

# The Data-centric AI Approach

Andrew Ng



# Shifting from model-centric to data-centric AI

Conventional model-centric approach:

$$\text{AI} = \text{Code} + \text{Data}$$

(algorithm/model)

Work on this

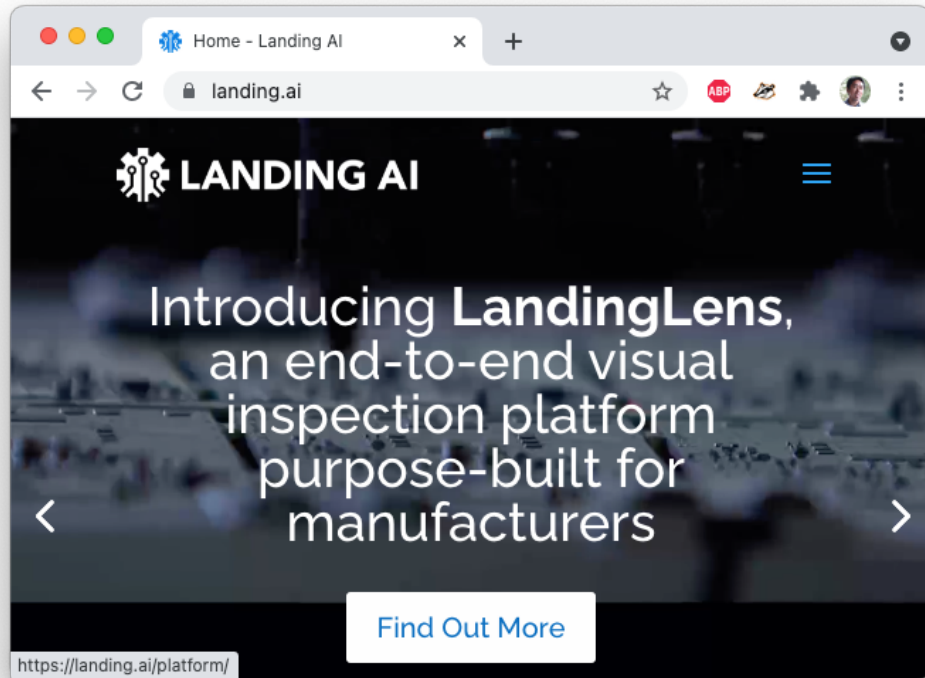
Data-centric approach:

