Structuring Machine Learning Projects

Machine Learning Strategies

Course 3 Week 2

Maryam Jameela

Agenda

- Error Analysis
- Mismatched training and dev/test set
- Learning from multiple tasks
- End to End Deep learning

Error Analysis

Look at dev examples to evaluate ideas

90% accuracy
→ 10% error

Should you try to make your cat classifier do better on dogs?

Error analysis: → 5-10 min

- Get ~100 mislabeled dev set examples.
- Count up how many are dogs.

"ceiling"

→ 5% 5/100

10% → 9.5%

→ 50% 50/100

10% → 5%

Andrew Ng

Contd..

Evaluate multiple ideas in parallel

Ideas for cat detection:

- Fix pictures of dogs being recognized as cats
- Fix great cats (lions, panthers, etc..) being misrecognized
- Improve performance on blurry images

Image	Dog	Great Cats	Blurry	Instagram	Comments
1	✓			✓	Pitbull
2			✓	✓	
3		✓	✓		Rainy day at zoo
⋮	⋮	⋮	⋮		
% of total	8%	43%	61%	12%	

Andrew Ng

Incorrectly Labeled Data

Incorrectly labeled examples

DL algorithms are quite robust to random errors in the training set.

Systematic errors

Contd..

Error analysis

Image	Dog	Great Cat	Blurry	Incorrectly labeled	Comments
...					
98				✓	Labeler missed cat in background
99		✓			
100				✓	Drawing of a cat; Not a real cat.
% of total	8%	43%	61%	6%	

Overall dev set error ..... 10% | 2%

Errors due incorrect labels ..... 0.6% ← | 0.6%

Errors due to other causes ..... 9.4% ← | 1.4%

2.1% 1.9%

Goal of dev set is to help you select between two classifiers A & B.

Contd..

Correcting incorrect dev/test set examples

- Apply same process to your dev and test sets to make sure they continue to come from the same distribution
- Consider examining examples your algorithm got right as well as ones it got wrong.
- Train and dev/test data may now come from slightly different distributions.

Build your first system quickly and then iterate

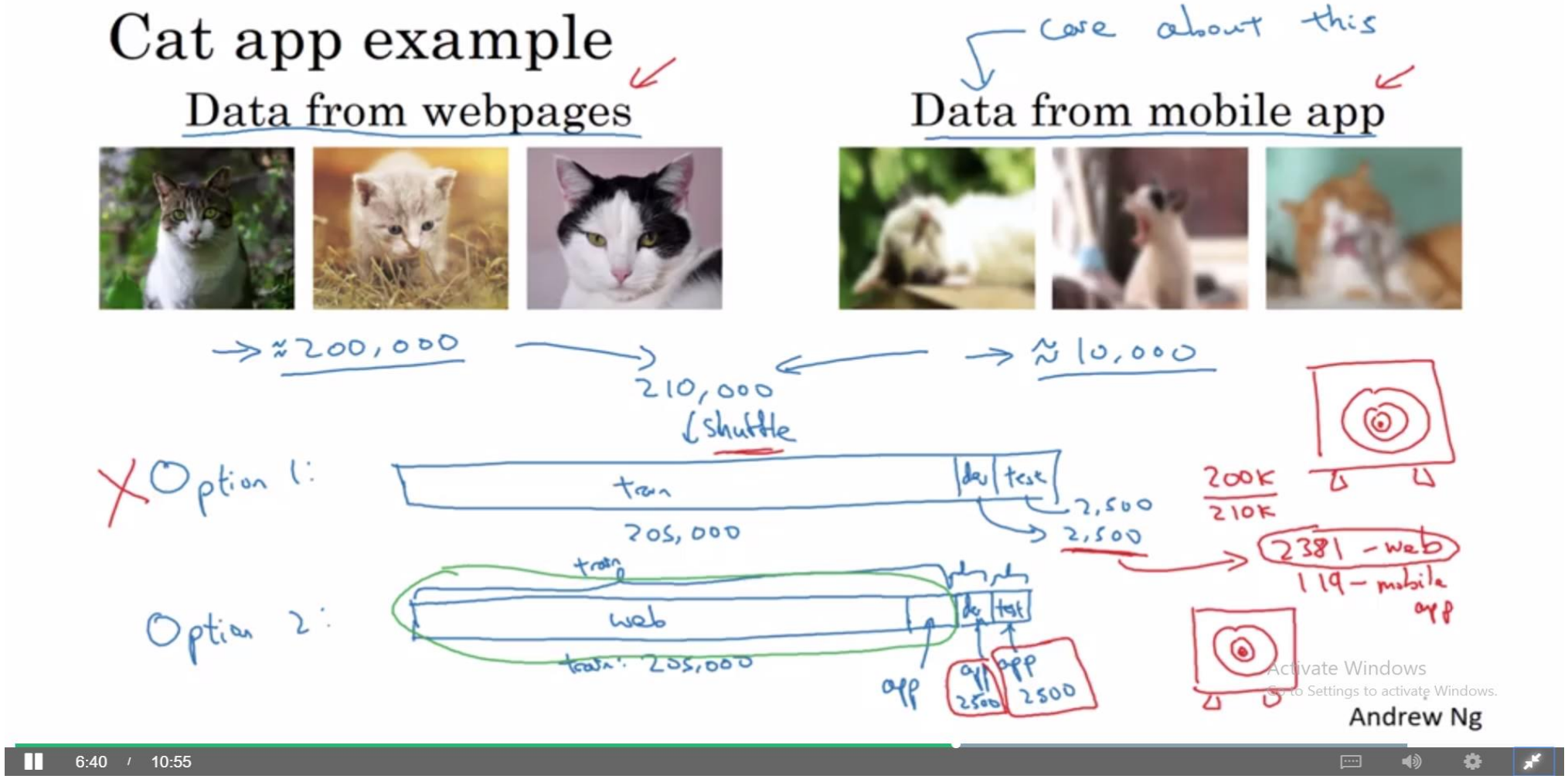
Speech recognition example

- Noisy background
  - Café noise
  - Car noise
- Accented speech
- Far from microphone
- Young children's speech
- Stuttering (uh, ah, um, ...)
- ...

