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
    - …

Andrew Ng

Chain of assumptions in ML

→ Fit training set well on cost function — bigger network, Adam, ...
(≈ human-level performance)

$\boxed{\text{Width}}$

→ Fit dev set well on cost function — Regularization, Bigger training set

$\boxed{\text{Depth}}$

→ Fit test set well on cost function — Bigger dev set

→ Performs well in real world — Change dev set or
(Happy cat pic app users.)

Using a single number evaluation metric

Idea

Experiment

Code

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

→ What % of actual cats are correctly recognized?

Classifier	Precision	Recall	F1 Score
A	95%	90%	92.4%
B	98%	85%	91.0%

F1 Score = "Average" of P and R.

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

Another example

Algorithm	US	China	India	Other	Average
A	3%	7%	5%	9%	6%
B	5%	6%	5%	10%	6.5%
C	2%	3%	4%	5%	3.5%
D	5%	8%	7%	2%	5.25%
E	4%	5%	2%	4%	3.75%
F	7%	11%	8%	12%	9.5%

How could we combine an optimizing and satisficing metrics?

Another cat classification example

optimizing — Satisficing

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

cost = accuracy − 0.5 × running Time

maximize accuracy

subject to running Time ≤ 100 ms.

Similar examples of wake words/ trigger words

Wakewords / Trigger words

Alexa, OK Google,

Hey Siri, nihaobaidu

你好百度

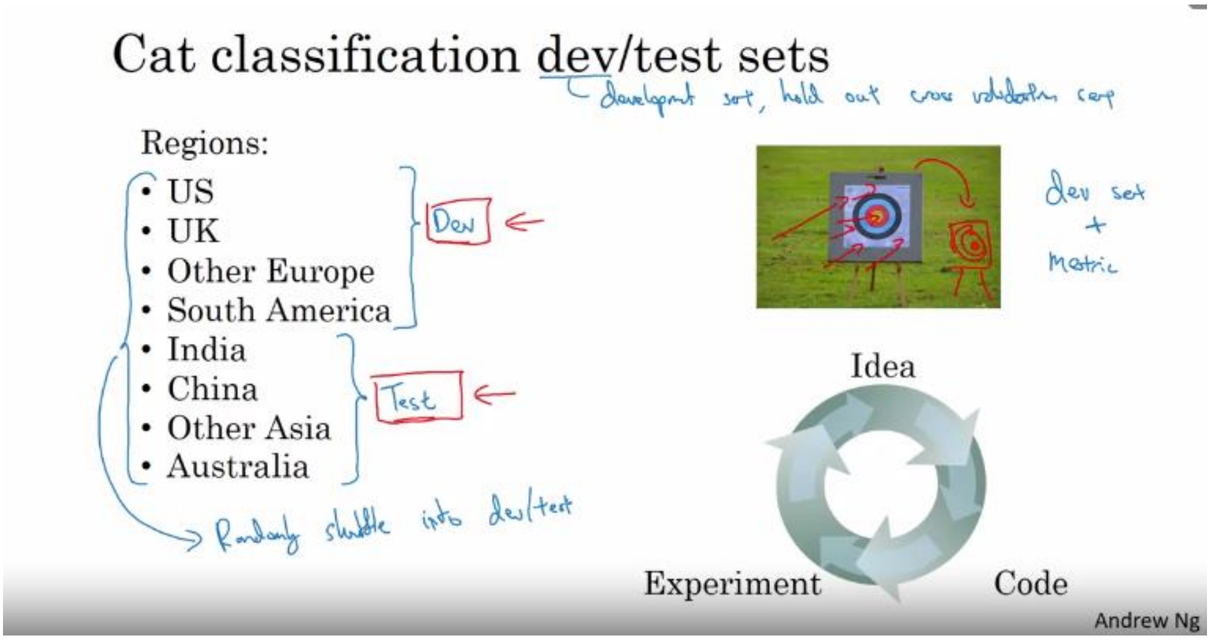
accuracy.

