Data-Centric AI Development: From Big Data to Good Data

Andrew Ng
Landing AI and DeepLearning.AI

AI cannot reach its

full potential

until it's accessible

to everyone.

Data-centric AI is

key to democratizing

access to AI.

But what does that mean?

Shifting from model-centric to data-centric AI

Conventional model-centric approach:

AI = Code + Data

(algorithm/model)

Work on this

Data-centric approach:

AI = Code + Data

(algorithm/model)

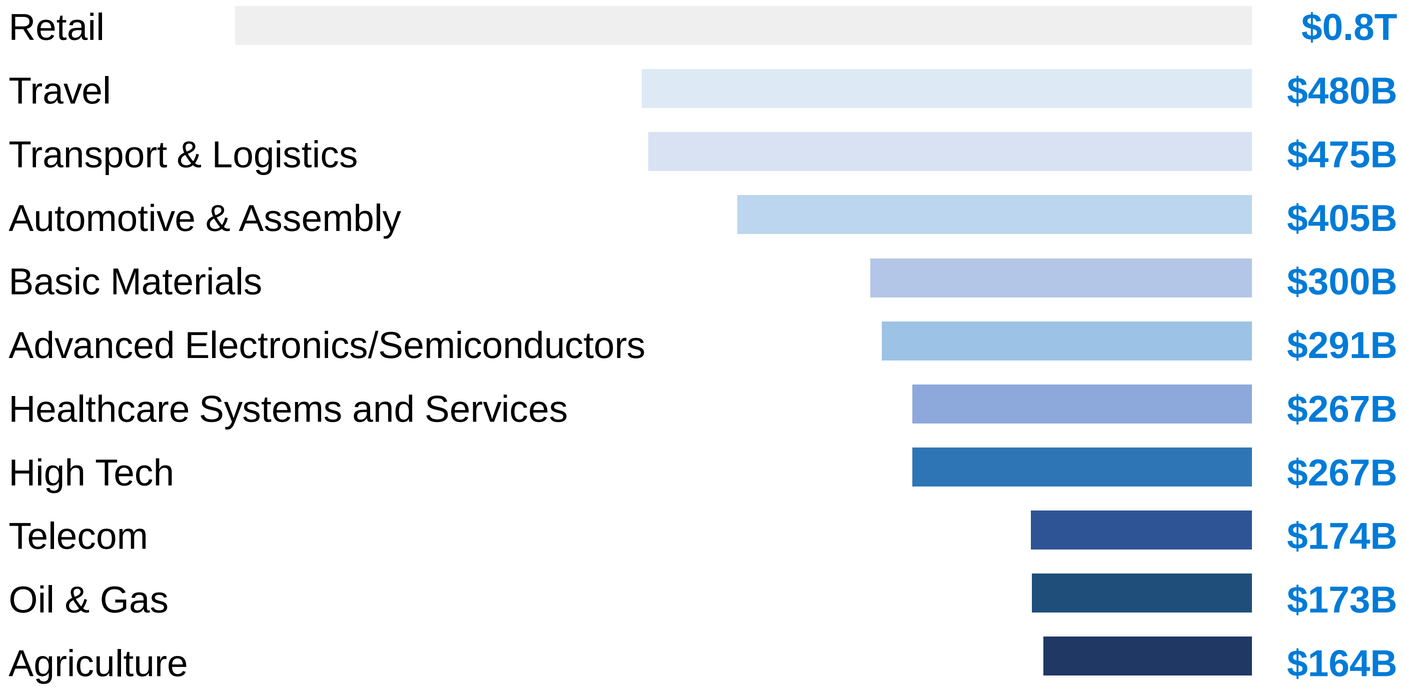
Work on this

Data-centric AI

is the discipline of systematically engineering the data used to build an AI system.