- Set up dev/test set and metric
- Build initial system quickly
- Use Bias/Variance analysis & Error analysis to prioritize next steps.

Andrew Ng

Training and Testing on different distribution..



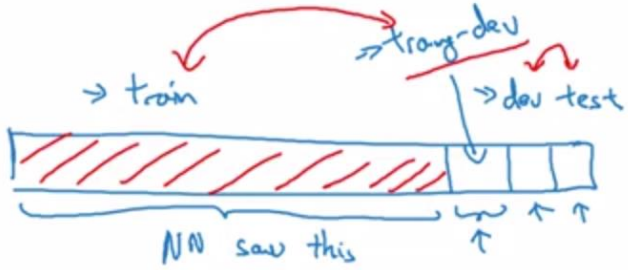
Bias and Variance with mismatched data distribution

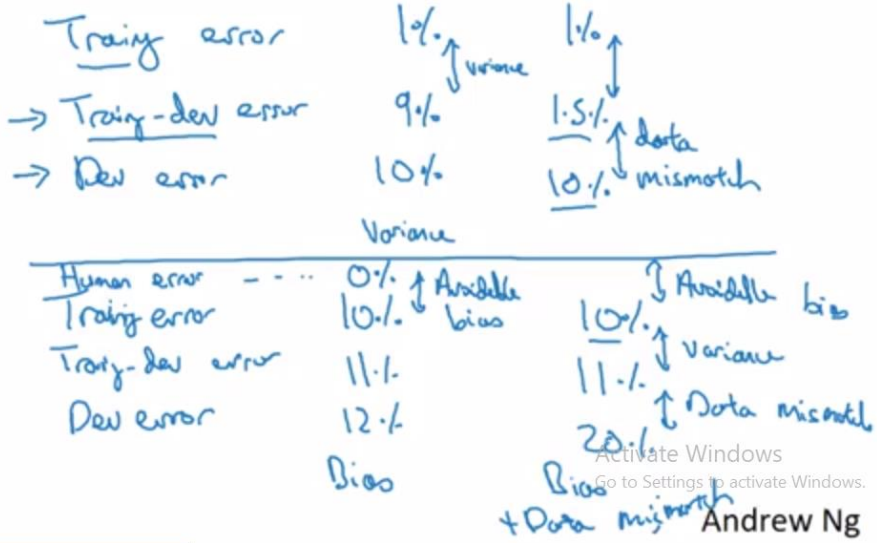
Cat classifier example

Assume humans get ≈ 0% error.

Training error 1%
Dev error 10%

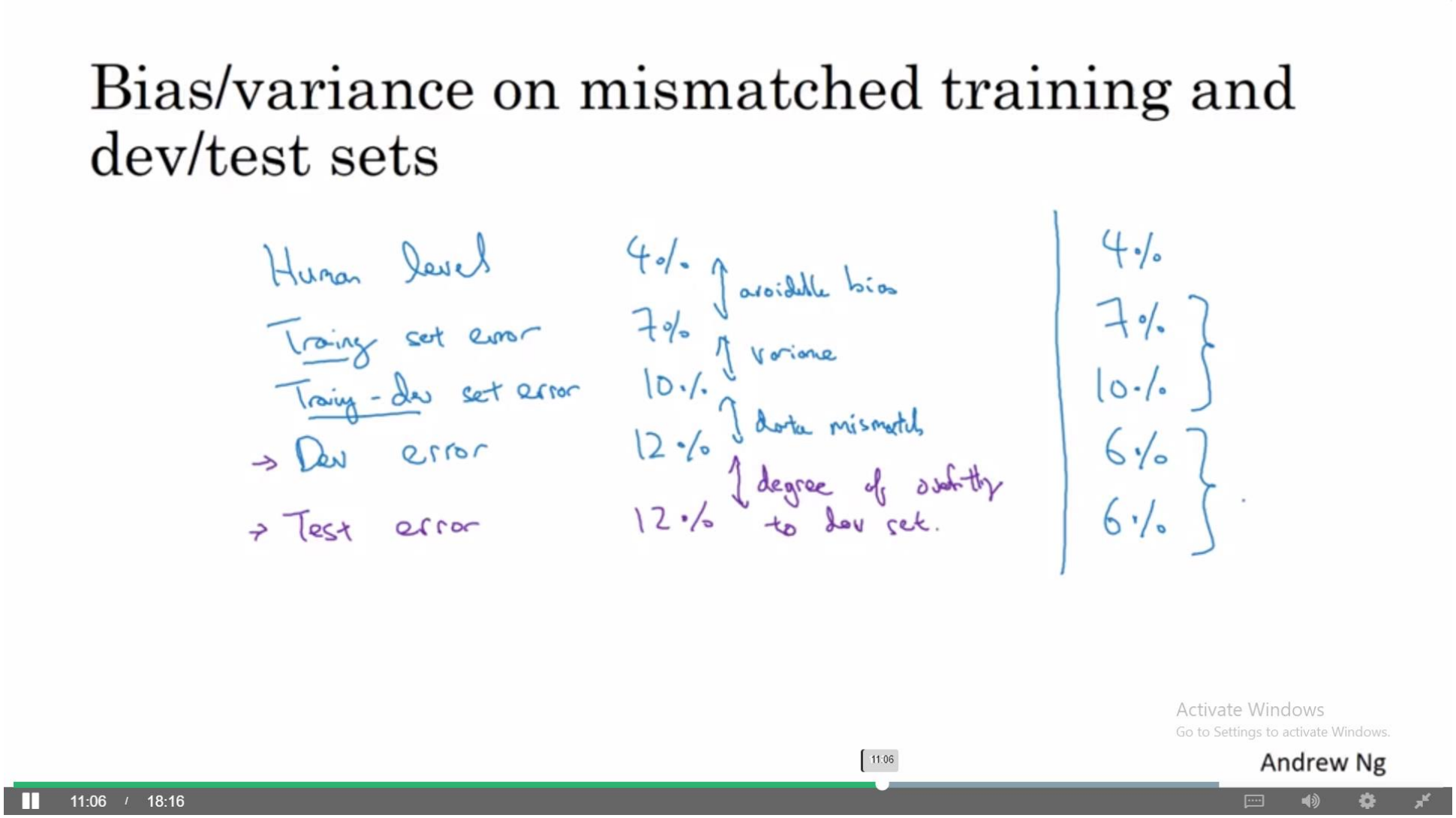
Training-dev set: Same distribution as training set, but not used for training