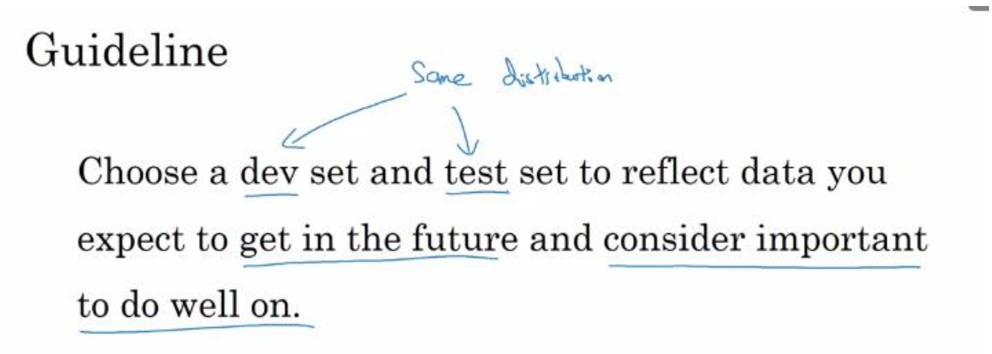
#false positive

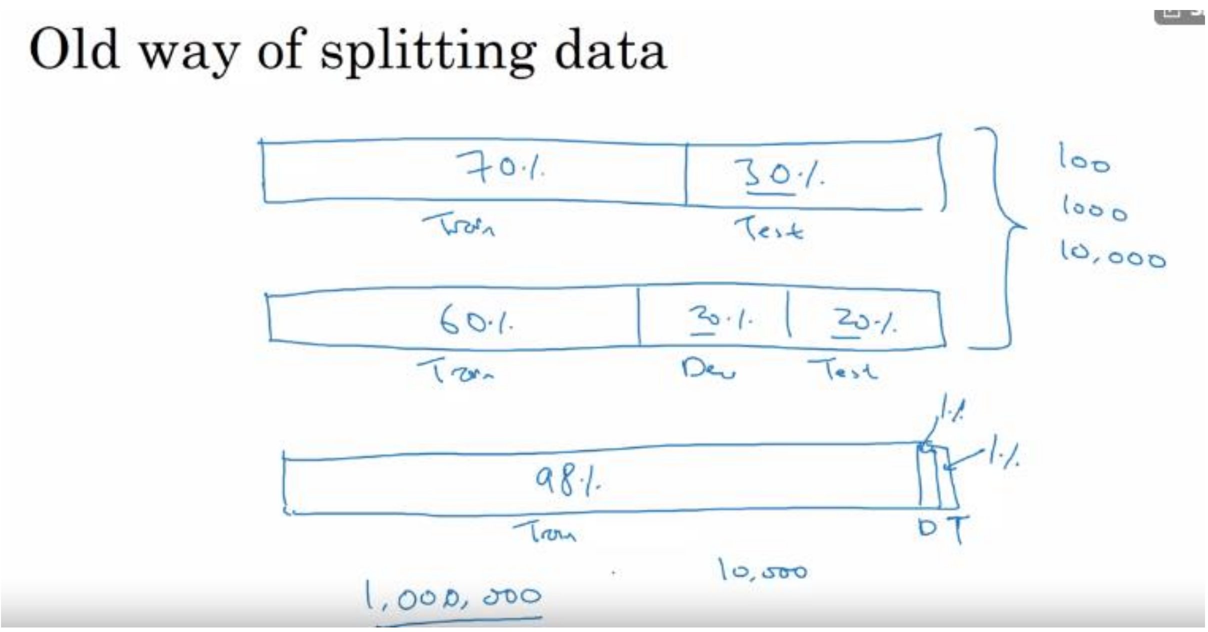
maximize accuracy.