$$\text{AI} = \text{Code} + \text{Data}$$

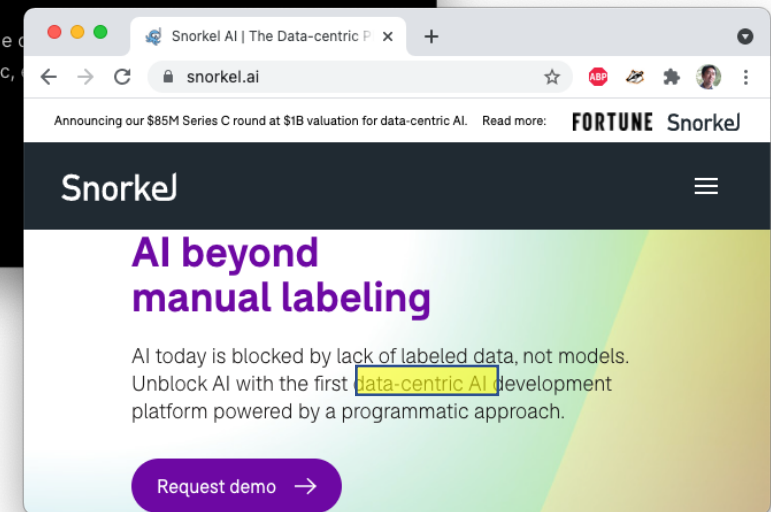
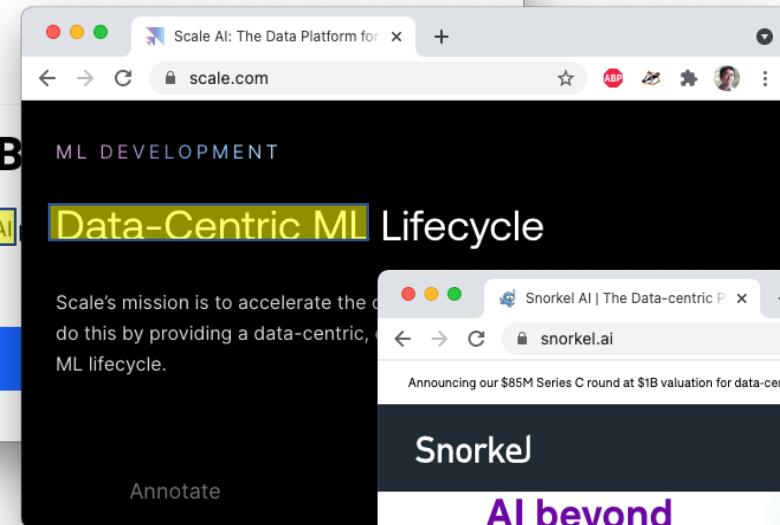
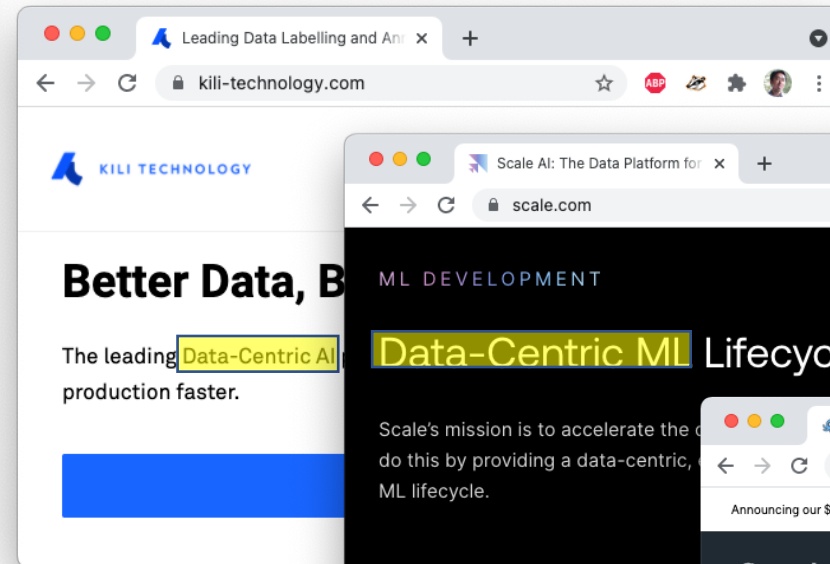
(algorithm/model)

Work on this

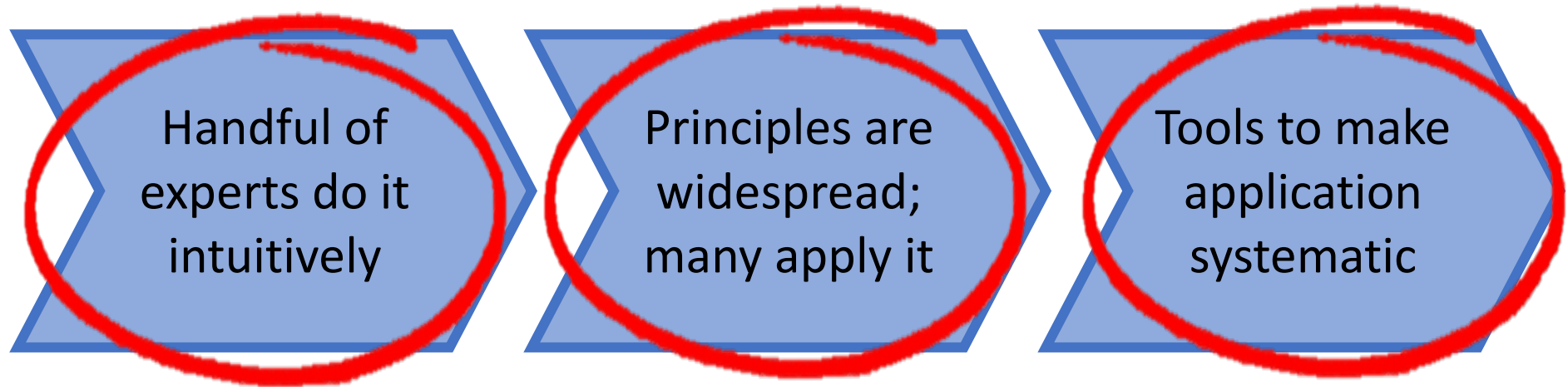
# Data-centric AI movement growth in 6 months



Landing AI: Data-centric MLOps platform for computer vision



# Evolution of new technology approach



# Tips for Data-centric AI development

Tip 1: Make the labels  $y$  consistent

Tip 2: Use consensus labeling to spot inconsistencies

Tip 3: Clarify labeling instructions by tracking down ambiguous examples

Tip 4: Toss out noisy examples. More data is not always better!

