Copyright Notice

DeepLearning.AI

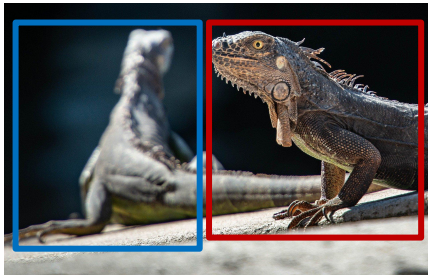
C1W3 Slides

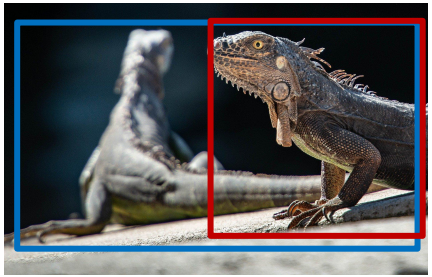
DeepLearning.AI

Define data and establish baseline

Why is data definition hard?

Iguana detection example

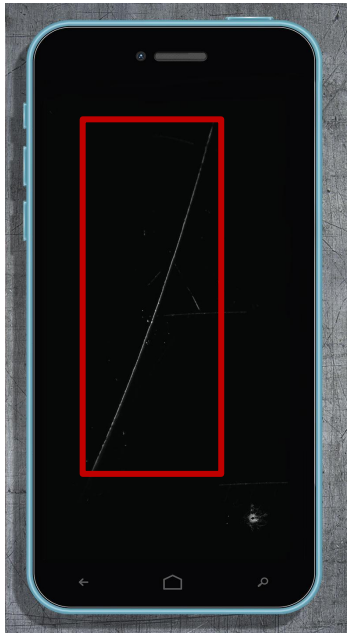


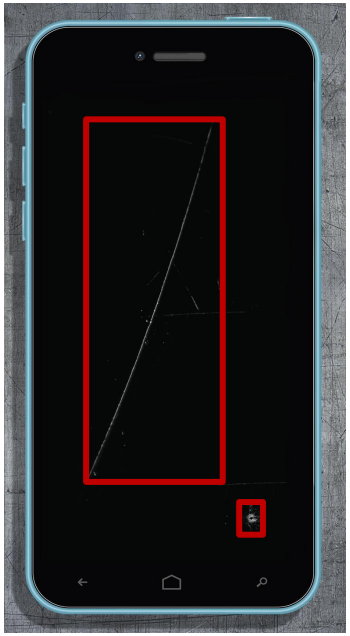


Labeling instructions: "Use bounding boxes to indicate the position of iguanas"

if some labellers use different convention, then its not good for algo.
→ learning is difficult.

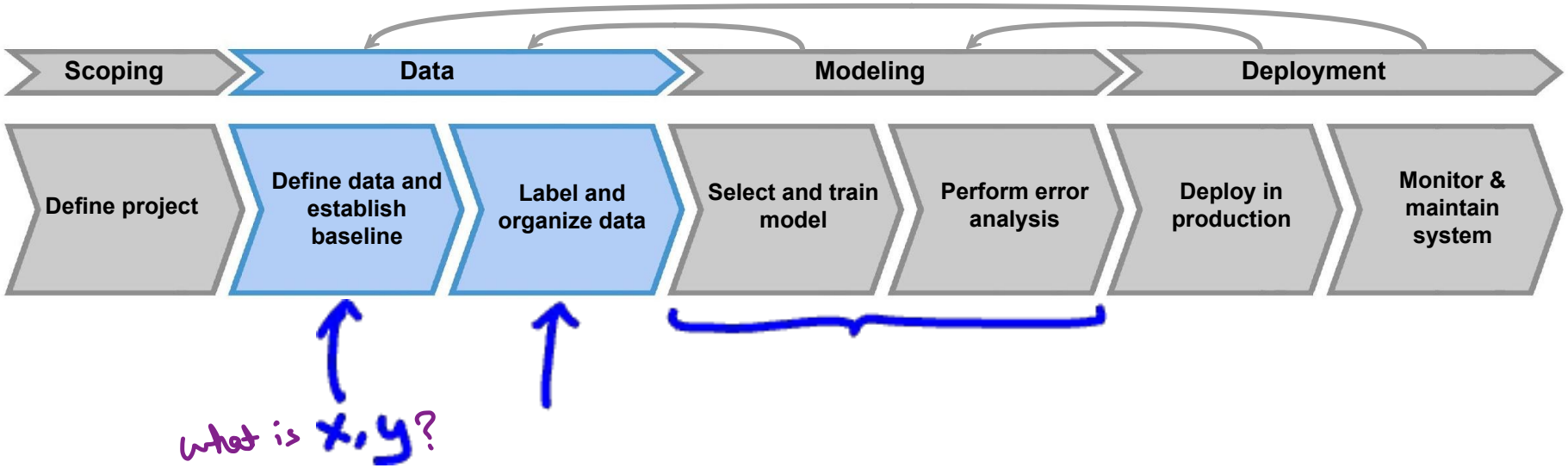
Phone defect detection





→ inconsistent labelling

Data stage



→ Best Practices for Data Stage

→ Many Practitioners downloads data from internet to start with. Data makes huge impact in success of Project

DeepLearning.AI

Define data and establish baseline

More label ambiguity examples

Speech recognition example

"Um, nearest gas station"

"Umm, nearest gas station"

”Nearest gas station [unintelligible]"

* Standardise one convention for the labels.