AI is changing all industries

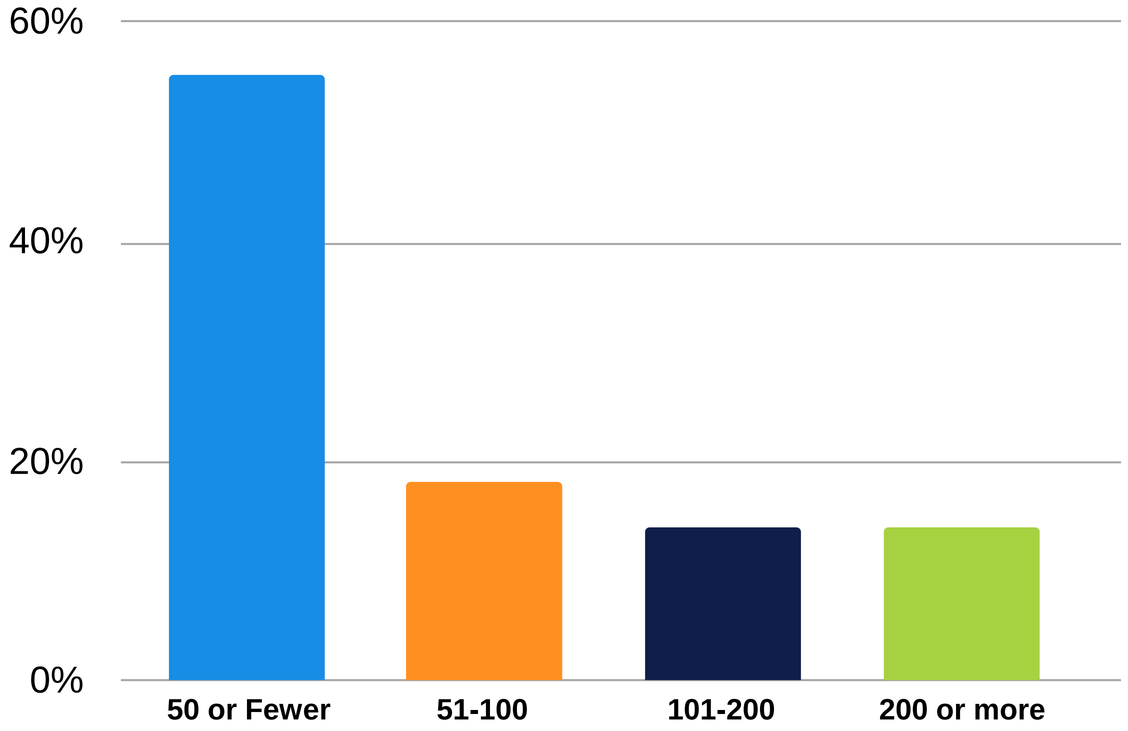
$13
TRILLION
AI value creation by 2030



Source: McKinsey

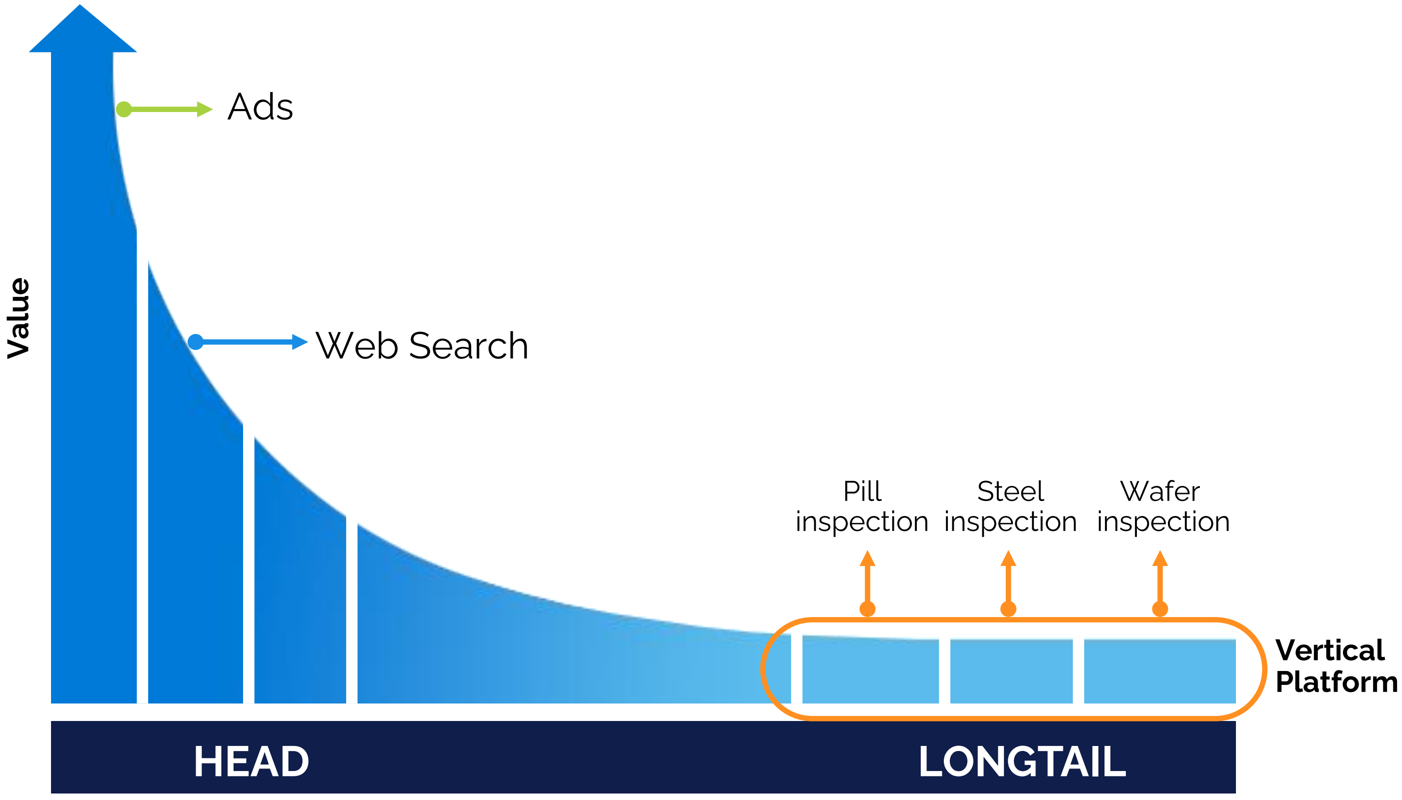
Barriers to widespread adoption #1: Small datasets

Manufacturing audience: How many images do you typically have of each defect type you want to detect?



Technology built for 100M images does not work for other industries.

Barriers to widespread adoption #2: Customization (long tail) problem



All potential AI projects, sorted in decreasing order of value

We need vertical platforms that **enable the end customer** to build the custom AI system they need.

They will do this by **engineering the data**, rather than the model.

From Big Data to Good Data

Supervised learning to learn x -> y mapping.

What makes a good dataset?

Consistent and accurate labels y

Representative and high-quality inputs x

Reflects post deployment changes (concept/data drift)