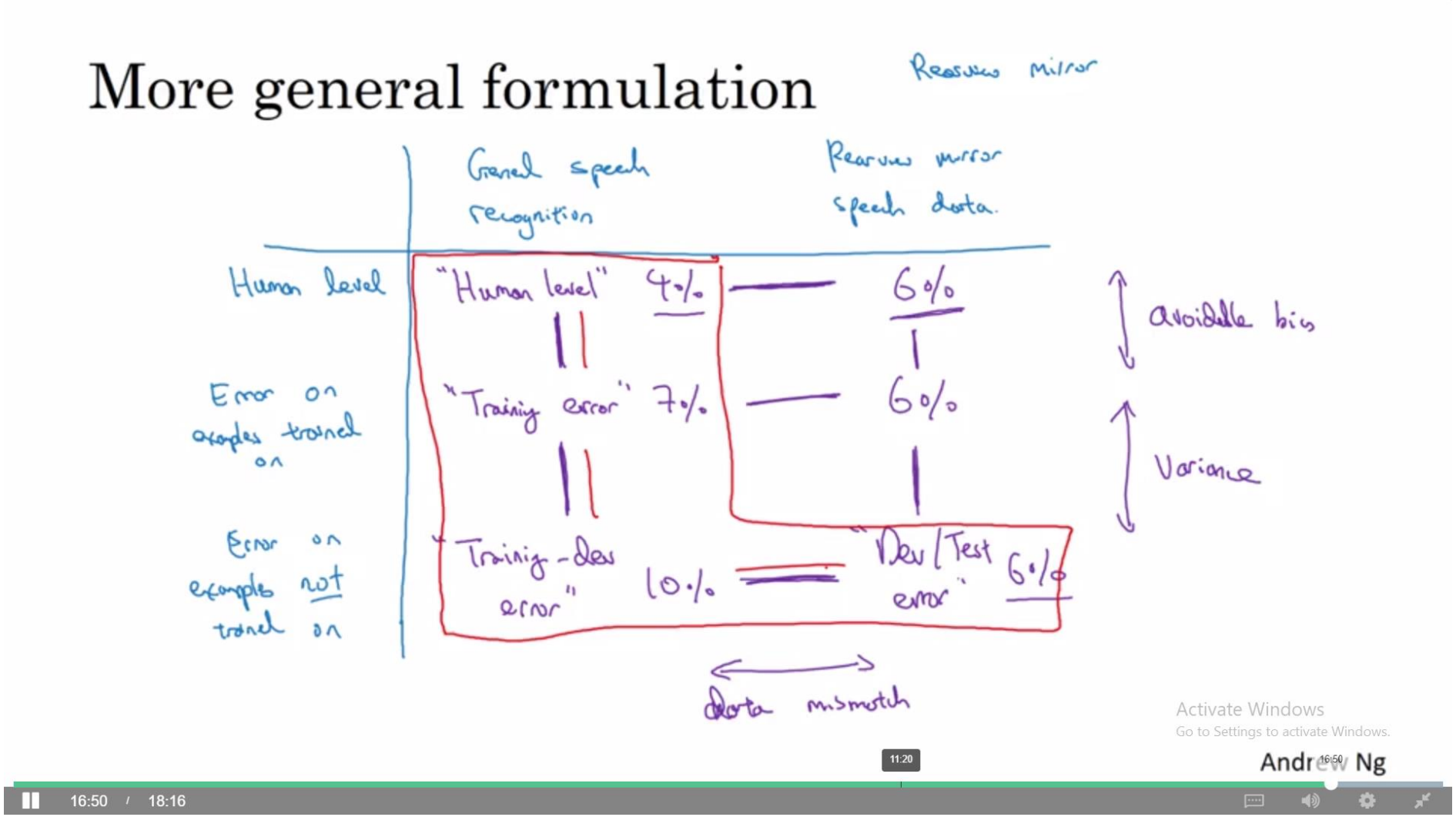


Andrew Ng

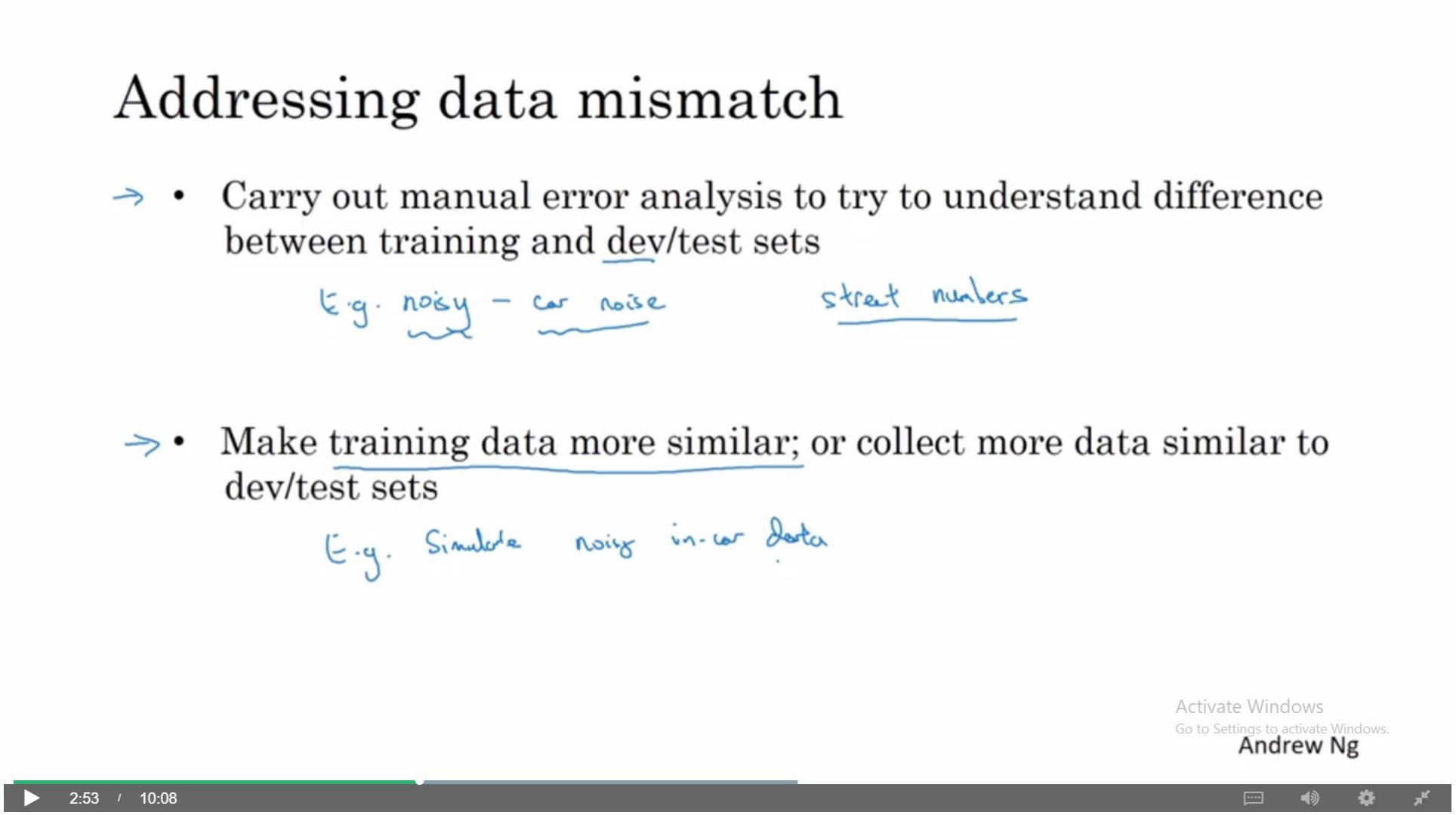
Contd...


Bias/variance on mismatched training and dev/test sets
Human level
Training set error
Training-dev set error
→ Dev error
→ Test error
4%
7%
10%
12%
12%
avoidable bias
variance
data mismatch
degree of overfitting to dev set.
4%
7%
10%
6%
6%
Activate Windows
Go to Settings to activate Windows.
Andrew Ng
11:06
11:06 / 18:16

