deeplearning.ai

Introduction to ML strategy

Why ML Strategy?

Motivating example

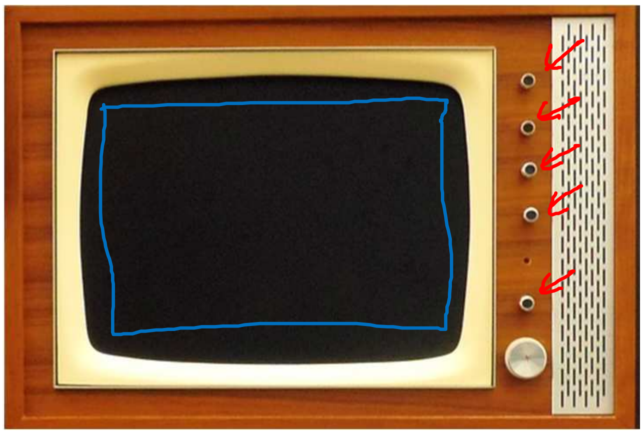
     

90%

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
    - …

TV tuning example



Orthogonalization

$0.1 \times$

$+ 0.3 \times$

$- 1.7 \times$

$+ 0.8 \times$

$+ \ldots$

Car

→ Steering

→ Acceleration

Braking

→ $0.3 \times$ angle $-$ 0.8 speed

→ $2 \times$ angle $+$ 0.9 speed.

speed

angle

Andrew Ng

Chain of assumptions in ML

→ Fit training set well on cost function

(≈ human-level performance)

bigger network
Adam
...

width

→ Fit dev set well on cost function

early stopping

Regularization
Bigger train set

height

→ Fit test set well on cost function

Bigger dev set

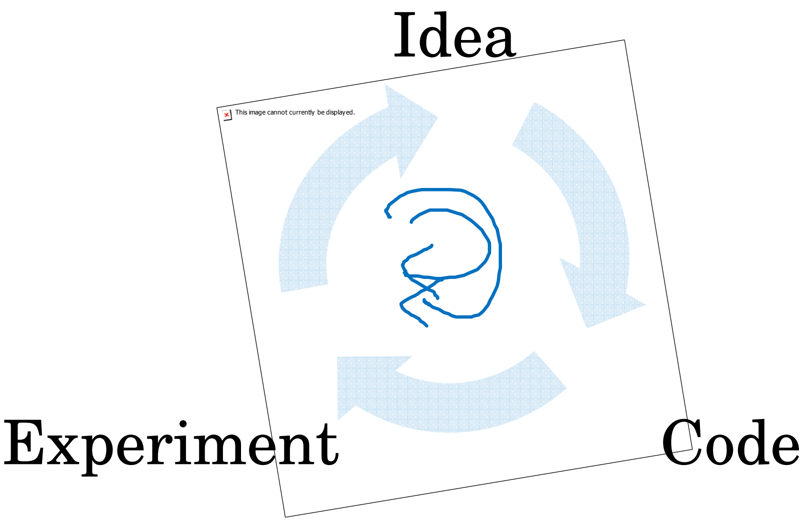
→ Performs well in real world

(Happy cat pic app users.)

Change dev set or cost function

Andrew Ng

Using a single number evaluation metric



→ Of examples recognized as cat, what % actually are cats?

→ what % of actual cats are correctly recognized

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

F_1 score = "Average" of P and R.