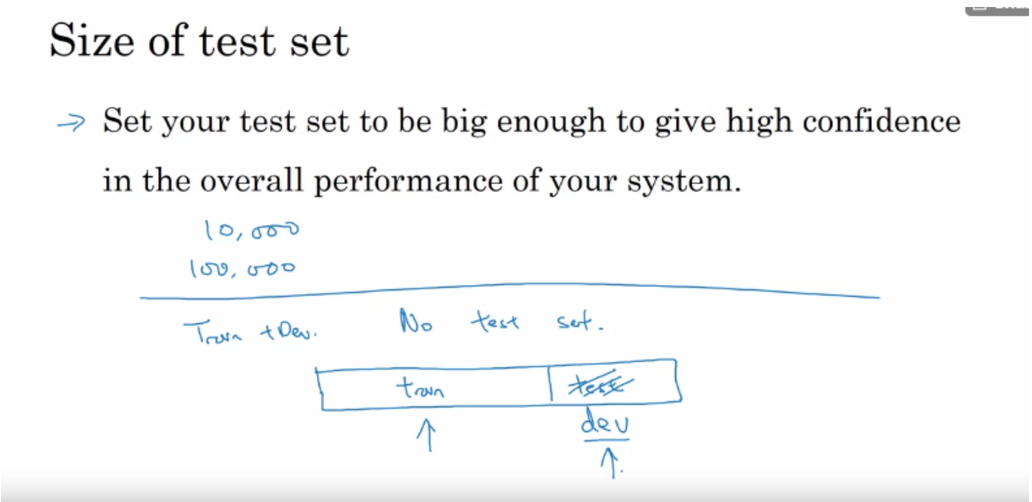
s.t. ≤ 1 false positive every 24 hours.