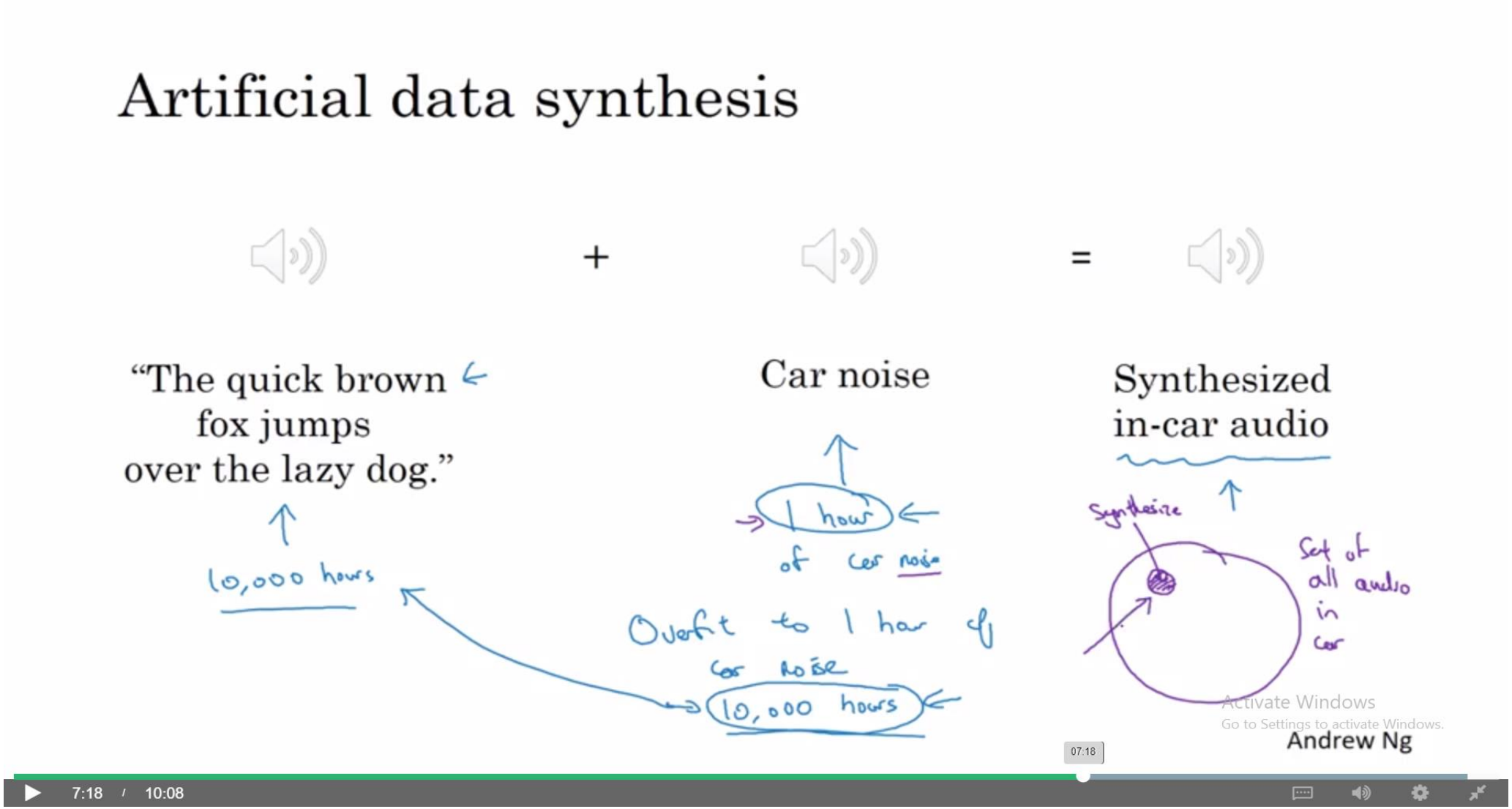
Contd..


More general formulation
Rearview mirror
General speech recognition
Rearview mirror speech data
Human level
"Human level" 4%
6%
avoidable bias
Error on examples trained on
"Training error" 7%
6%
Variance
Error on examples not trained on
"Training-dev error" 10%
"Dev/Test error" 6%
data mismatch
Activate Windows
Go to Settings to activate Windows.
Andrew Ng
11:20
16:50 / 18:16