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

Dev set + Single number evaluation metric

real Speed up iterating

Another example

Algorithm	US	China	India	Other
A	3%	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%

Cat classification dev/test sets

development set, hold out cross validation set

Regions:

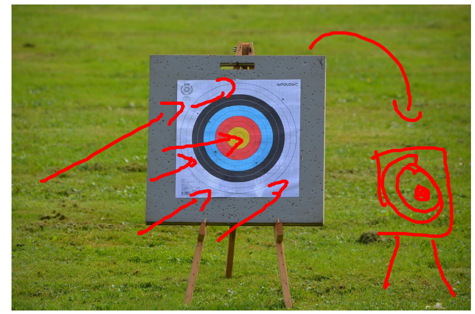
- US
- UK
- Other Europe
- South America

Dev

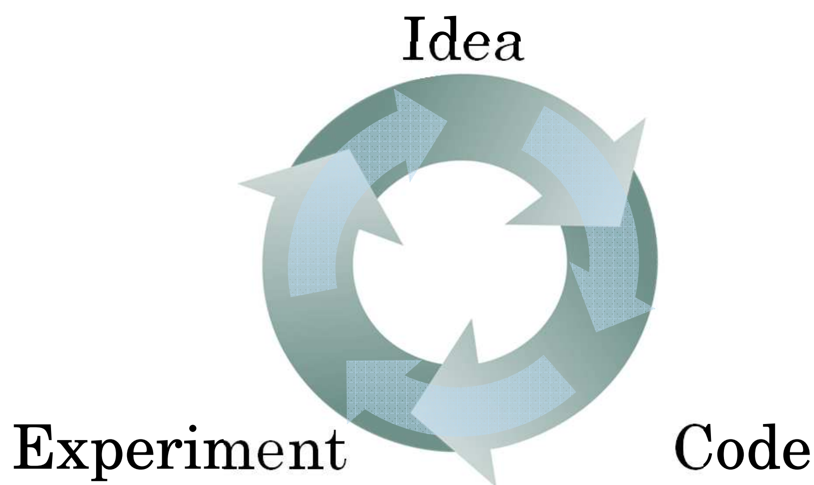
- India
- China
- Other Asia
- Australia