Train/dev/test distributions



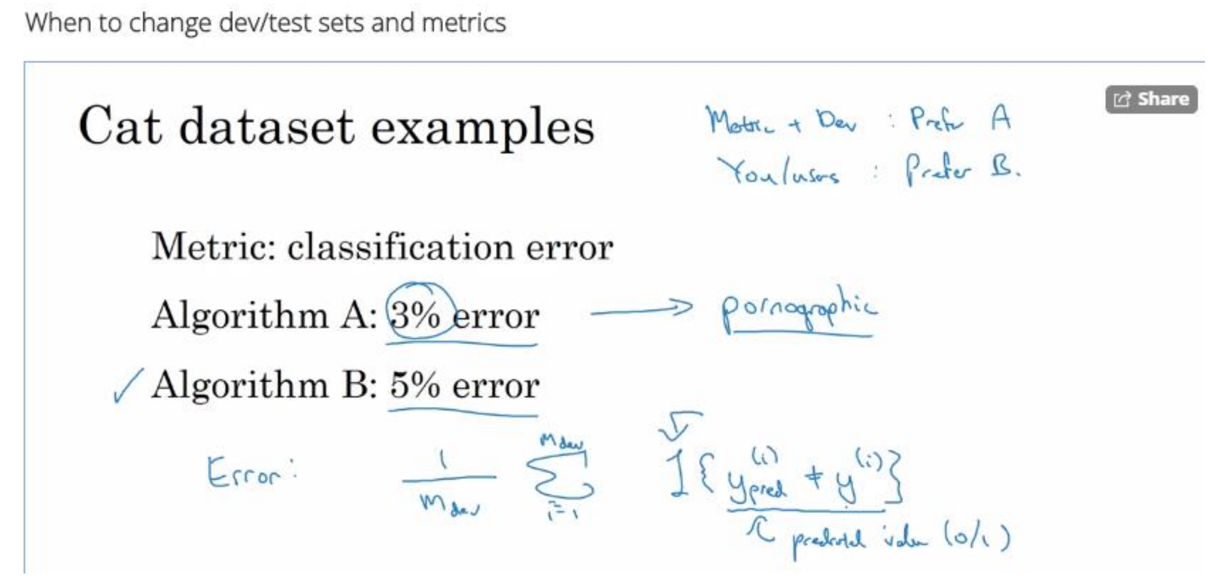




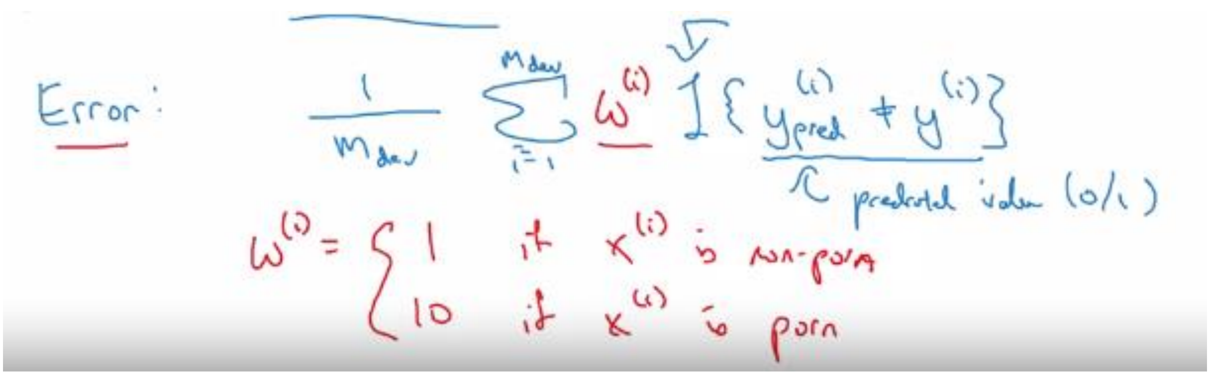


An unconventional option might be to only have train & dev sets, with dev set large enough that you won't overfit the train set.

In the blow, algo A is doing better with lower error, but we may also want to cost it for displaying the occasional pornographic image:



Weigh it with a w_i which will magnify the error by 10 if the image is pornographic



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

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

Another example

Algorithm A: 3% error

✓ Algorithm B: 5% error ←

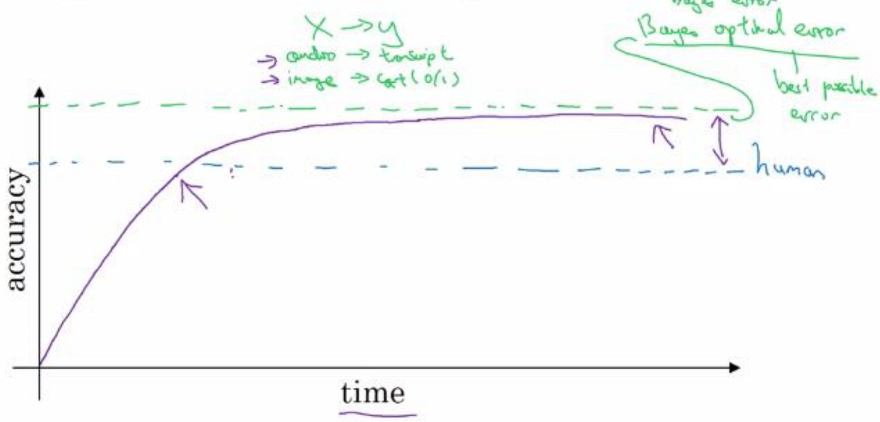
→ Dev/test → User images

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.

Why human-level performance?