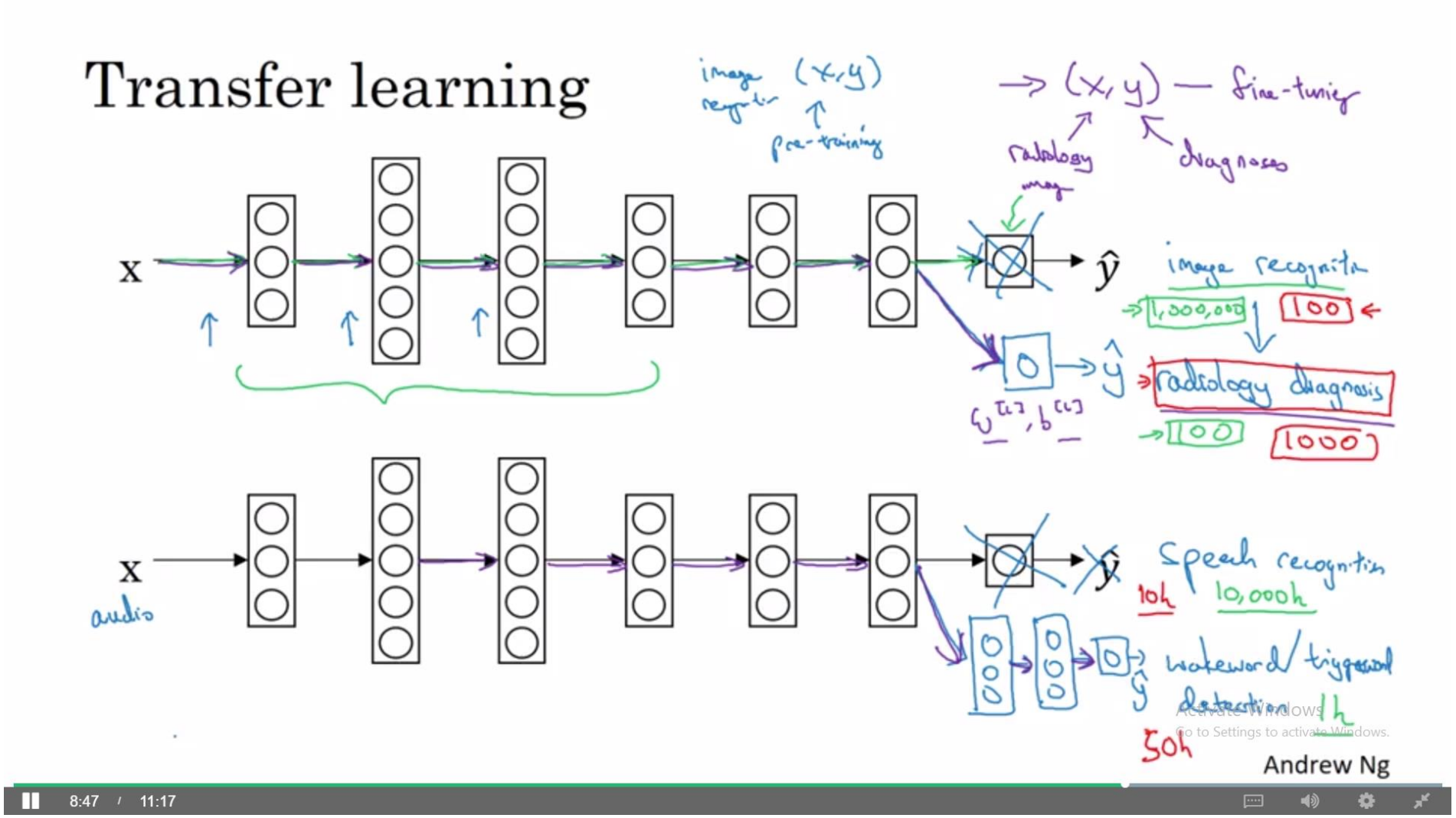
Addressing data mismatched



Contd.


Artificial data synthesis
+
=
"The quick brown fox jumps over the lazy dog."
10,000 hours
Car noise
1 hour
of car noise
Overfit to 1 hour of car noise
10,000 hours
Synthesized in-car audio
Synthesize
Set of all audio in car
Activate Windows
Go to Settings to activate Windows.
Andrew Ng
07:18
7:18 / 10:08

Learning from multiple tasks



Contd..