Test

Randomly shuffle into dev/test



dev set + metric

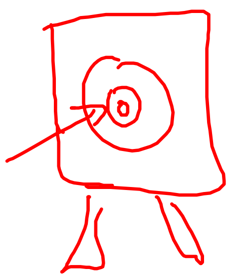


Andrew Ng

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

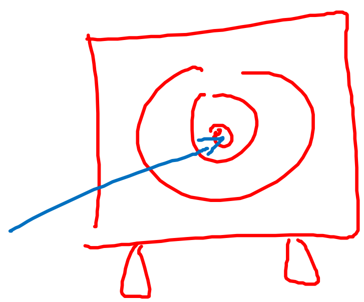
~ 3 month

Guideline

Same distribution

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



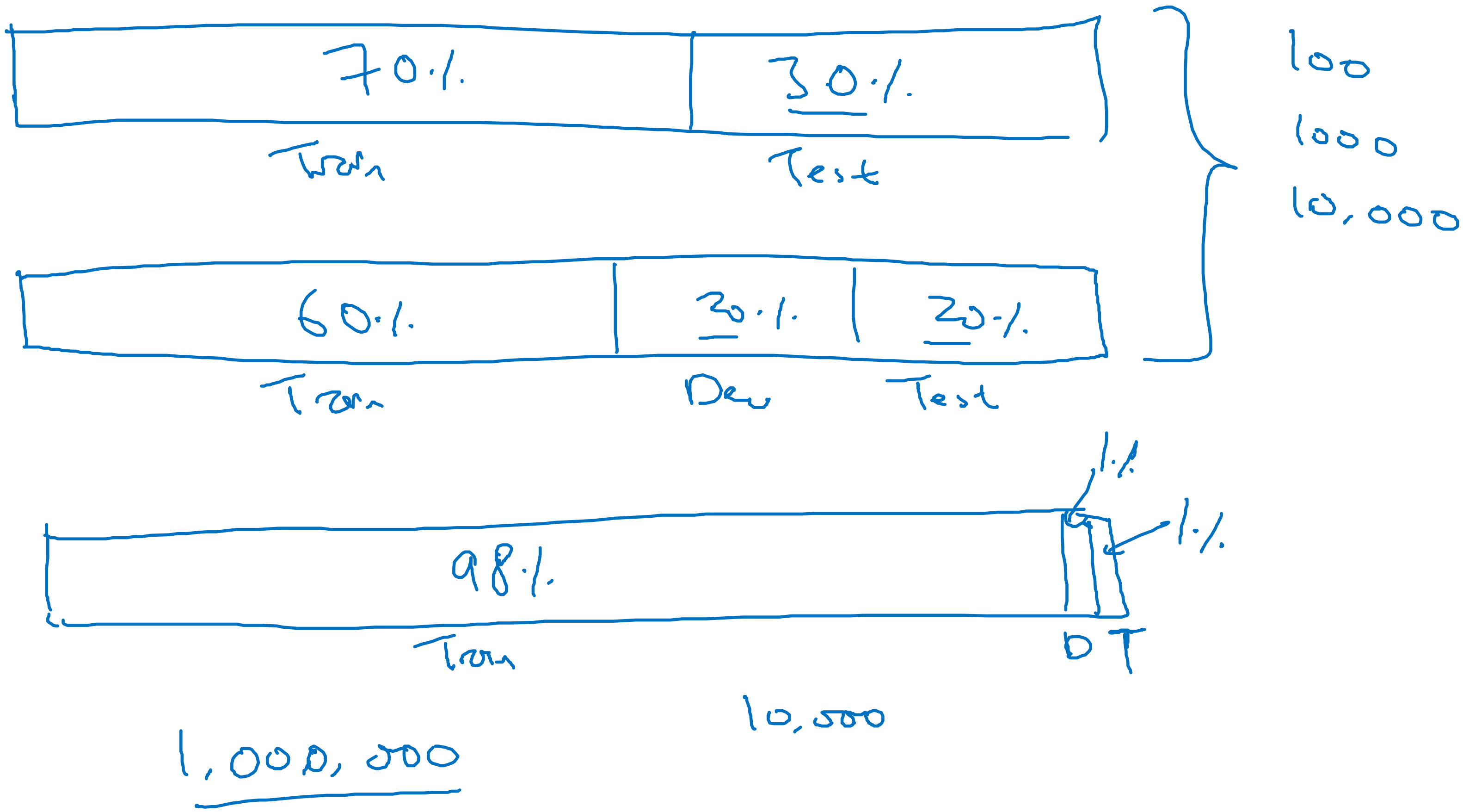
dev

metric

test

Andrew Ng

Old way of splitting data



Andrew Ng

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%

A: 97% → B: 97.1%

0.1%

1,000

10,000

100,000

0.01%

0.001%

Online advertising

Andrew Ng

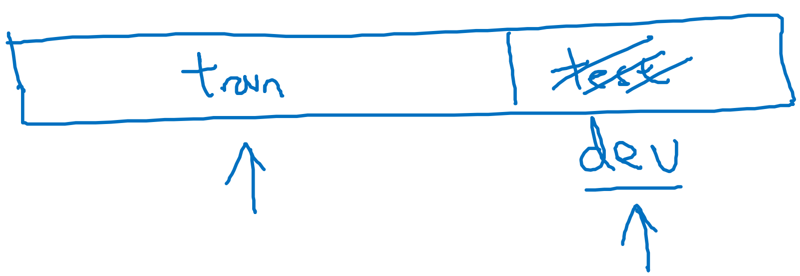
Size of test set

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

10,000

100,000

Train + Dev. No test set.