Tip 5: Use error analysis to focus on subset of data to improve

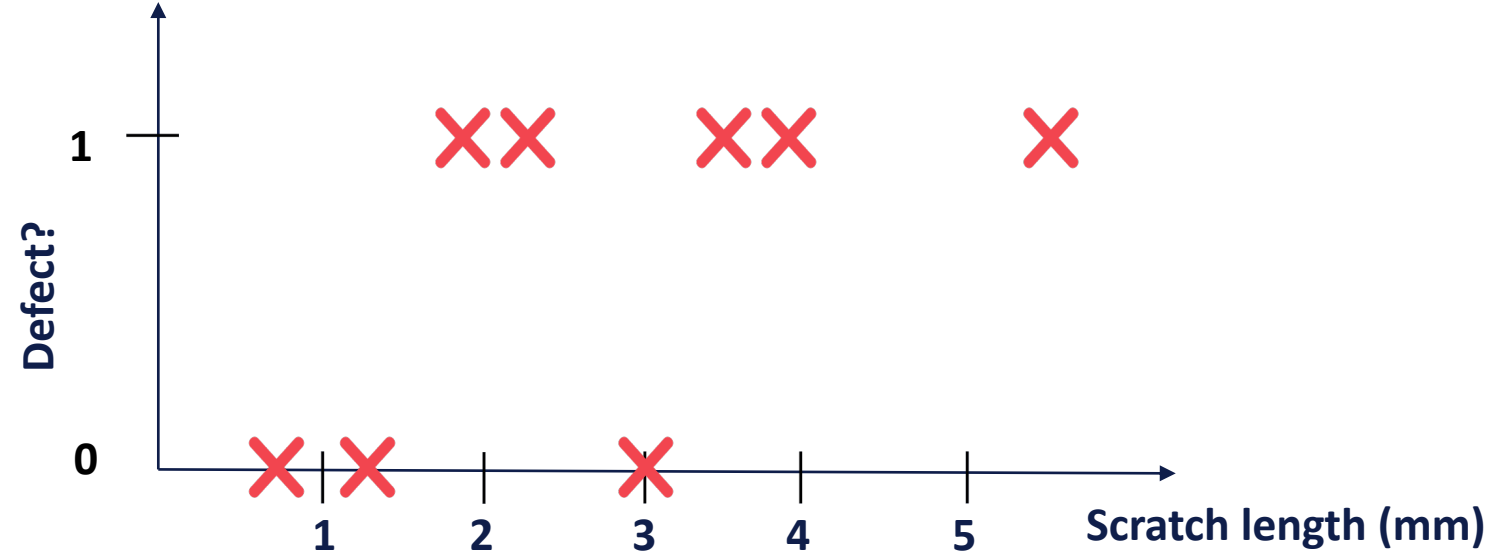
# Tip 1: Make the labels $y$ consistent

**Ideal:** There is some deterministic (non-random) function mapping from  $x \rightarrow y$ , and the labels are consistent with this function.



# Tip 1: Make the labels $y$ consistent

**Ideal:** There is some deterministic (non-random) function mapping from  $x \rightarrow y$ , and the labels are consistent with this function.