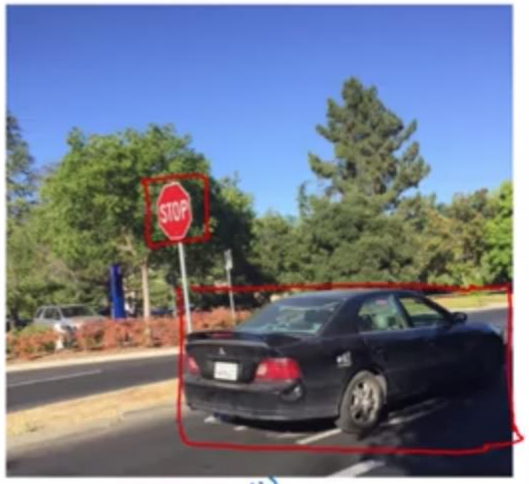
When transfer learning makes sense

Transfer from A → B

- Task A and B have the same input x.
- You have a lot more data for Task A than Task B.
- Low level features from A could be helpful for learning B.

Multitask Learning

Simplified autonomous driving example



$x^{(i)}$

$y^{(i)}$ $(4,1)$

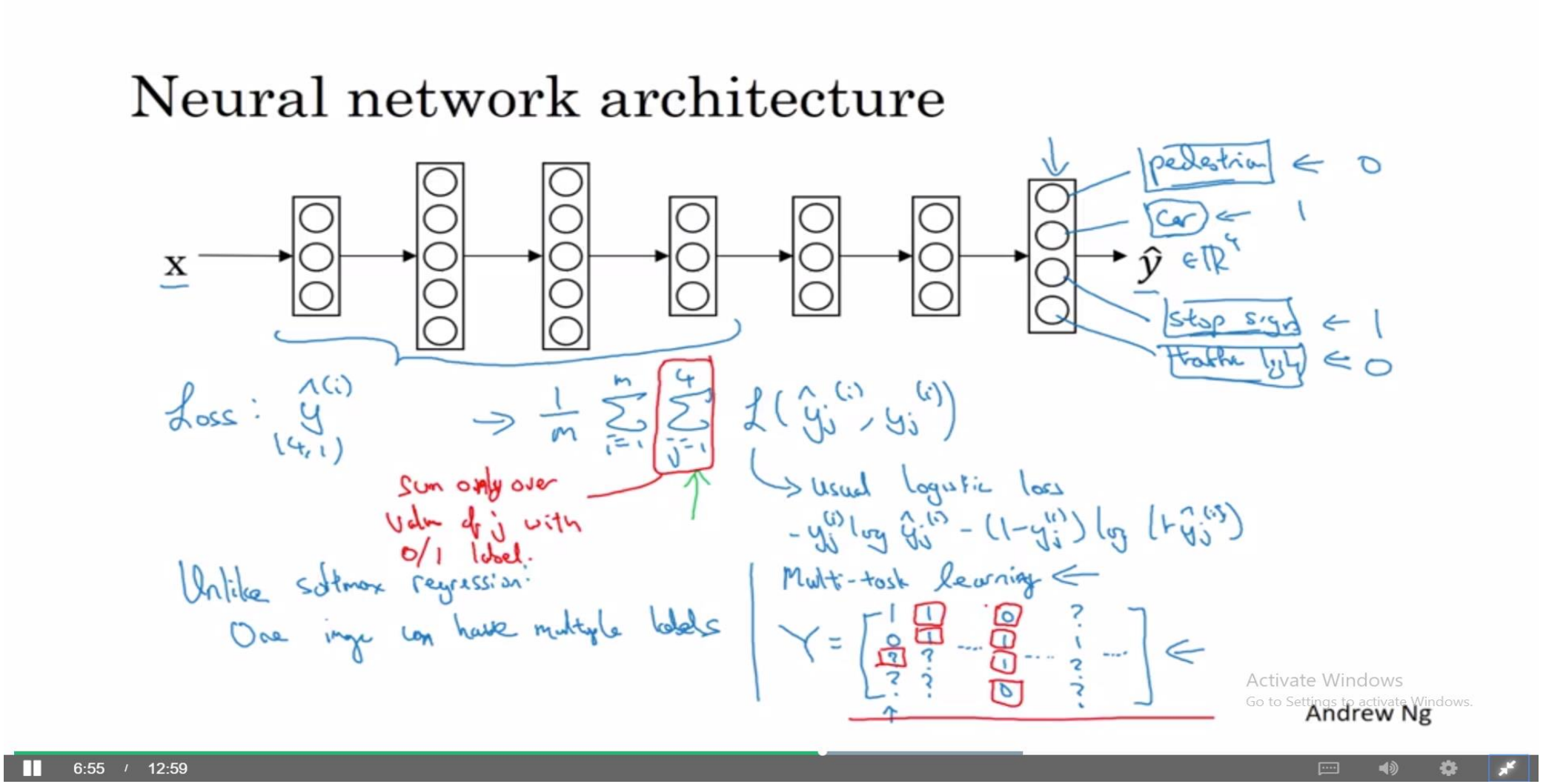
Pedestrians 0
Cars 1
Stop signs 1
Traffic lights 0
⋮ ⋮