Andrew Ng

Cat dataset examples

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

- Metric: classification error

  Algorithm A: 3% error → pornographic

  Algorithm B: 5% error

Error: $\frac{1}{\sum_i w^{(i)}} \cancel{\frac{1}{m_{dev}}} \sum_{i=1}^{m_{dev}} w^{(i)} \, I\{ \underbrace{y_{pred}^{(i)}}_{\text{predicted value (0/1)}} \neq y^{(i)} \}$

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

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

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

Another example

Algorithm A: 3% error

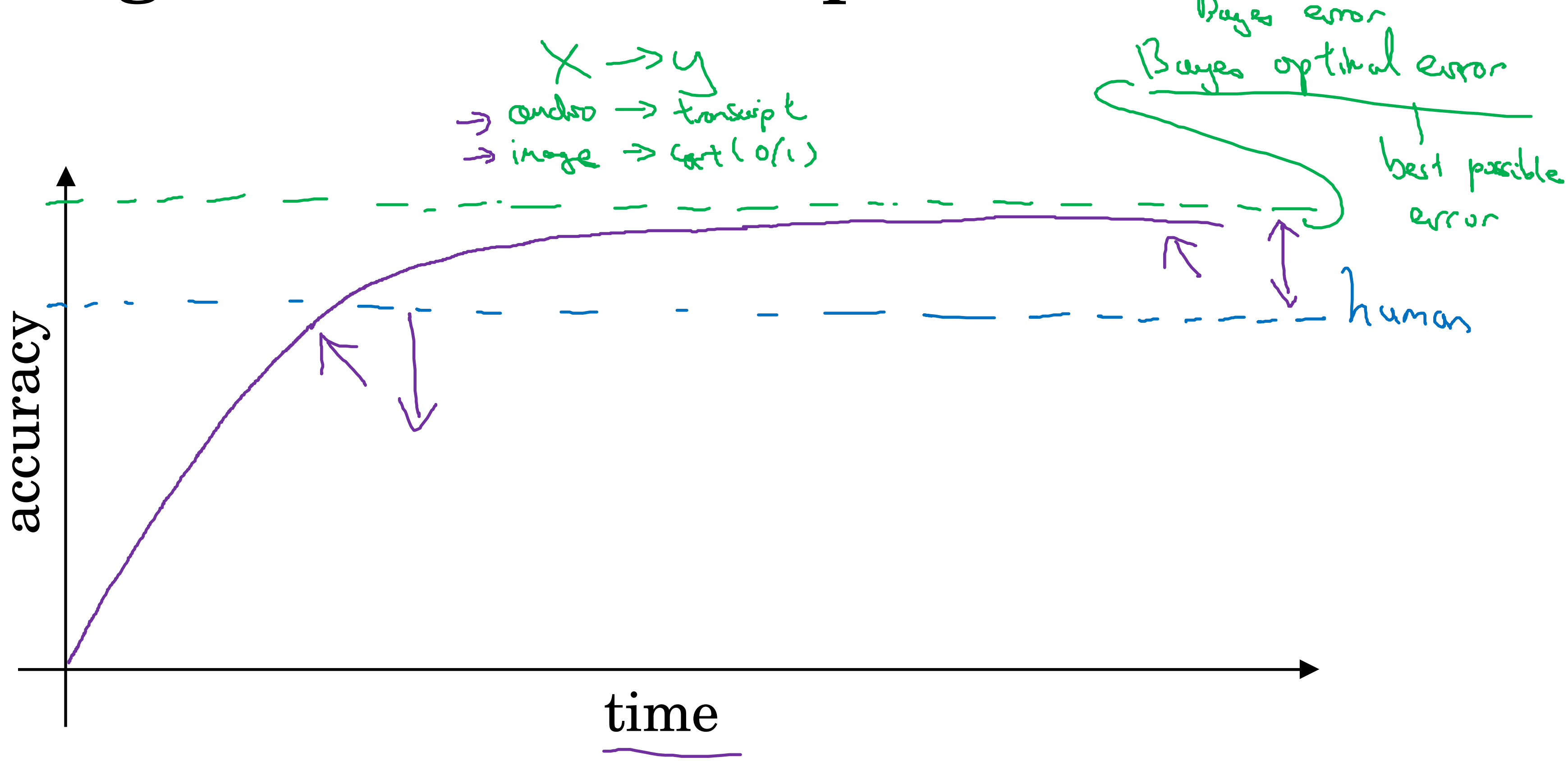
Algorithm B: 5% error

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.