User ID merge example

Merge Data (New Company)

Job website

Job App

	Job Board (website)	Resume chat (app)
Email	nova@deeplearning.ai	nova@chatapp.com
First Name	Nova	Nova
Last Name	Ng	Ng
Address	1234 Jane Way	?
State	CA	?
Zip	94304	94304

- is it a bot/spam account?
- fraudulent transaction?
- looking for job?

→ bot/spam.
→ fraud transc.
→ looking for job?

ambiguous

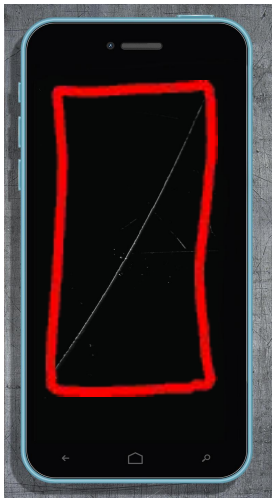
Build a model

1 if same

0 if different

are these 2 same person?

Data definition questions



- What is the input x?
  - Lightning? Contrast? Resolution?
  - What features need to be included?
- What is the target label y?
  - How can we ensure labelers give consistent labels?

Quality of input

for structure e.g. (gps location sought)

Hard for person. → ignore this image

get better image

DeepLearning.AI

Define data and establish baseline

Major types of data problems

Major types of data problems

If you're working on a problem from one of these 4 quadrants. Advice from someone who has worked on problems in the same quadrant will probably be more useful than someone from diff. quadrants.

	Unstructured	Structured
Small data	Scratch Detection 100 phones	House Price Prediction 50 data pts
Big data	Speech recog. 50M.	Online Shopping Rec. sys. 1M users

≤ 10,000

Clean labels are critical. (small data)

↓ Consistent

> 10,000

Emphasis on data process. (Hard to clean labels)

How:
- Collect & store data
- labelling instructions

Humans can label data.

Data augmentation.

Harder to obtain more data. & augmentation (for structured data)

Unstructured vs. structured data

Unstructured data

- May or may not have huge collection of unlabeled examples x.
- Humans can label more data. (doesn't apply to all problems)
- Data augmentation more likely to be helpful.

Structured data

- May be more difficult to obtain more data.
- Human labeling may not be possible (with some exceptions). [Harder]

Small data vs. big data

< 10,000 > 10,000

Small data

- Clean labels are critical.
- Can manually look through dataset and fix labels.
- Can get all the labelers to talk to each other. (for consistency)

Big data

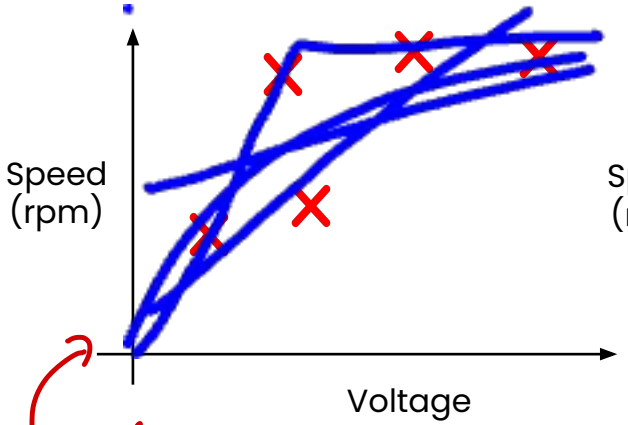
- Emphasis data process.