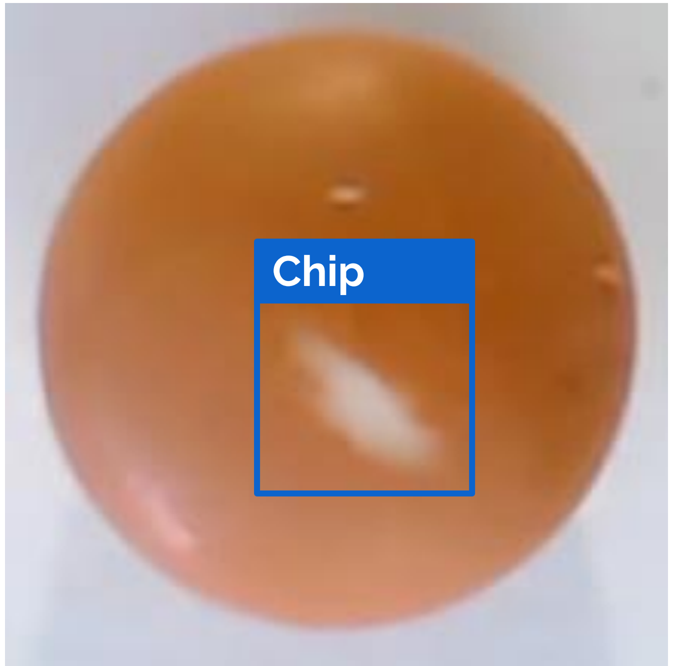
Consistent and accurate labels y

Examples of inconsistencies

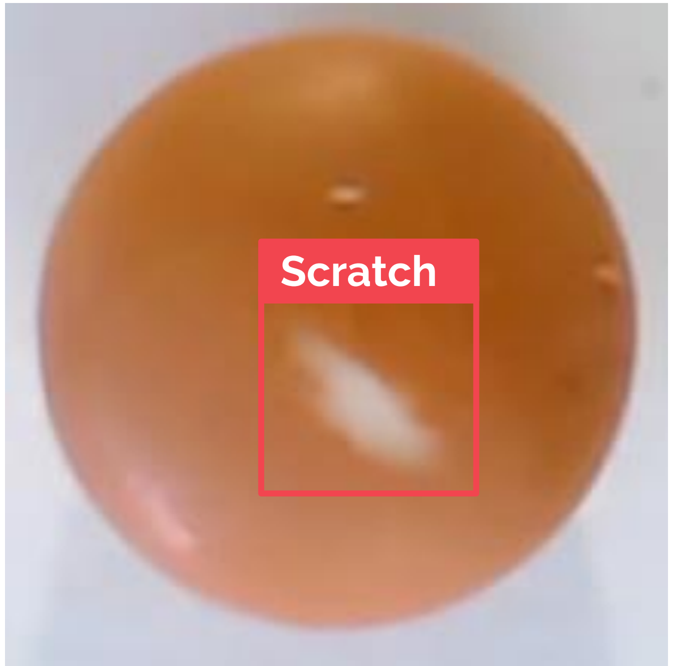
Label name

Bounding box size

Number of bounding boxes



Labeler 1



Labeler 2

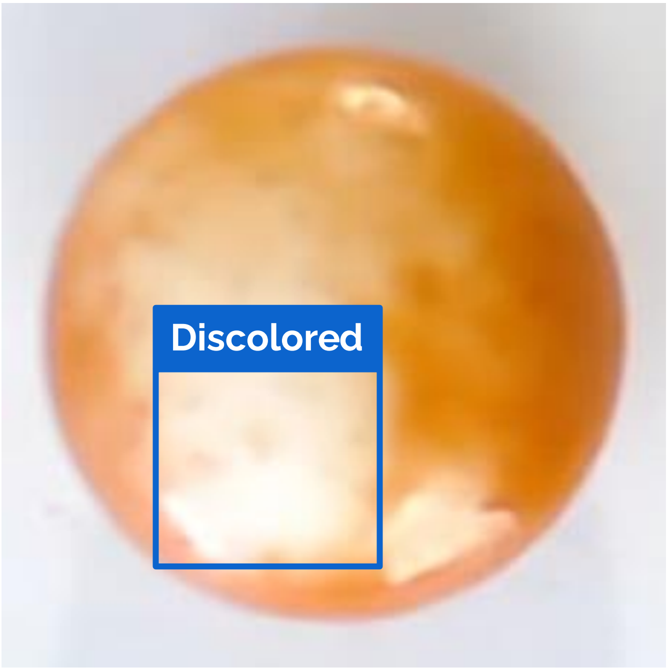
Consistent and accurate labels y

Examples of inconsistencies

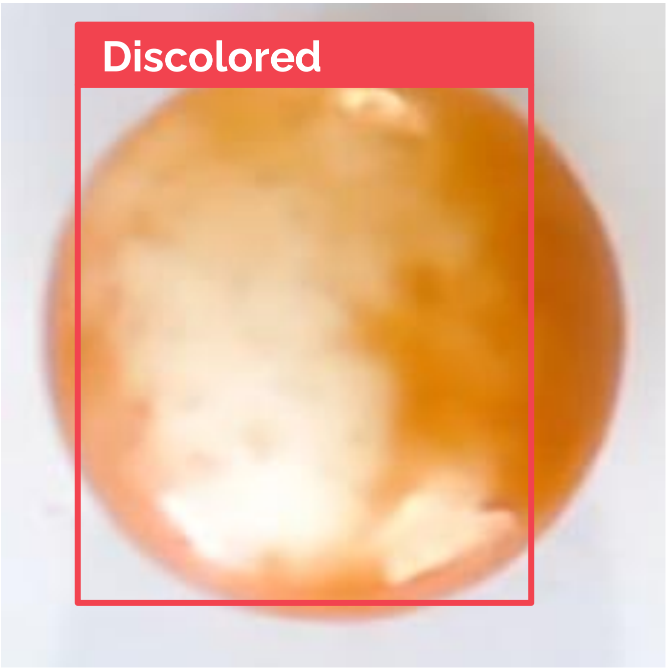
Label name

Bounding box size

Number of bounding boxes



Labeler 1



Labeler 2

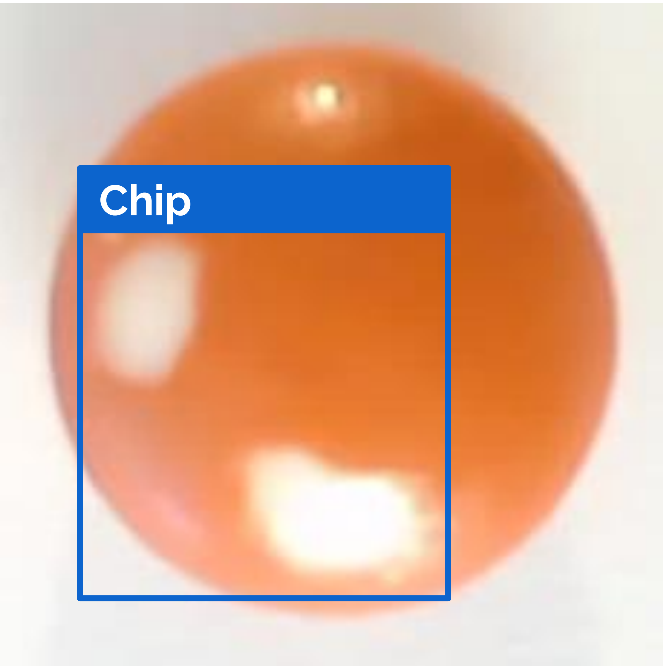
Consistent and accurate labels y

Examples of inconsistencies

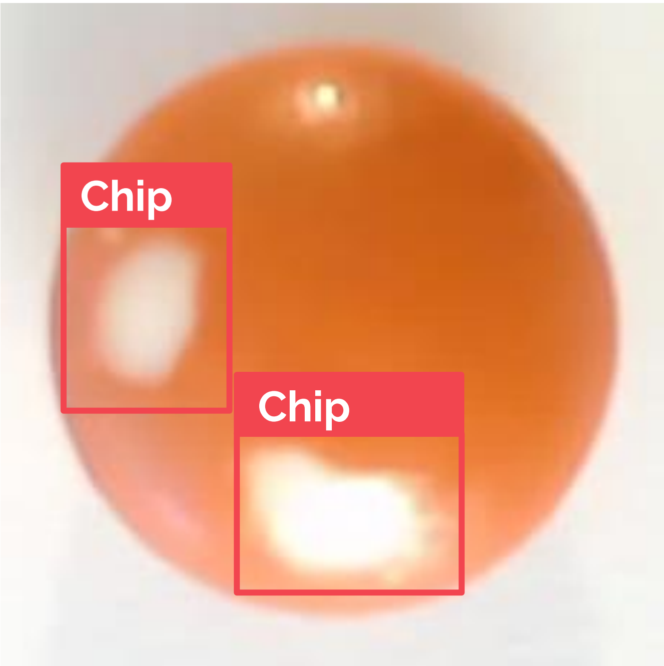
Label name

Bounding box size

Number of bounding boxes



Labeler 1



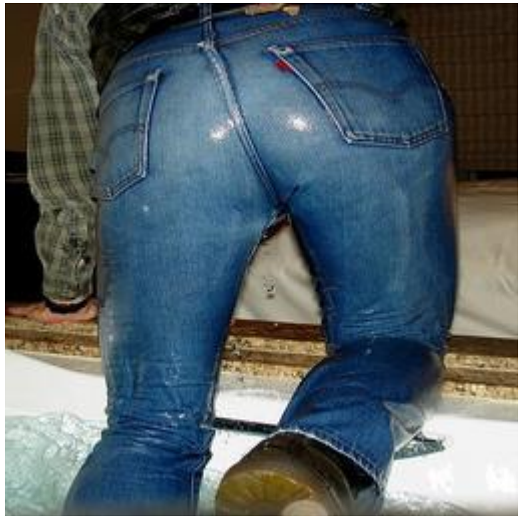
Labeler 2

LandingLens

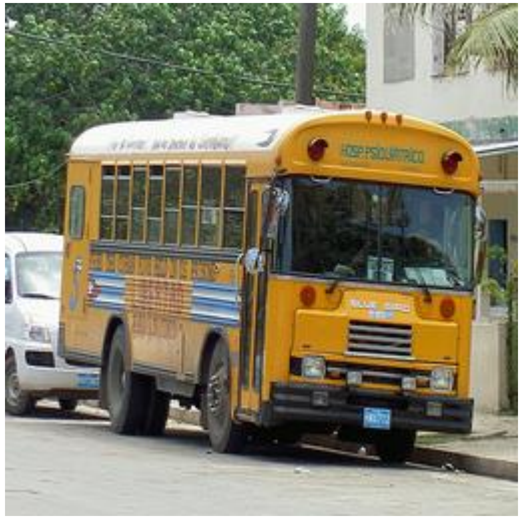