Share

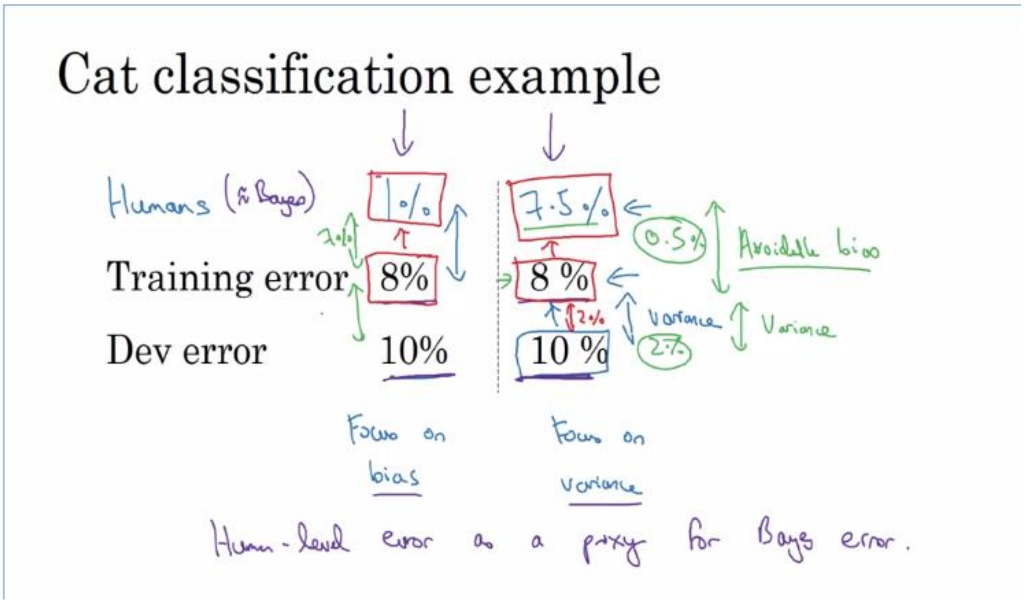
Comparing to human-level performance



Why the slowdoen after surpassing human level performance?

1. For many tasks, human level performance is not that far from Bayes' optimal error. So, by the time you surpass human level performance maybe there's not that much head room to still improve.
2. As long as your performance is worse than human level performance, you can:
    a. Get labeled data from humans
    b. Gain insight from a manual error analysis
    c. Better analysis of bias/variance