DeepLearning.AI

Define data and establish baseline

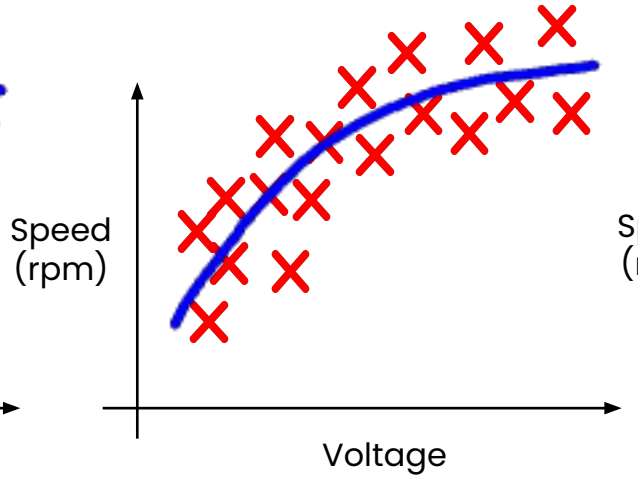
Small data and label consistency

Why label consistency is important

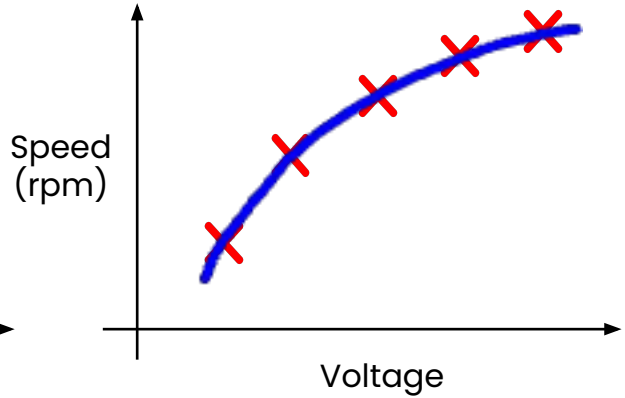


Not sure what should be the function

- Small data (5 Points)
- Noisy labels



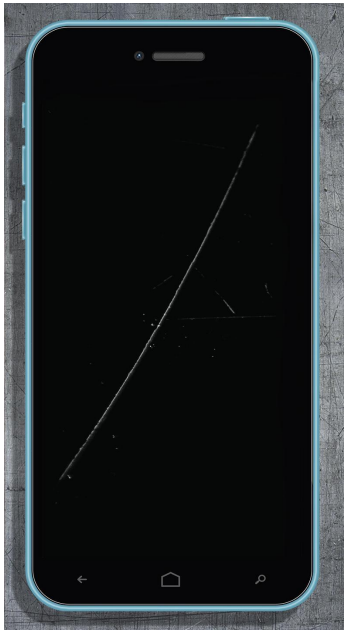
- Big data
- Noisy labels

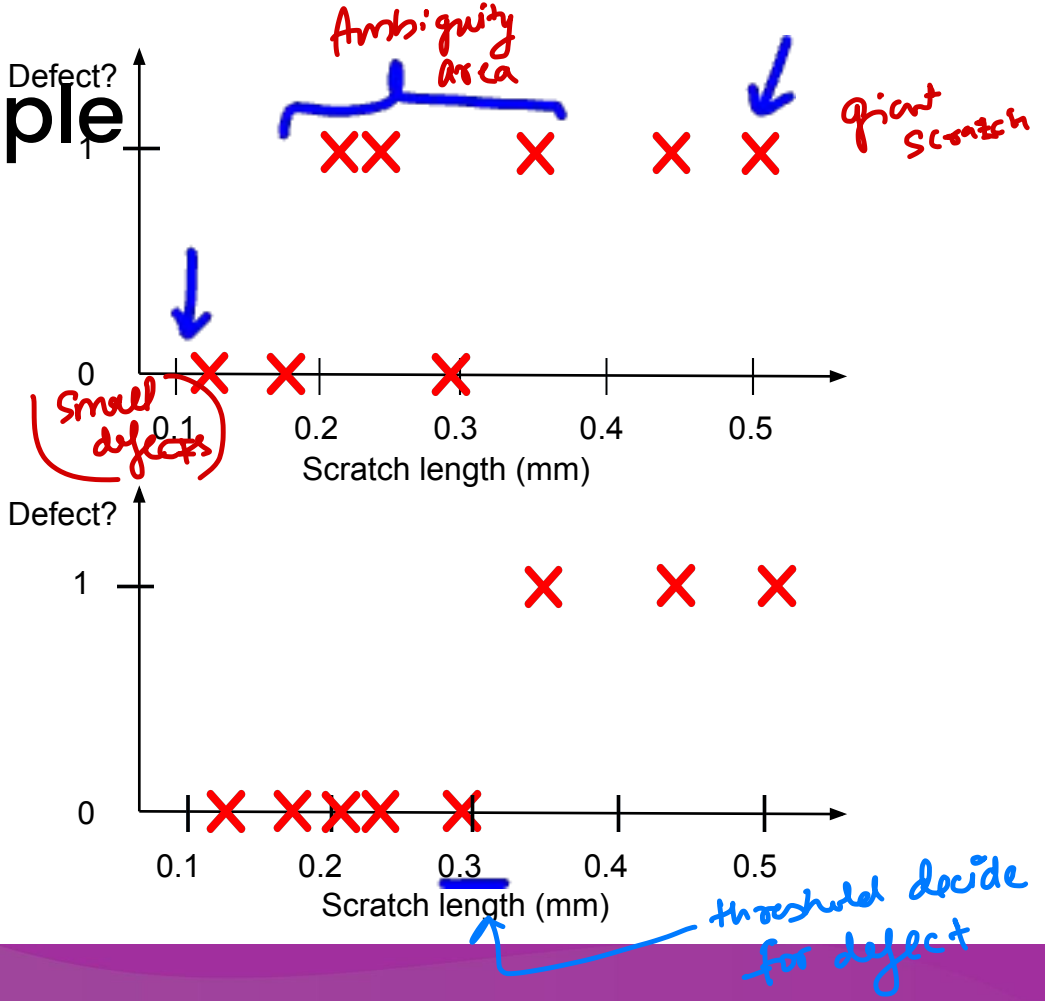


- Small data
- Clean (consistent) labels

[V. good data & model]

Phone defect example





Big data problems can have small data challenges too

Problems with a large dataset but where there's a long tail of rare events in the input will have small data challenges too.