Demo: Defect book

Even the more revered datasets have errors

ImageNet

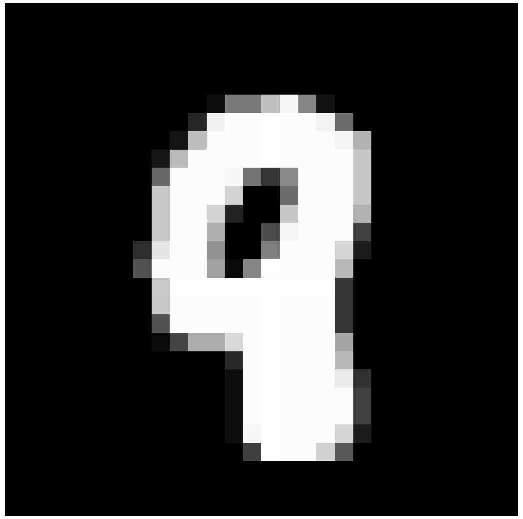


Label: Tub
Corrected: Jeans



Label: Passenger car
Corrected: School bus

MNIST



Label: 8
Corrected: 9

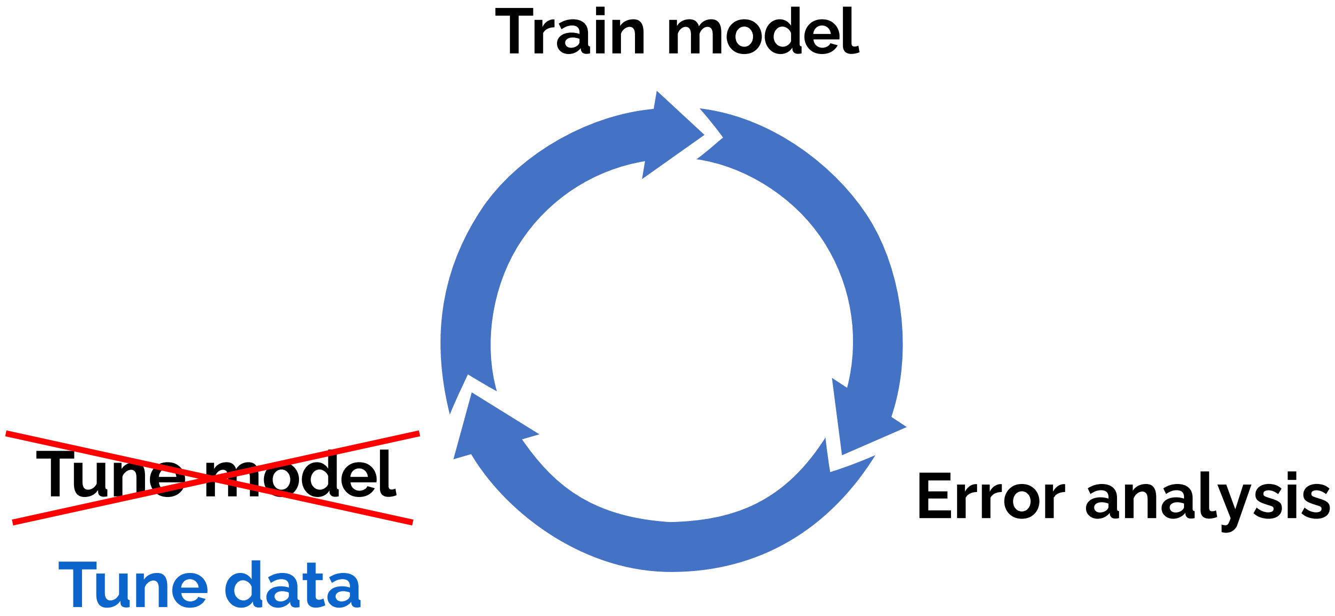
Amazon Reviews

"I've had this for over a year, and it works very well. I am very happy with this purchase."

Label: 1 star
Corrected: 5 stars (?)

[Northcutt et al., 2021. Confident Learning: Estimating uncertainty in dataset labels]

Data engineering data as part of ML workflow



Train model

Error analysis

~~Tune model~~

Tune data

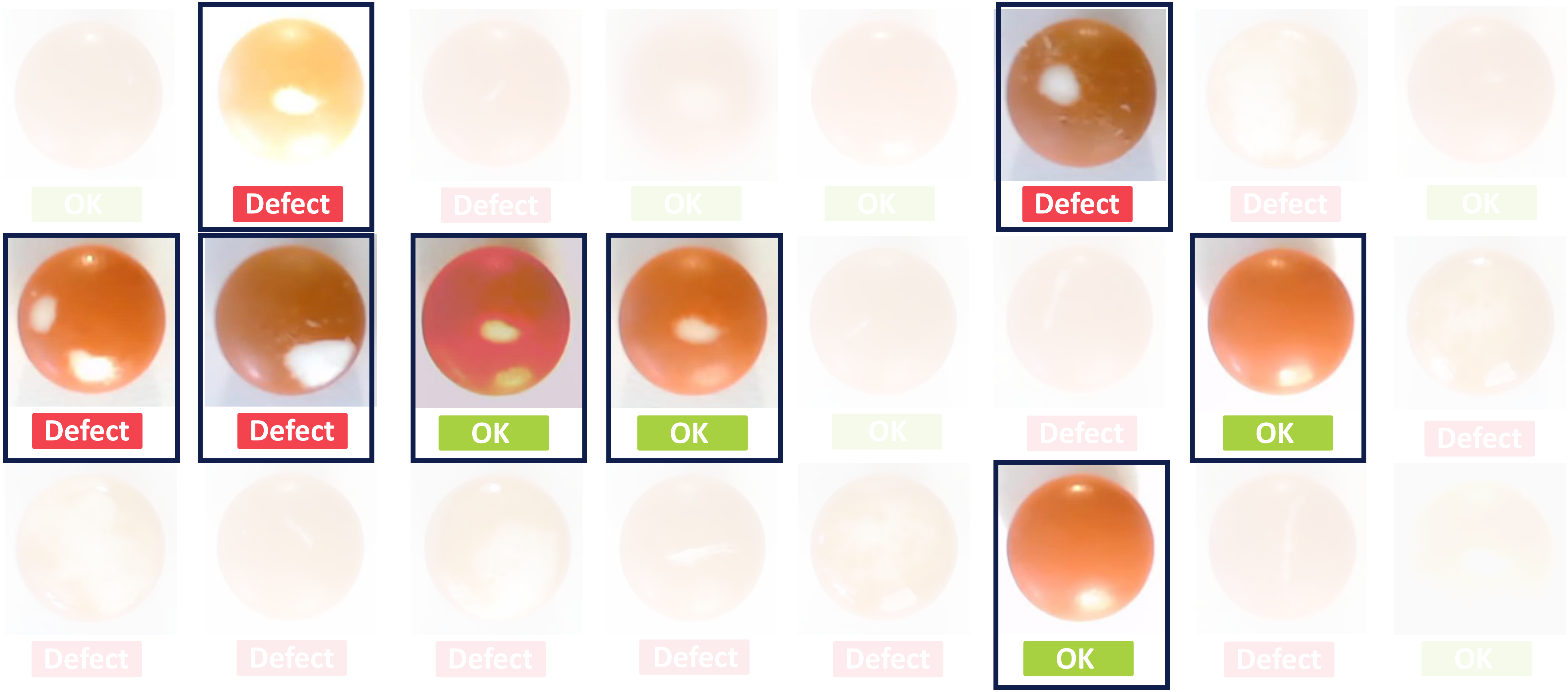
Data cleaning isn't a "pre-processing" step that you do once. It should be part of the iterative process of ML development.