Comparing to human-level performance



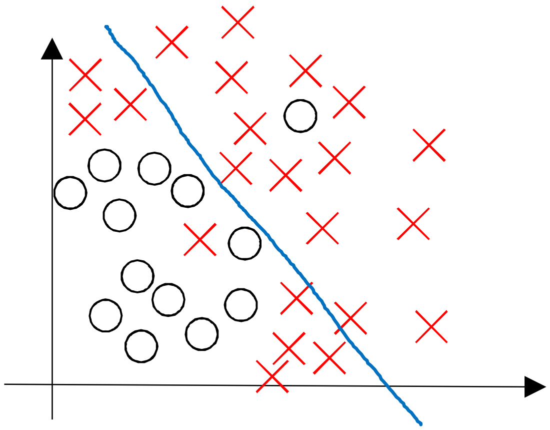
Andrew Ng

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:

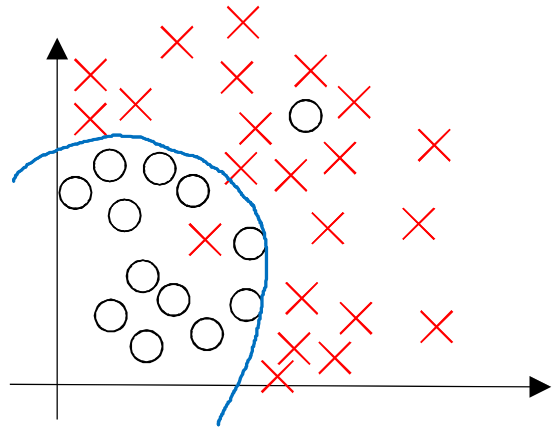
- Get labeled data from humans. (x, y)
- Gain insight from manual error analysis:
  Why did a person get this right?
- Better analysis of bias/variance.

Bias and Variance

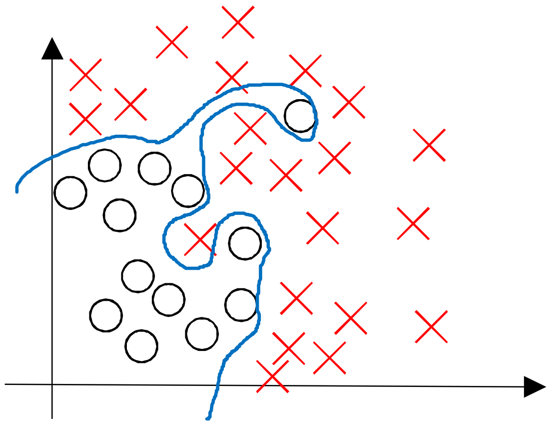


high bias

underfitting



"just right"



high variance

overfitting.

Andrew Ng

Bias and Variance

Cat classification

$\boxed{\text{Human-level} \approx 0\%}$ . . . .

Training set error:

Dev set error:

high variance high bias high bias high variance low bias low variance

Cat classification example

Humans (≈ Bayes)	1%	7.5%
Training error	8%	8%
Dev error	10%	10%
	Focus on bias	Focus on variance

7%

2%

0.5%

Avoidable bias

2%

Variance

2%

Variance

Human-level error as a proxy for Bayes error.

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?