$$Y = \begin{bmatrix} | & | & | & & | \\ y^{(1)} & y^{(2)} & y^{(3)} & \cdots & y^{(m)} \\ | & | & | & & | \end{bmatrix}$$

$(4, m)$

Andrew Ng

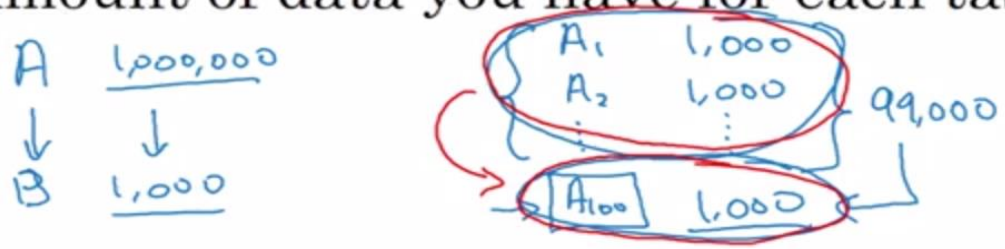
Contd..


Neural network architecture
x
pedestrian ← 0
car ← 1
$\hat{y} \in \mathbb{R}^4$
stop sign ← 1
traffic light ← 0
Loss: $\hat{y}^{(i)}$, $y^{(i)}$, $(4,1)$
$\rightarrow \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{4} \mathcal{L}(\hat{y}_j^{(i)}, y_j^{(i)})$
Sum only over values of j with 0/1 label.
Usual logistic loss
$-y_j^{(i)} \log \hat{y}_j^{(i)} - (1-y_j^{(i)}) \log (1-\hat{y}_j^{(i)})$
Unlike softmax regression:
One image can have multiple labels
Multi-task learning ⇐
$Y = \begin{bmatrix} 1 & 1 & \cdots & 0 & \cdots & ? & \cdots \\ 0 & ? & & 0 & & 1 & \\ ? & ? & & 0 & & ? & \end{bmatrix}$
Activate Windows
Go to Settings to activate Windows.
Andrew Ng
6:55 / 12:59

Contd..