- Web search
- Self-driving cars
- Product recommendation systems

DeepLearning.AI

Define data and establish baseline

Improving label consistency

Improving label consistency

If - some Points have inconsistency :-

- Have multiple labelers label same example.
- When there is disagreement, have MLE, subject matter expert (SME) and/or labelers discuss definition of y to reach agreement.
- If labelers believe that x doesn't contain enough information, consider changing x. (v. dark Phones) → get Brighter Phone images.
- Iterate until it is hard to significantly increase agreement.

Examples

- Standardize labels

  "Um, nearest gas station"

  "Umm, nearest gas station"

  ”Nearest gas station [unintelligible]”

[Standardise label]

➡ "Um, nearest gas station"

- Merge classes

These can be same class 'scratch'

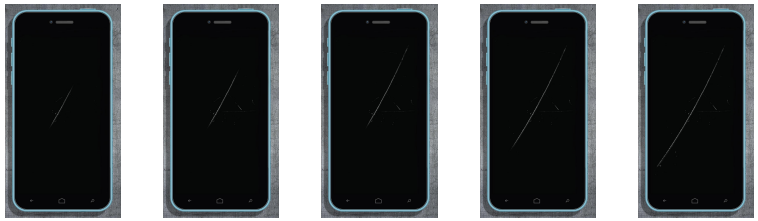
Deep scratch

Shallow scratch

➡ Scratch

Have a new class/label to capture uncertainty

- Defect: 0 or 1



Alternative: 0, Borderline, 1

No

defect

- Unintelligible audio

"nearest go"

"nearest grocery"

"nearest [unintelligible]"

"New Tag" → More Consistent across add data Points

Small data vs. big data (unstructured data)

Small data

- Usually small number of labelers.
- Can ask labelers to discuss specific labels.

Big data

- Get to consistent definition with a small group.
- Then send labeling instructions to labelers.
- Can consider having multiple labelers label every example and using voting or consensus labels to increase accuracy.

⇒ first & foremost is having clear consistent labelling instructions

DeepLearning.AI

Define data and establish baseline

Human level performance (HLP)

Why measure HLP?

is it how 2 humans agree with each other?

or Max. Performance that is possible?

- Estimate Bayes error / irreducible error to help with error analysis and prioritization.

expectation 99%

Ground Truth Label (given by another human)	Inspector
1	✓
1	✗
1	✓
0	✓
0	✓
0	✗

Human²

Human level Performance

66.7% accuracy

Other uses of HLP

- In academia, establish and beat a respectable benchmark to support publication. Beat HLP & get your paper published.
- Business or product owner asks for 99% accuracy. HLP helps establish a more reasonable target.
- "Prove" the ML system is superior to humans doing the job and thus the business or product owner should adopt it.

The problem with beating HLP as a "proof" of ML "superiority"

"Um... nearest gas station" ← 70% of labels

"Um, nearest gas station" ← 30%

Two random labelers agree: $0.7^2 + 0.3^2 = 0.58$

ML agrees with humans: 0.70 ← +12%

This is ML improvement. But its not significant. Not of any help. It's due to inconsistent labelling

The 12% better performance is not important for anything! This can also mask more significant errors ML may be making.

Side effect → This will hide/mask the errors made by Model on other types of inputs. Bcoz this is in majority number.

DeepLearning.AI

Define data and establish baseline

Raising HLP

Raising HLP

When the ground truth label is externally defined, HLP gives an estimate for Bayes error / irreducible error.

But often ground truth is just another human label.

Raising HLP

- When the label *y* comes from a human label, HLP << 100% may indicate ambiguous labeling instructions.

Um, Um... → inconsistency

- Improving label consistency will raise HLP. upto 100%.

- This makes it harder for ML to beat HLP. But the more consistent labels will raise ML performance, which is ultimately likely to benefit the actual application performance.

Get the labeller & Inspector agree of data consistency of the size of scratch for defect.
then HLP is increased to 100% from 66%.

Mostly low HLP stems from inconsistent labels.

HLP on structured data

Structured data problems are less likely to involve human labelers, thus HLP is less frequently used.

Some exceptions:

- User ID merging: Same person? from 2 [Company] Bisusiness merging
- Based on network traffic, is the computer hacked?
- Is the transaction fraudulent?
- Spam account? Bot?
- From GPS, what is the mode of transportation – on foot, bike, car, bus?

