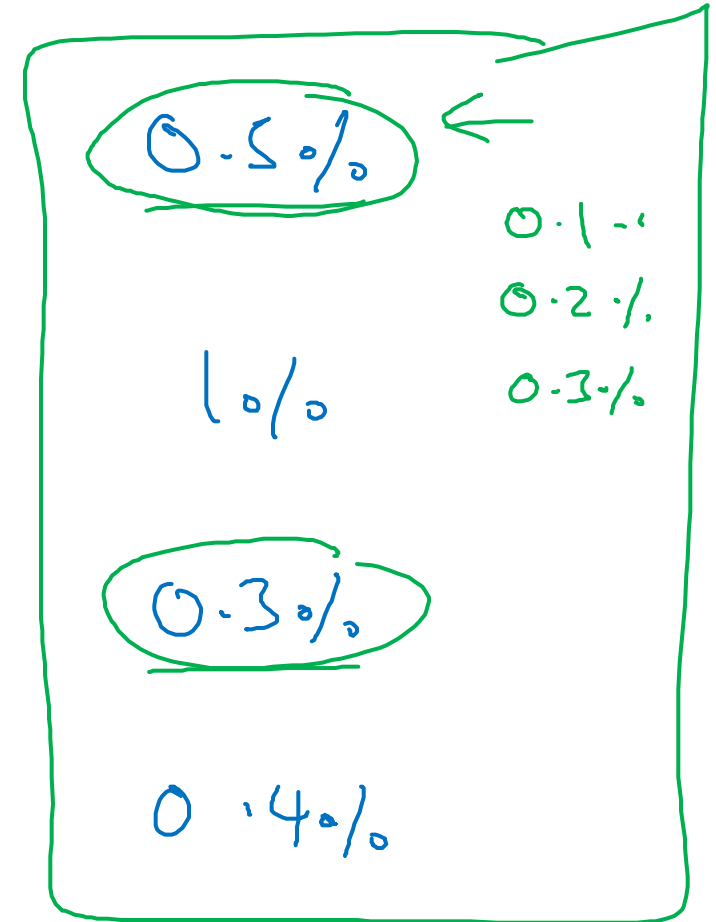
Training error

0.6%

Dev error

0.2  $\updownarrow$  0.8%

What is avoidable bias?



# Problems where ML significantly surpasses human-level performance

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

Structured data

Not natural perception

Lots of data

- Speech recognition
- Some image recognition
- Medical
  - ECG, Skin cancer, ...



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.



~ Avoidable bias

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



~ Variance

# Reducing (avoidable) bias and variance

Human-level



Avoidable bias

Training error



Variance

Dev error

Train bigger model

Train longer/better optimization algorithms

- momentum, RMSprop, Adam

NN architecture/hyperparameters search

RNN  
CNN

More data

Regularization

-  $L_2$ , dropout, data augmentation

NN architecture/hyperparameters search