Key idea: identifying what slice (subset) of data to improve

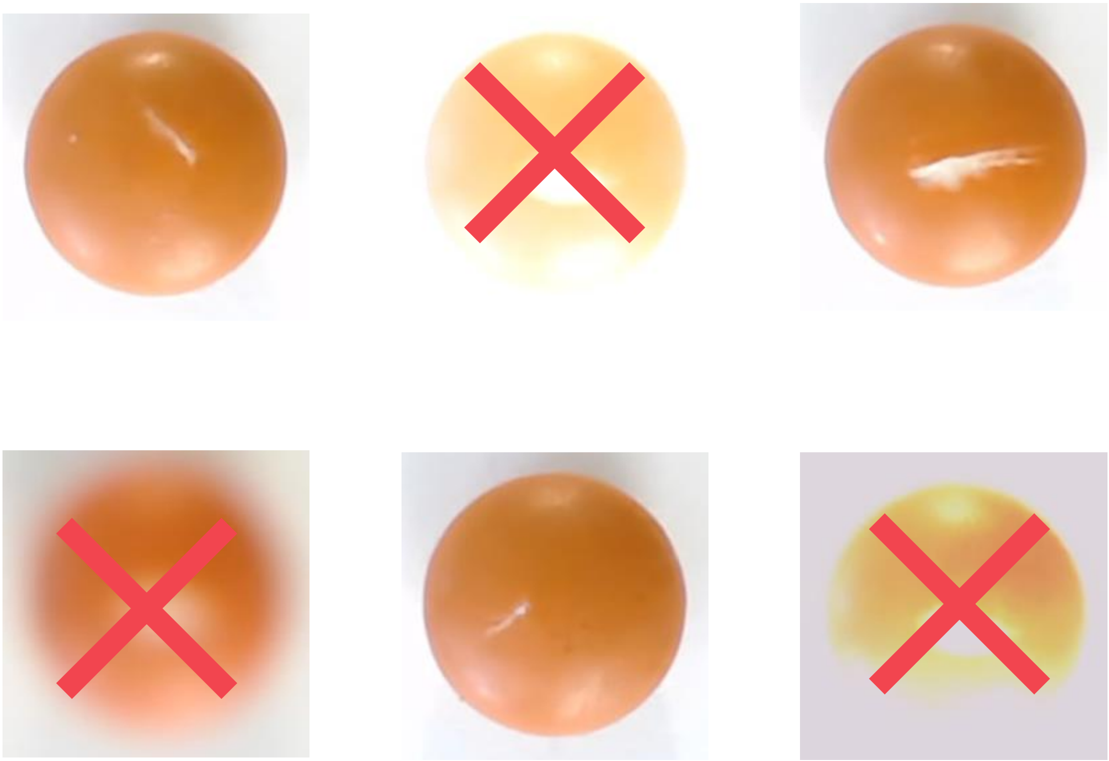
LandingLens

Demo: Agreement based labeling

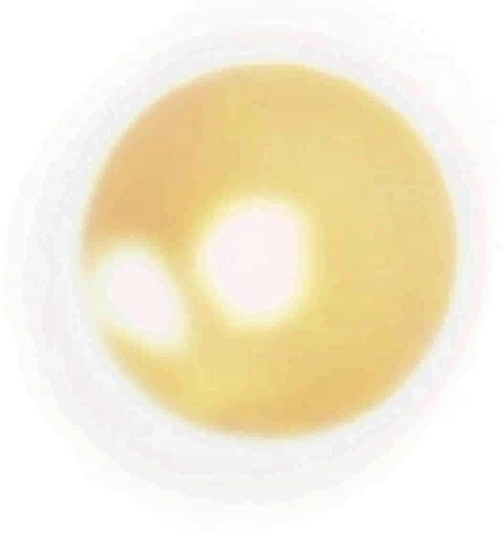
Decide which slice (subset) of data to prioritize improving via error analysis