Andrew Ng

Error analysis example

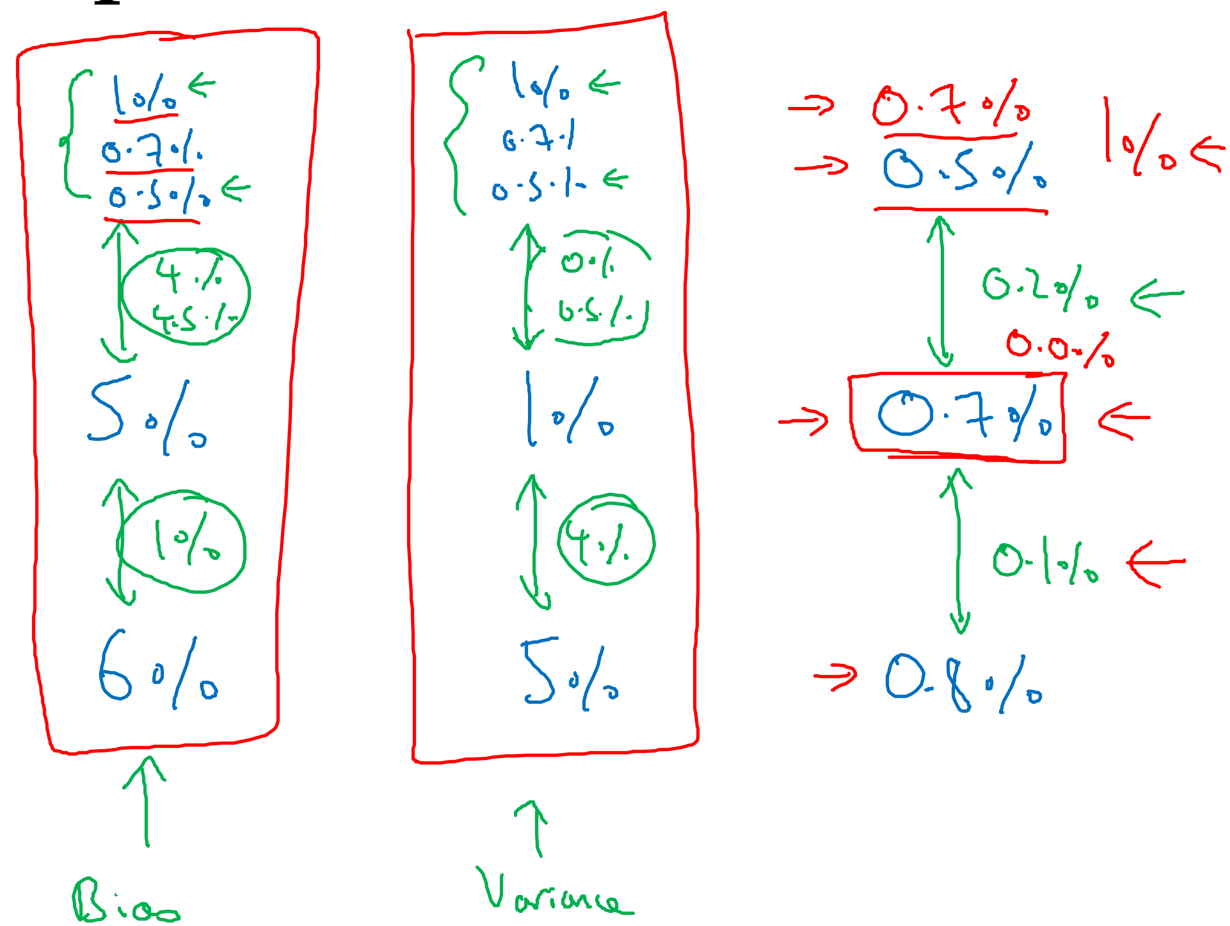
Human (proxy for Bayes error)

Avoidable bias

Training error

Variance

Dev error



Andrew Ng

Summary of bias/variance with human-level performance

0%

"Bias"

Human-level error

(proxy for Bayes error)

"Avoidable bias"

Training error

"Variance"

Dev error

Surpassing human-level performance

Team of humans	0.5%	0.5%
One human	~~1%~~	1%
Training error	0.6%	0.3%
Dev error	0.8%	0.4%

0.1

0.2

0.1%

0.2%

0.3%

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, ...

Andrew Ng

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

Andrew Ng