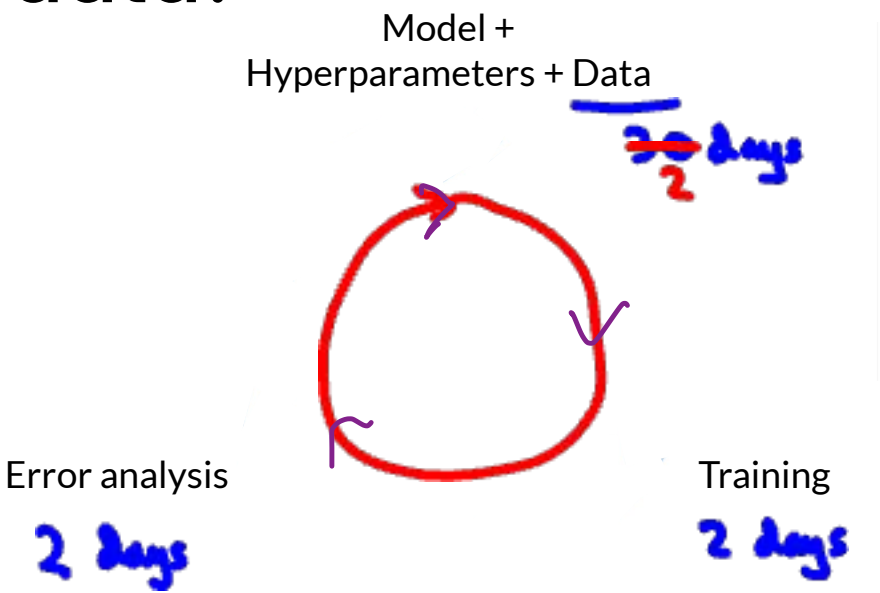
↓ Stops at Bus stops

DeepLearning.AI

Label and organize data

Obtaining data

How long (time) should you spend obtaining data?



- Get into this iteration loop as quickly possible.
- Instead of asking: How long it would take to obtain m examples?
  Ask: How much data can we obtain in k days.
- Exception: If you have worked on the problem before and from experience you know you need m examples.

Inventory data

Brainstorm list of data sources ( speech recognition)

Source	Amount	Cost	Time (to execute)
Owned	100h	$0	0 ✓
Crowdsourced – Reading	1000h	$10000	14 days
Pay for labels	100h	$6000	7 days
Purchase data	1000h	$10000	1 day ✓

Other factors: Data quality, privacy, regulatory constraints

Labeling data

specialized company

multiple people labelling

- Options: In-house vs. outsourced vs. crowdsourced
- Having MLEs label data is expensive. But doing this for just a few days is usually fine.
- Who is qualified to label?

Speech recognition – any reasonably fluent speaker

Factory inspection, medical image diagnosis – SME (subject matter expert)

Recommender systems – maybe impossible to label well

- Don't increase data by more than 10x at a time

generally → 5x, 2x

DeepLearning.AI

Label and organize data

Data pipeline

(Preprocessing steps)

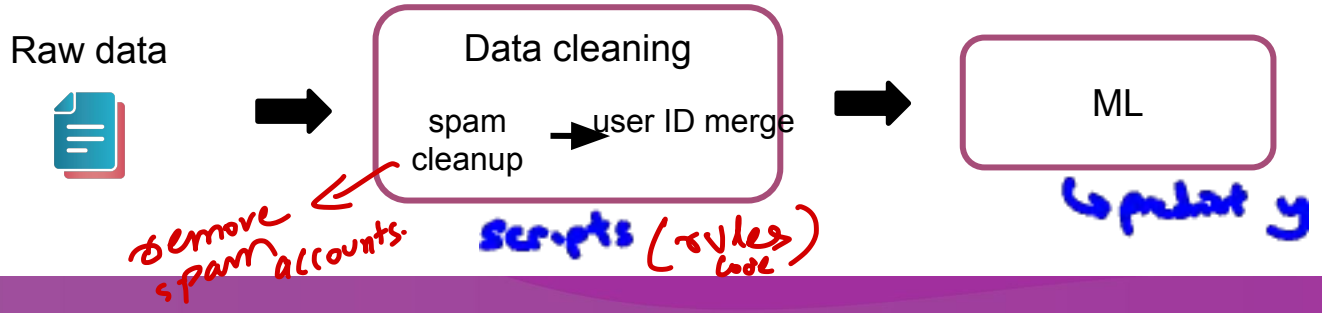
Data pipeline example

	Job Board (website)	Resume chat (app)
Email	nova@deeplearning.ai	nova@chatapp.com
First Name	Nova	Nova
Last Name	Ng	Ng
Address	1234 Jane Way	?
State	CA	?
Zip	94304	94304