When multi-task learning makes sense

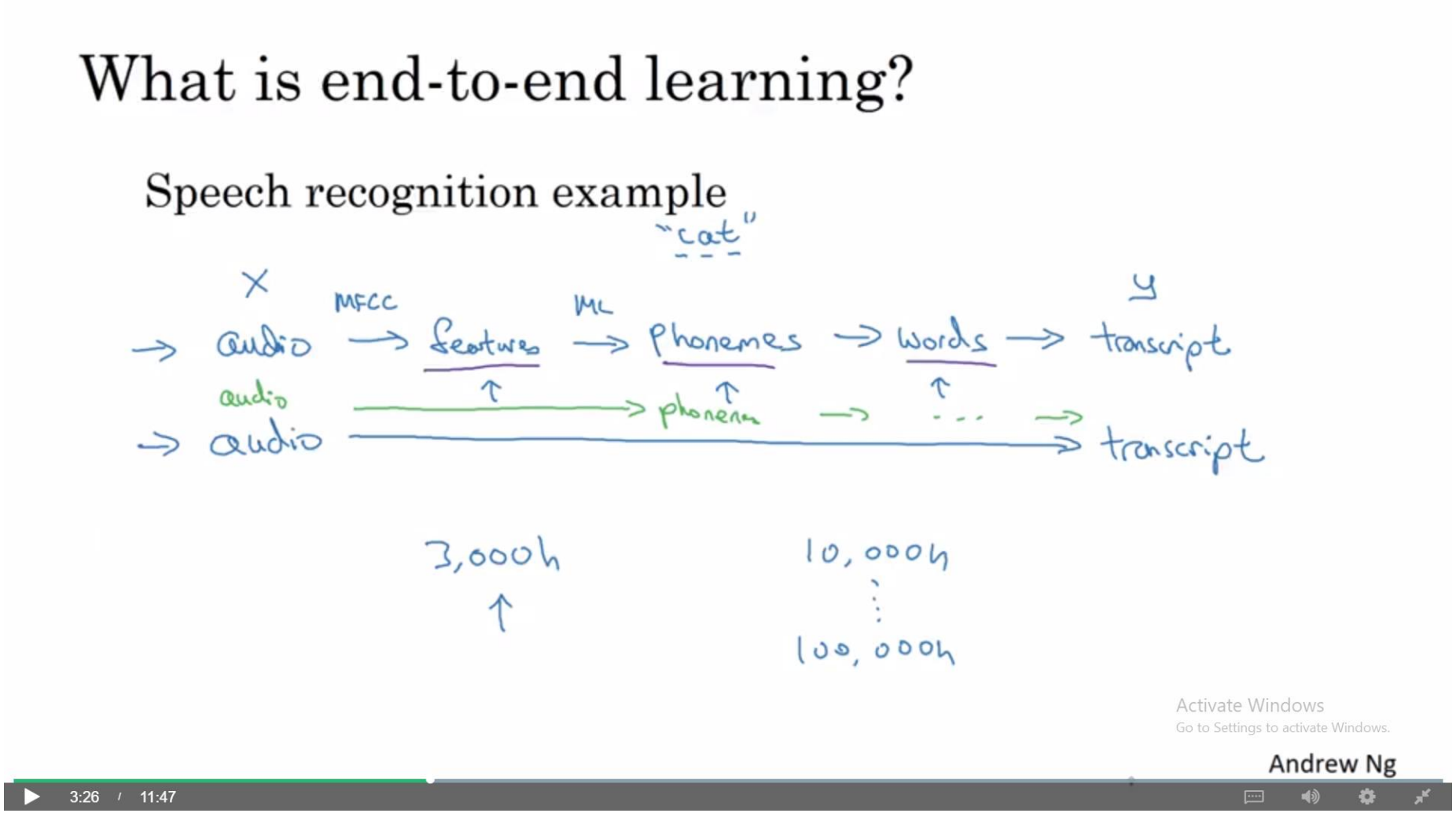
- Training on a set of tasks that could benefit from having shared lower-level features.
- Usually: Amount of data you have for each task is quite similar.



- Can train a big enough neural network to do well on all the tasks.

Andrew Ng

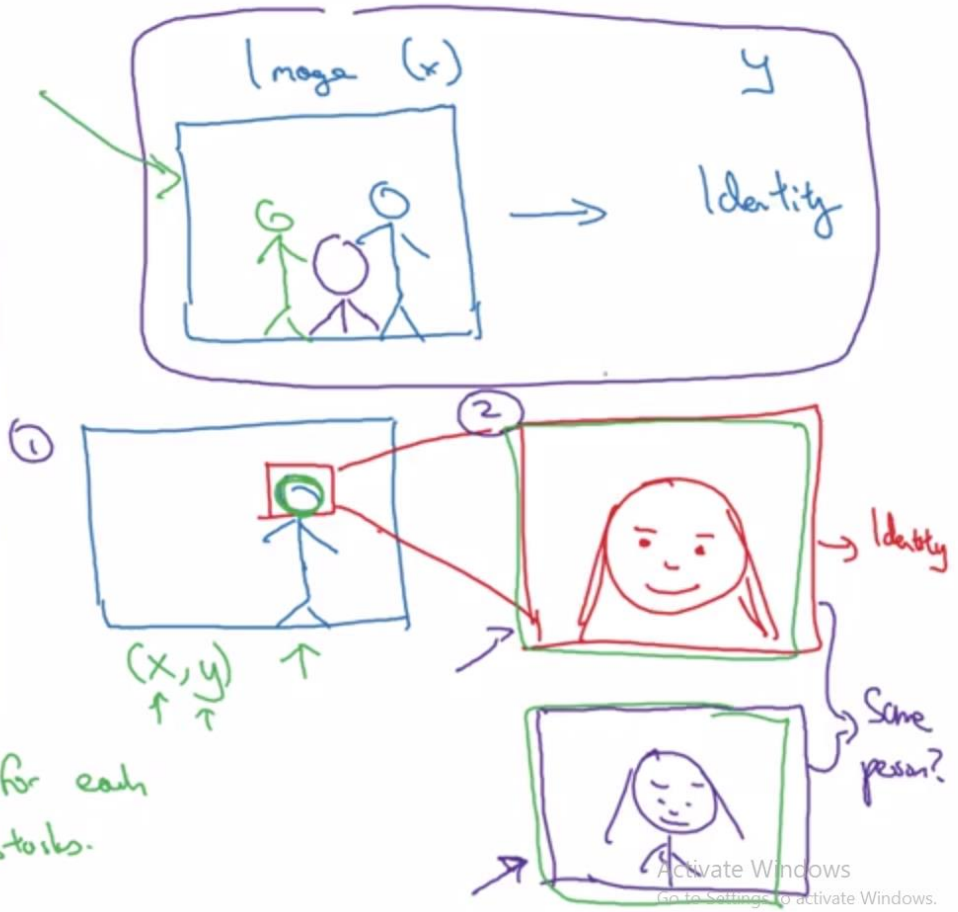
End to End Learning



Contd..

Face recognition

[Image courtesy of Baidu]



Andrew Ng

Pros and Cons..