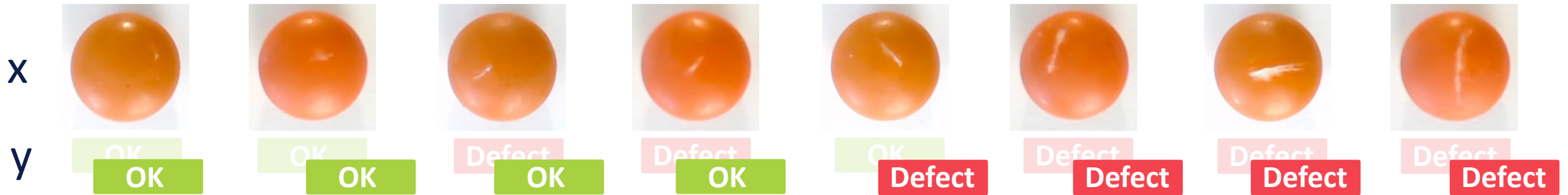




# Tip 1: Make the labels $y$ consistent

$$O\left(\frac{1}{\sqrt{m}}\right) \quad O\left(\frac{1}{m}\right)$$

**Ideal:** There is some deterministic (non-random) function mapping from  $x \rightarrow y$ , and the labels are consistent with this function.



Increasing scratch length





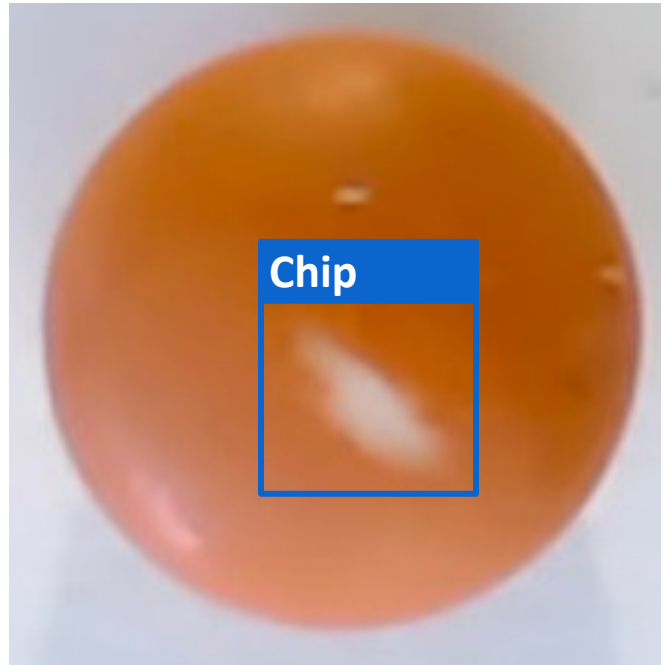
# Tip 2: Use multiple labelers to spot inconsistencies