Representative and high-quality inputs x: Improving input quality

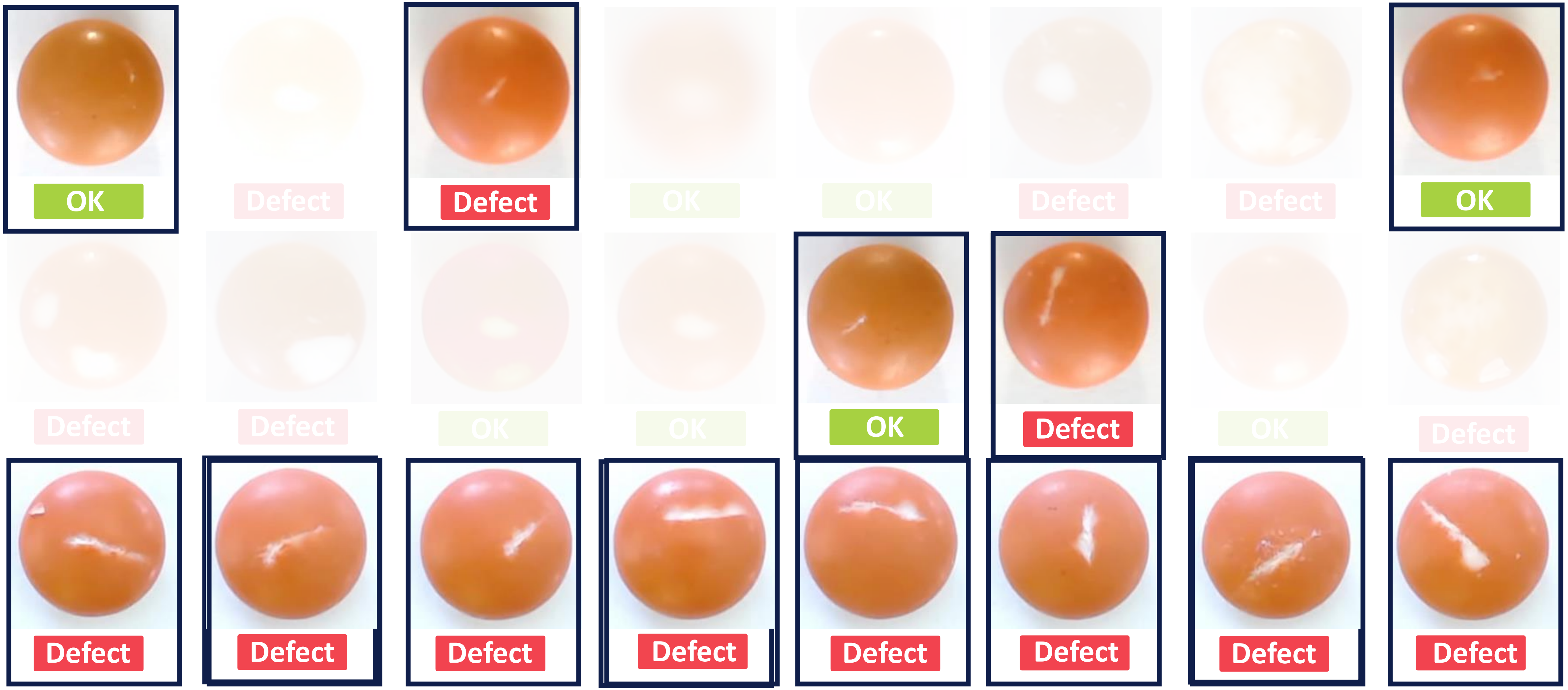


Poor imaging acquisition



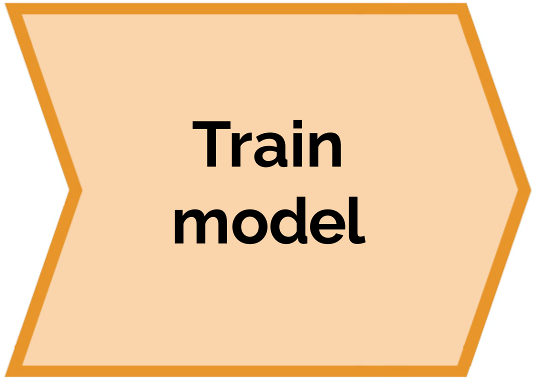
Improving imaging system design

Representative and high-quality inputs x:
Targeted data acquisition



Data reflects post deployment changes

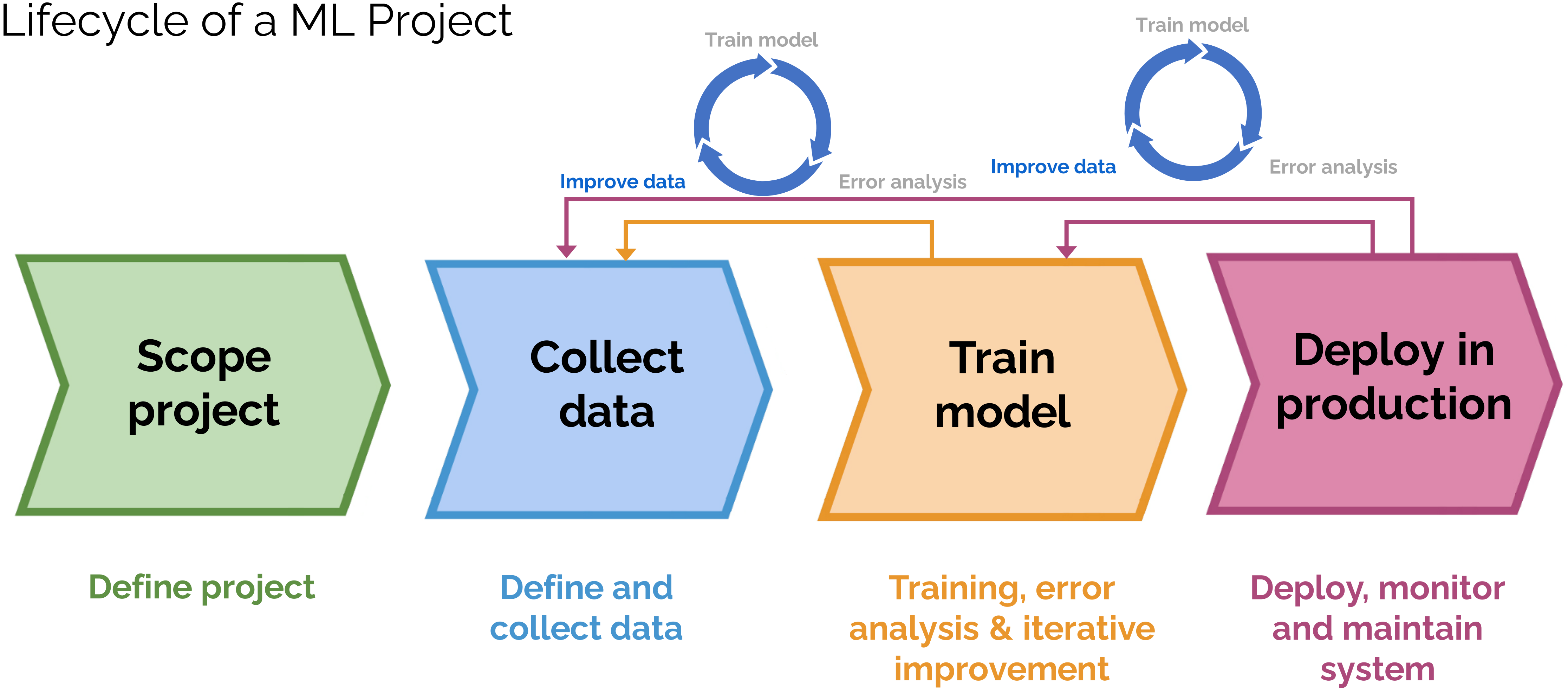
Lifecycle of a ML Project



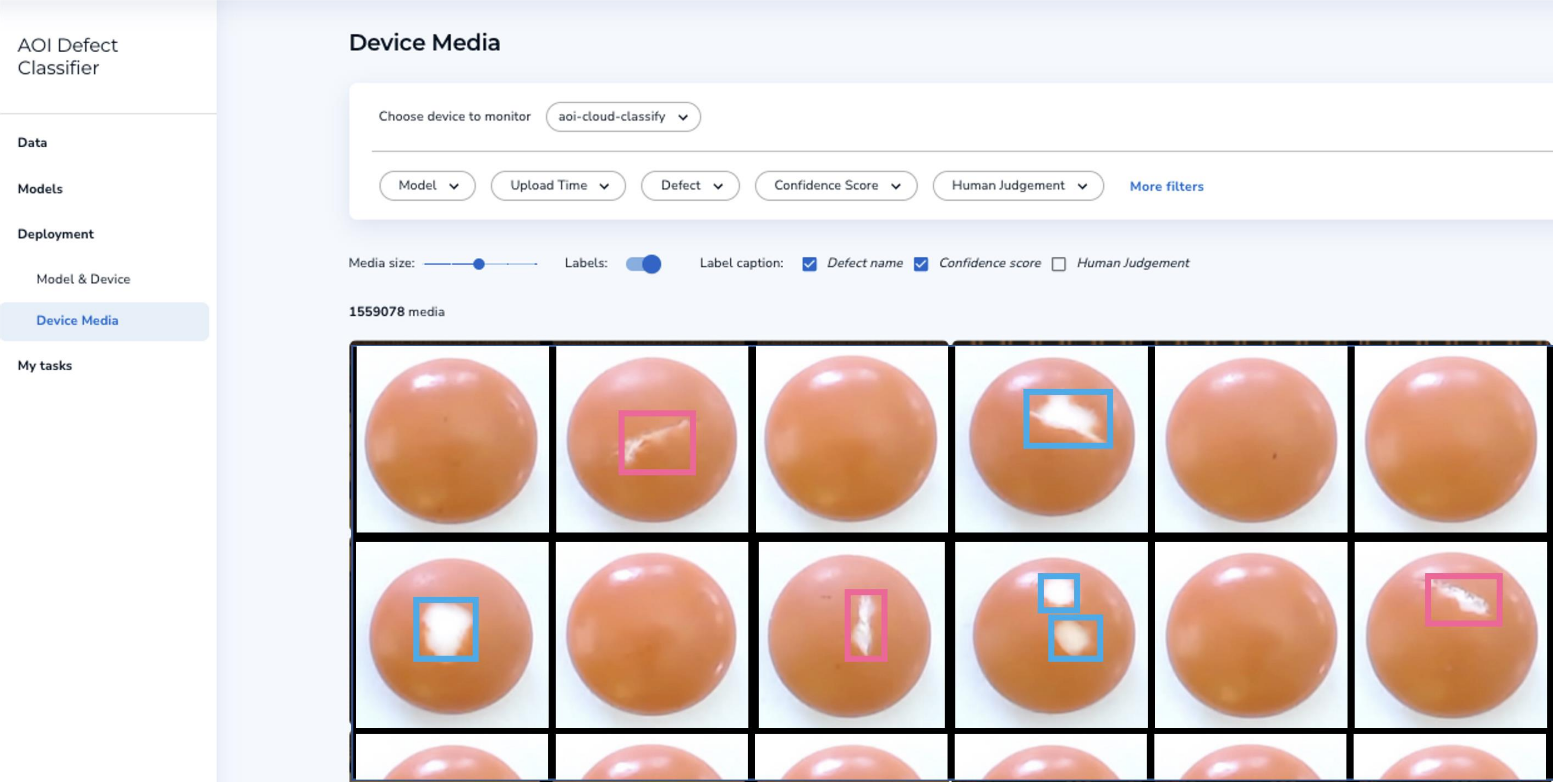
Training, error analysis & iterative improvement