x = user info

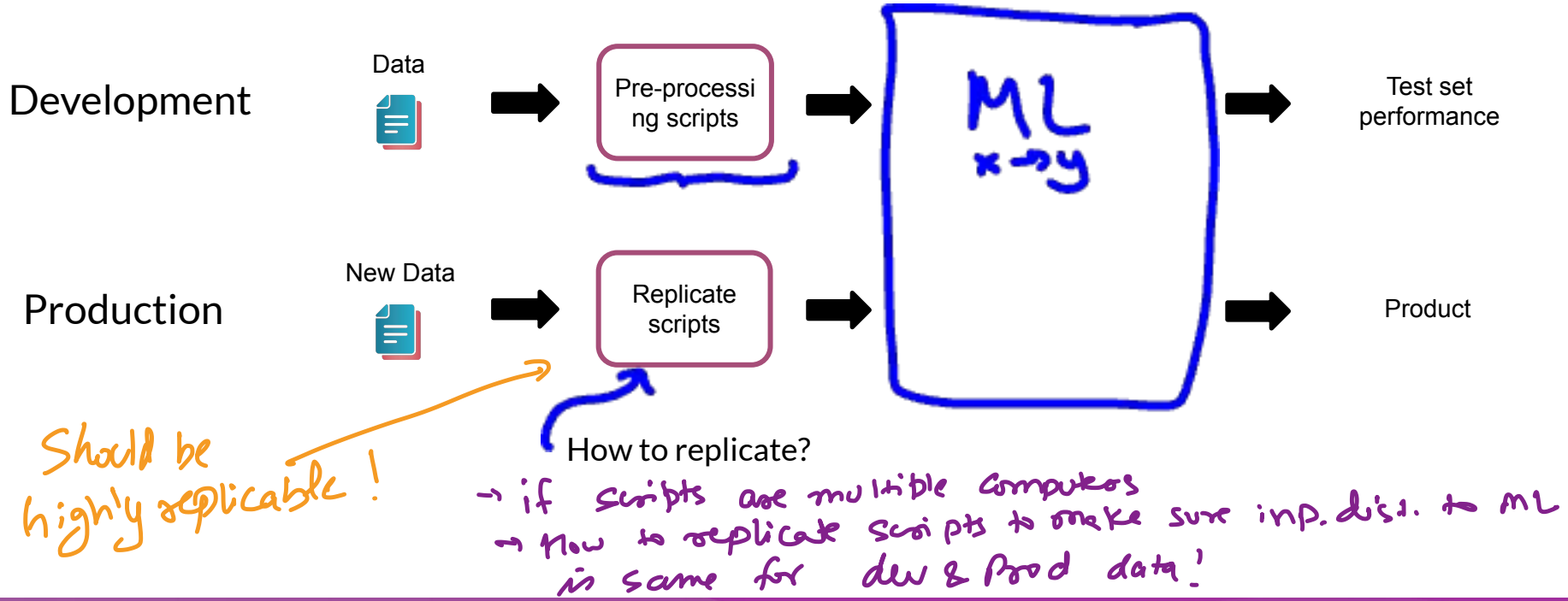
y = looking for job

→ want to give job ads



Issue in Production: "Replicability"

Data pipeline example



Should be highly replicable!

→ if scripts are multiple computers
→ how to replicate scripts to make sure inp. dist. to ML is same for dev & prod data!

POC and Production phases

POC (proof-of-concept):

- Goal is to decide if the application is workable and worth deploying.
- Focus on getting the prototype to work!
- It's ok if data pre-processing is manual. But take extensive notes/comments.

if Project succeed. then replicate scripts for Prod phase.

Production phase:

- After project utility is established, use more sophisticated tools to make sure the data pipeline is replicable.
- E.g., TensorFlow Transform, Apache Beam, Airflow,....