## Examples of inconsistencies

**Label name**

Bounding box size

Number of bounding boxes



**Labeler 1**



**Labeler 2**

# Tip 2: Use multiple labelers to spot inconsistencies

## Examples of inconsistencies

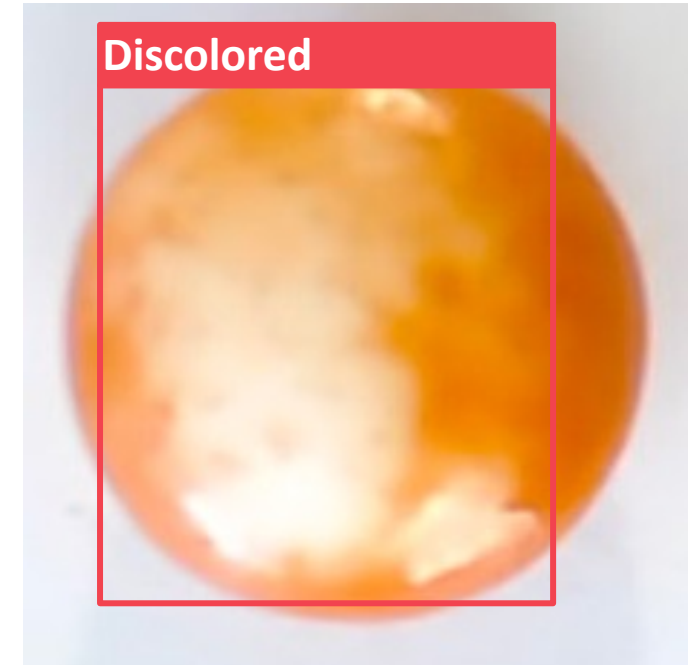
Label name

**Bounding box size**

Number of bounding boxes



**Labeler 1**



**Labeler 2**

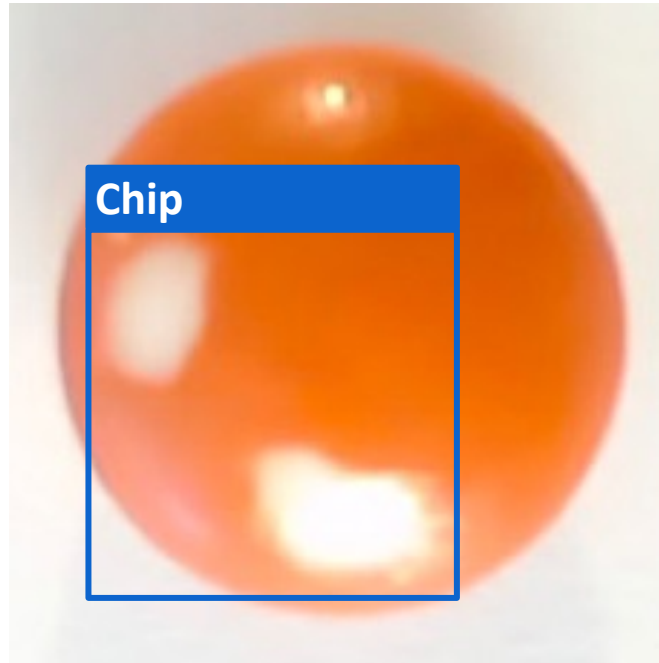
# Tip 2: Use multiple labelers to spot inconsistencies

## Examples of inconsistencies

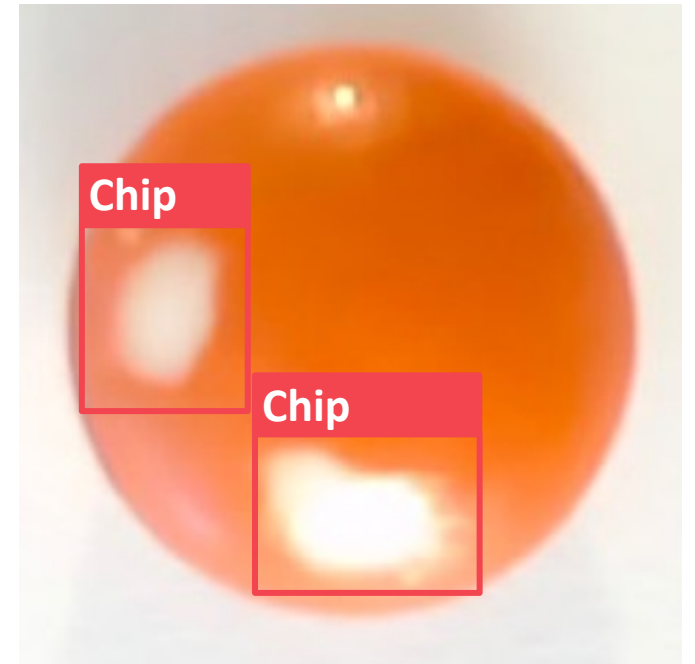
Label name

Bounding box size

**Number of bounding boxes**



Labeler 1



Labeler 2

# Tip 3: Repeatedly clarify labeling instructions by tracking down ambiguous examples

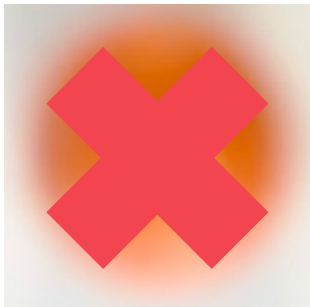
Repeatedly:

- Find examples where the label is ambiguous or inconsistent
- Make a decision on how they should be labeled
- Document that decision in your labeling instructions