Data reflects post deployment changes

Lifecycle of a ML Project



Deployment Dashboard (connected edge)



AOI Defect Classifier

Data

Models

Deployment

Model & Device

Device Media

My tasks

Device Media

Choose device to monitor: aoi-cloud-classify

Media size:

Labels:

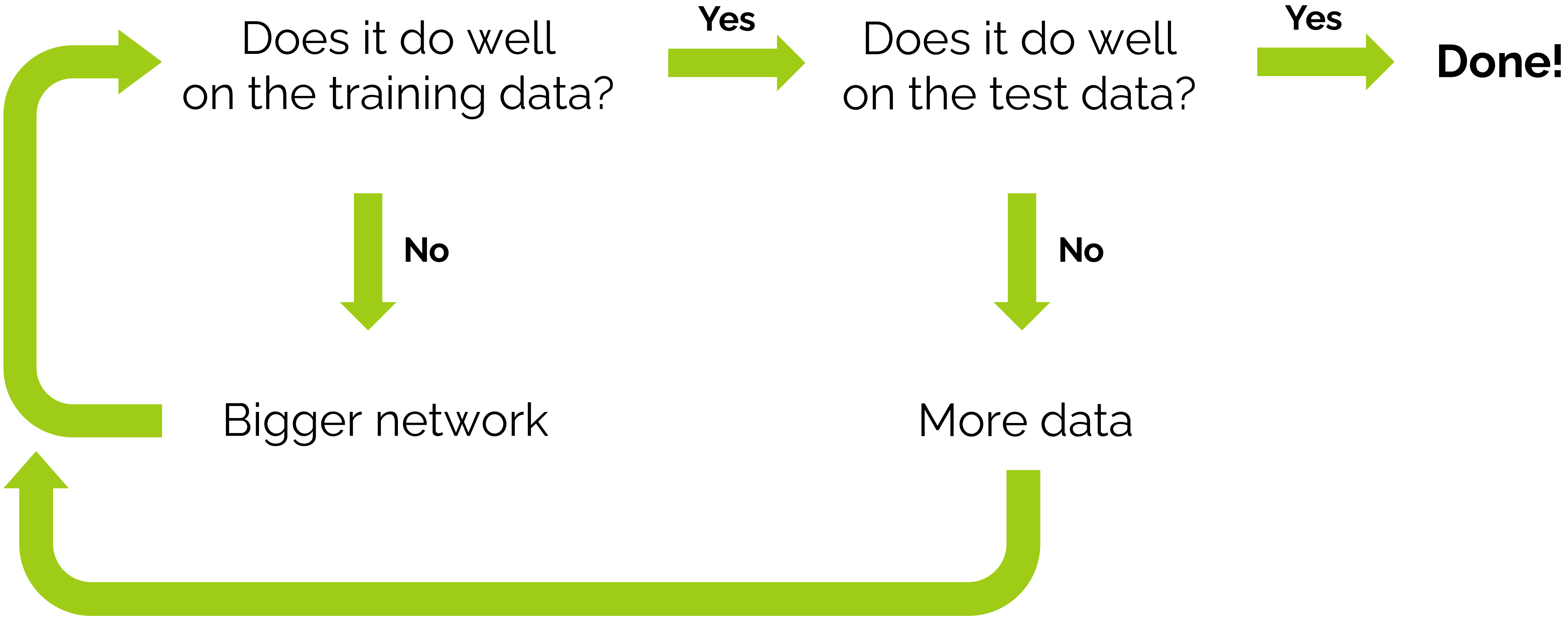
Label caption: ☑ *Defect name* ☑ *Confidence score* □ *Human Judgement*