DeepLearning.AI

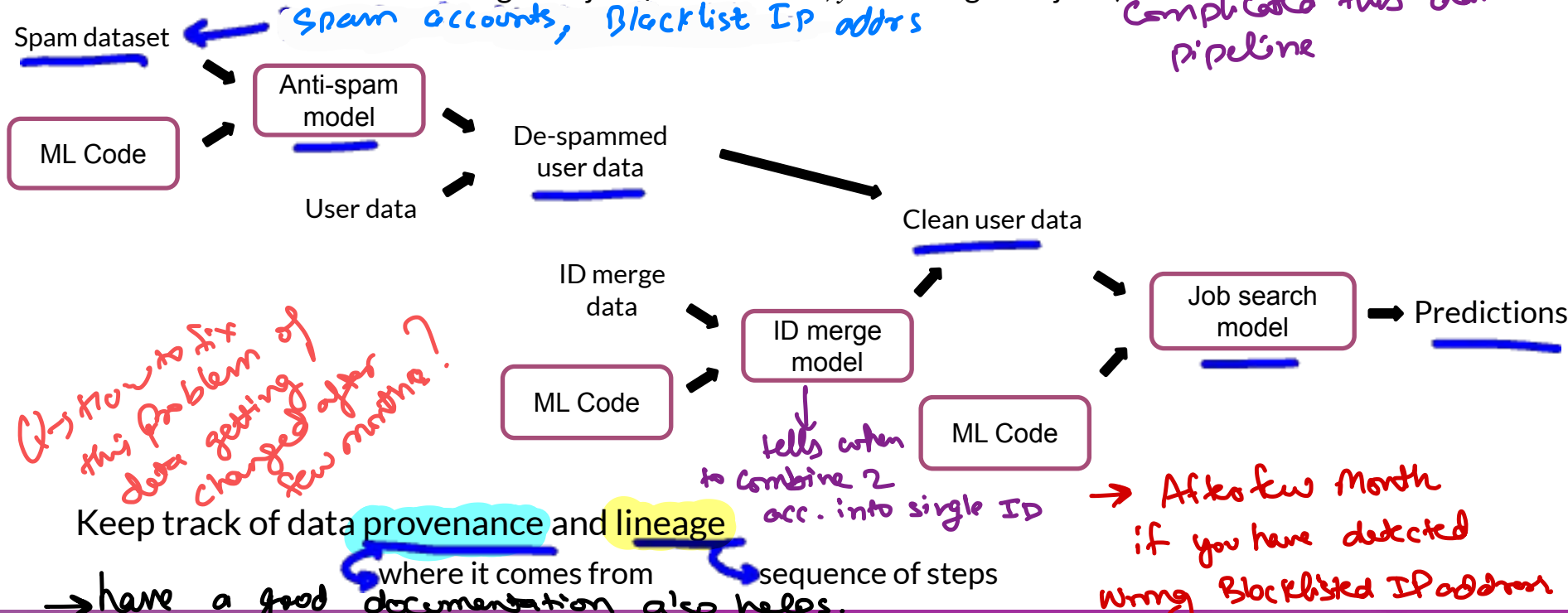
Label and organize data

Meta-data, data provenance and lineage

→ tracking this can help!

Data pipeline example

Task: Predict if someone is looking for a job. (x = user data, y = looking for a job?)



→ Make sure System is maintainable.

Meta-data ← Use this as much. → This helps in error analysis.

e.g Line 17, Lookup[2] generate a lot of logs.

Examples: how, where, when you collected data.

Manufacturing visual inspection: Time, factory, line #, camera settings, phone model, inspector ID,....

Speech recognition: Device type, labeler ID, VAD model ID,....

Key of Dic

Useful for:

- Error analysis. Spotting unexpected effects.
- Keeping track of data provenance.

Tools for Provenance & lineage are still immature & Not fully great.

TF transform

DeepLearning.AI

Label and organize data

Balanced train/dev/test splits

Balanced train/dev/test splits in small data problems

Visual inspection example: 100 examples, 30 positive (defective)

Train/dev/test: 60% / 20% / 20%

Random split: 21 / 2 / 7 positive example (by chance - not great)

35% 10% 35%

Want: 18 / 6 / 6 } balanced split (this is good)

30% / 30% / 30%

No need to worry about this with large datasets – a random split will be representative.

DeepLearning.AI

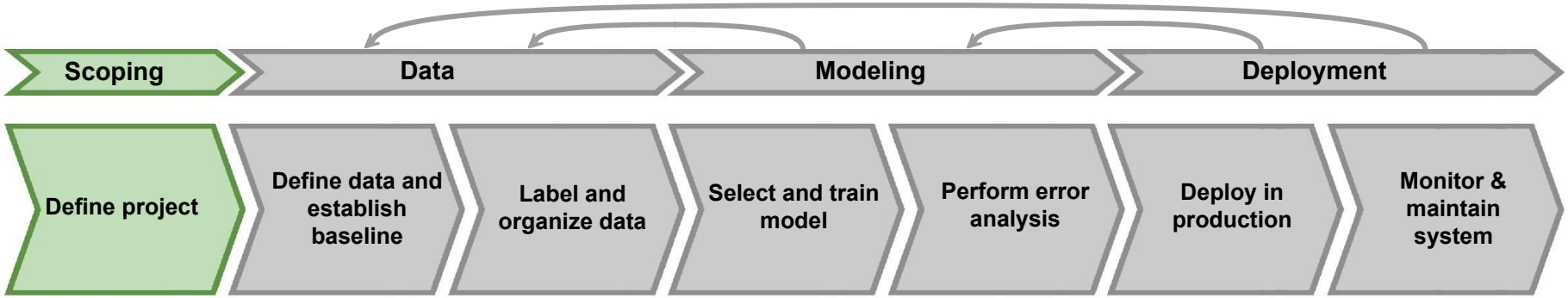
C1W3 Slides (Optional)

Scoping

DeepLearning.AI

Scoping (optional)

What is scoping?



Scoping example: Ecommerce retailer looking to increase sales

- Better recommender system
- Better search
- Improve catalog data
- Inventory management
- Price optimization

Questions:

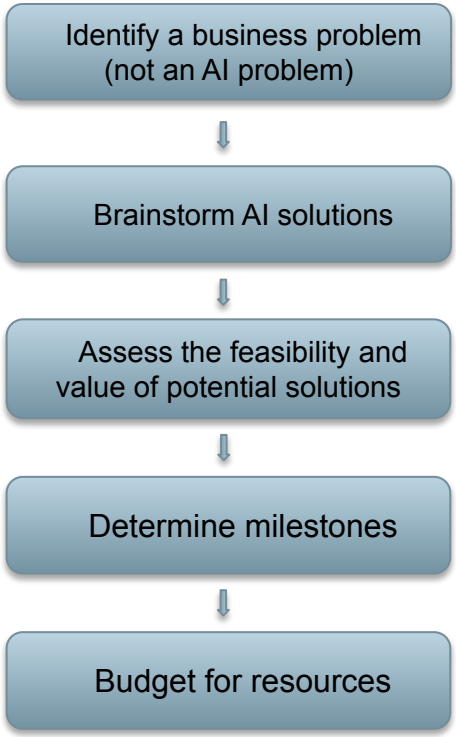
- What projects should we work on?
- What are the metrics for success?
- What are the resources (data, time, people) needed?