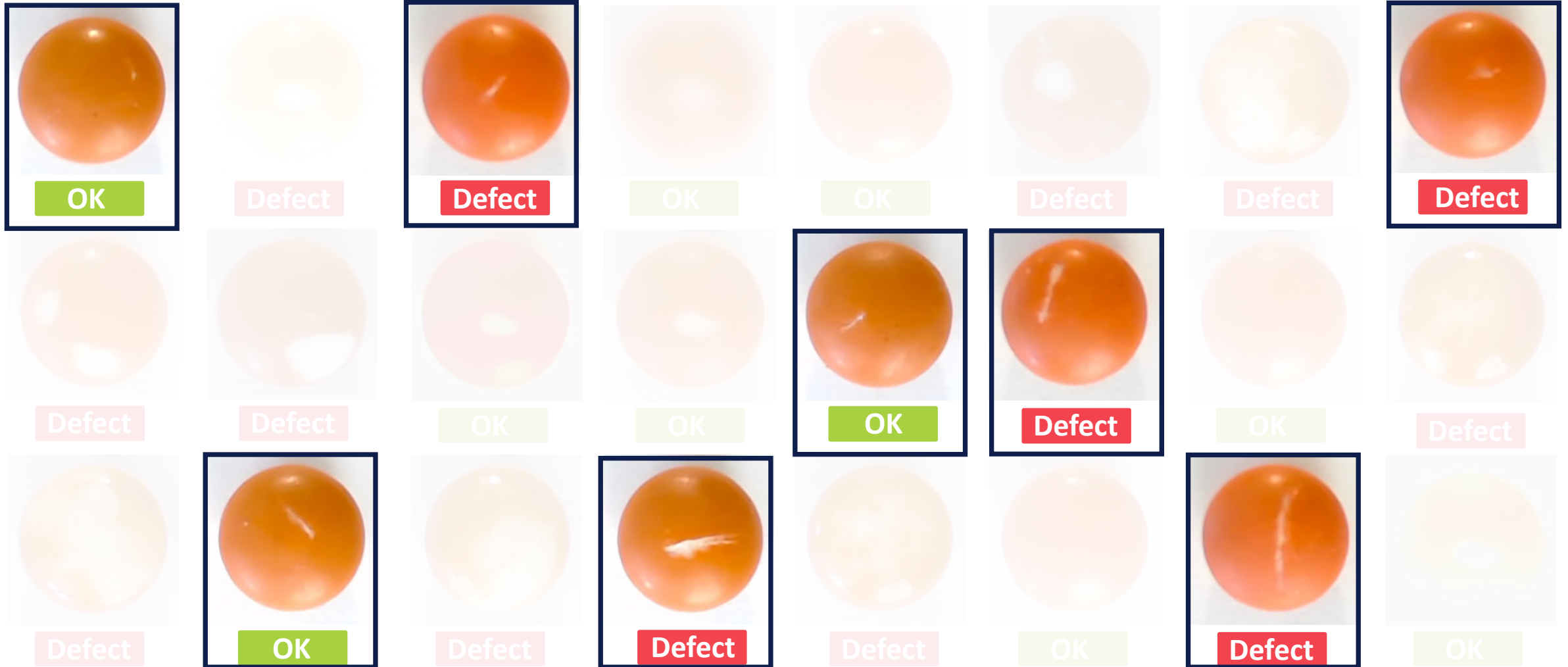
Labeling instructions should be illustrated with:

- Examples of concept (e.g., show some examples of scratched pills)
- Examples of borderline cases and near-misses
- Any other confusing examples

# Tip 4: Toss out bad examples. More data is not always better!



# Tip 5: Use error analysis to focus on subset of data to improve



Train model

# Iterative workflow

Improve data

Error analysis  
to decide on  
next step

Improve data via these methods:

- Multiple labelers to measure consistency
- Improve label definitions & relabel more consistently
- Toss out noisy examples or improve quality of input x
- Get more data through collection or data augmentation

improve y (labels)

improve x (images)



# Summary

Improving the data right is not a “preprocessing” step that you do once. It’s part of the iterative process of model development... as well as after that to deployment/monitoring/maintenance.

Tip 1: Make the labels  $y$  consistent

Tip 2: Use multiple labelers to spot inconsistencies

Tip 3: Clarify labeling instructions by tracking down ambiguous examples

Tip 4: Toss out noisy examples. More data is not always better!

Tip 5: Use error analysis to focus on subset of data to improve

# NEURIPS DATA-CENTRIC AI WORKSHOP

Date: 14 December 2021

Location: Virtual

<http://datacentricai.org/>



Early Submission Deadline

September 30, 2021



Notification of acceptance

October 22, 2021




Workshop

December 14, 2021

# Keep in touch!

 @AndrewYNg

 DeepLearning.AI

**THE BATCH**

*Essential news for deep learners*

thebatch.ai

Andrew Ng

**END END END**