1559078 media

LandingLens

Demo: Error Analysis

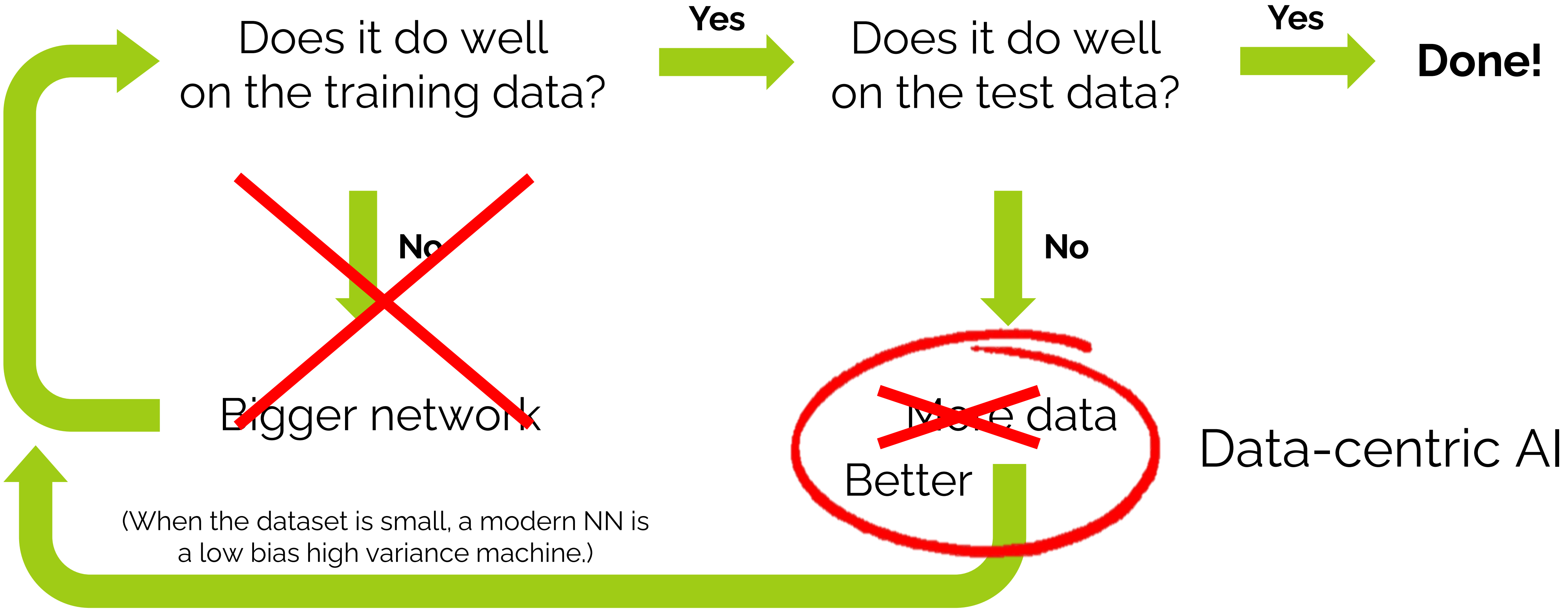
Big Data Era Recipe for Deep Learning



[From GTC 2015]

~~Big Data~~ Era Recipe for Deep Learning

Small Data



Does it do well on the training data?

Yes

Does it do well on the test data?

Yes

Done!

~~No~~

~~Bigger network~~

No

~~More~~ data

Better

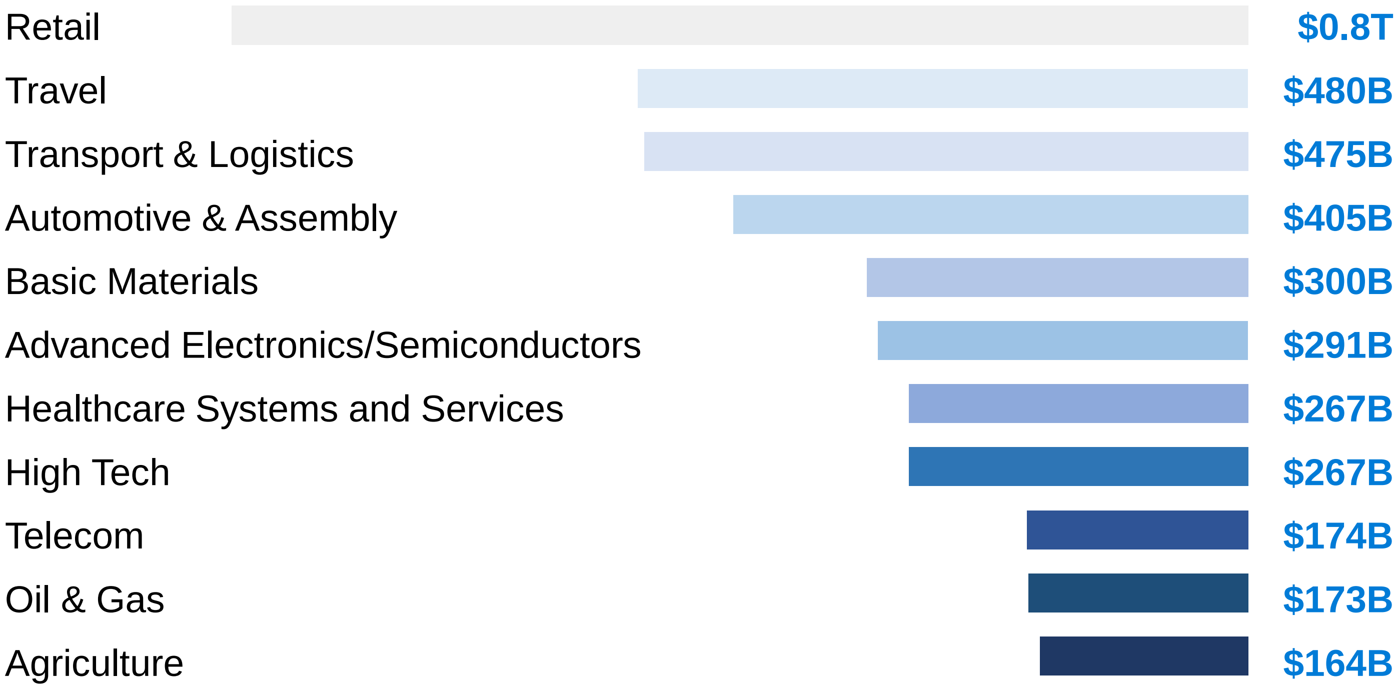
Data-centric AI

(When the dataset is small, a modern NN is a low bias high variance machine.)

[From GTC 2015]

AI is changing all industries

$13
TRILLION
AI value creation
by 2030

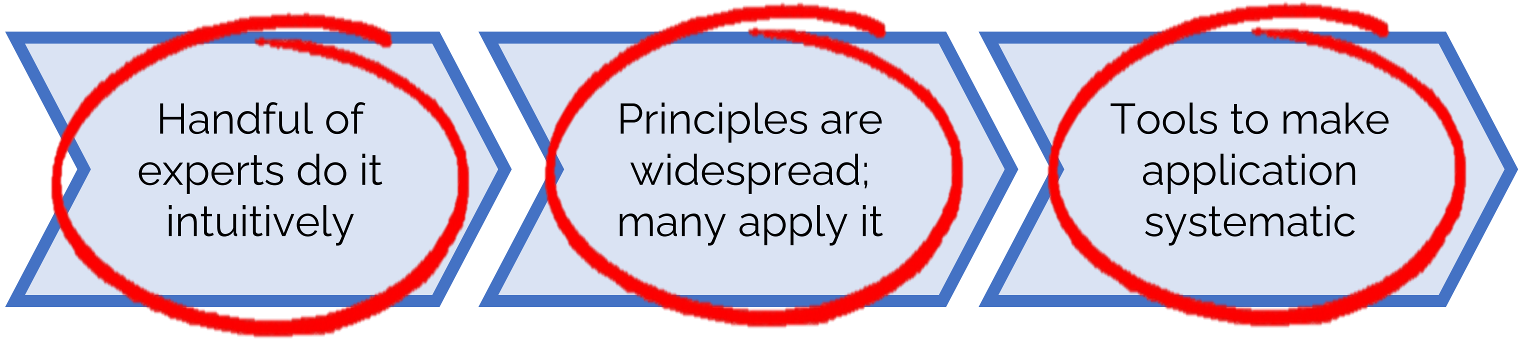


Data-centric AI will be particularly important to high-stakes applications, such as healthcare and loan approvals.

[Sambasivan, et al., 2021] Everyone wants to do the model work, not the data work.

Data-centric AI development – Summary

- Scaling up datasets and models has driven a lot of progress.
- But with the maturity of today's models, many applications require a shift to systematically engineering the data.


Handful of experts do it intuitively
Principles are widespread; many apply it
Tools to make application systematic

- Did not discuss in this talk: Structured data, Data cascades.
- Resources:
  - The Batch (thebatch.ai)
  - Data-centric AI resource hub (datacentricai.org)

Democratizing AI

benefits everyone.

Data-centric AI

is key to unlocking the

next era of AI.

Thank You

Andrew Ng

Landing AI and DeepLearning.AI