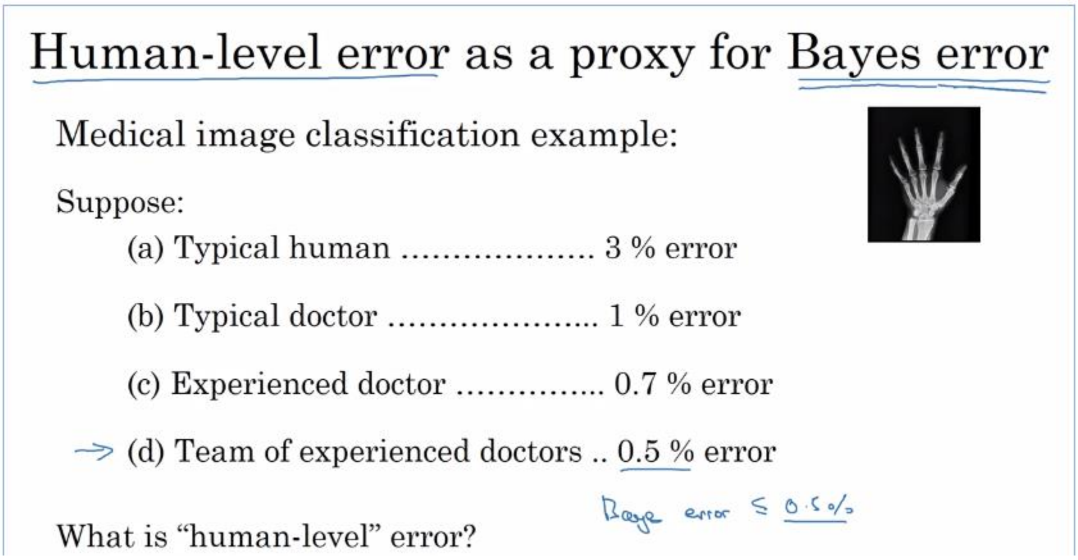
Avoidable bias



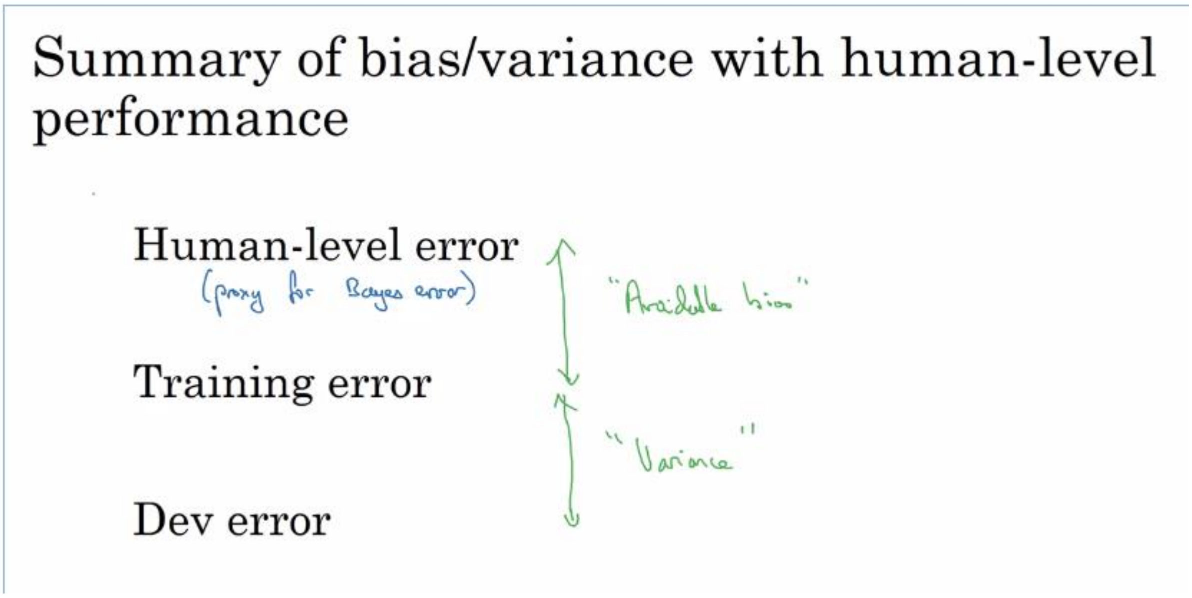
Avoidable bias: Diff between training error and the achievable min error level (Bayes error)

In scenario 2 on the right, much more scope to reduce the 2% variance than reducing the 0.5% avoidable bias.

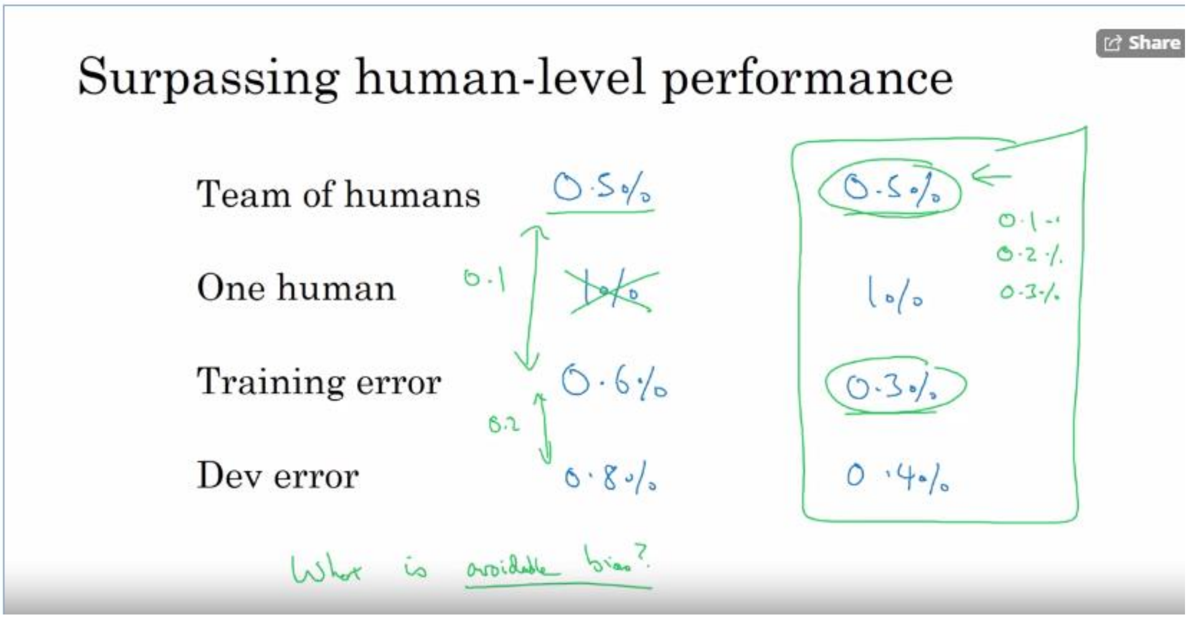
Understanding human-level performance



Understanding human-level performance



Surpassing human-level performance



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, ...

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

R
C

More data

Regularization
- L_2, dropout, data augmentation

NN architecture/hyperparameters search

3:54 / 4:36