DeepLearning.AI

Scoping (optional)

Scoping process

Scoping process



What are the top 3 things you wish were working better?

- Increase conversion
- Reduce inventory
- Increase margin (profit per item)

Separating problem identification from solution

Problem	Solution
Increase conversion	Search, recommendations
Reduce inventory	Demand prediction, marketing
Increase margin (profit per item)	Optimizing what to sell (e.g., merchandising), recommend bundles
What to achieve	How to achieve

DeepLearning.AI

Scoping (optional)

Diligence on feasibility and value

Feasibility: Is this project technically feasible?

Use external benchmark (literature, other company, competitor)

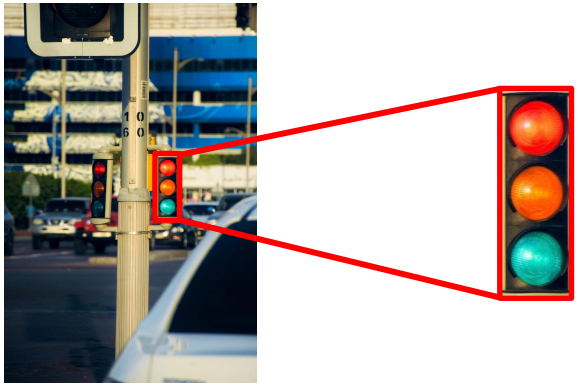
	Unstructured (e.g., speech, images)	Structured (e.g., transactions, records)
New	HLP	Predictive features available?
Existing	HLP History of project	New predictive features? History of project

HLP: Can a human, given the same data, perform the task?

Why use HLP to benchmark?

People are very good on unstructured data tasks

Criteria: Can a human, given the same data, perform the task?



Do we have features that are predictive?

x y

Given past purchases, predict future purchases ✓

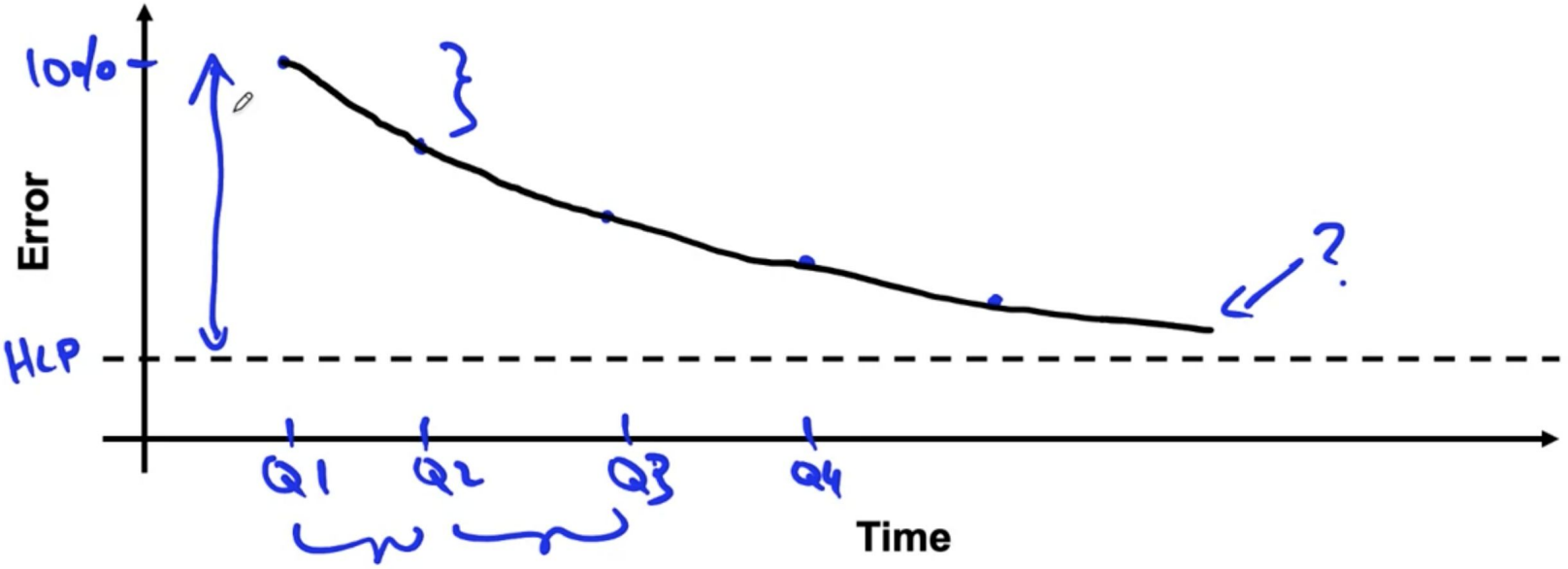
Given weather, predict shopping mall foot traffic ✓

Given DNA info, predict heart disease ?

Given social media chatter, predict demand for a clothing style ?

Given history of a stock's price, predict future price of that stock ✗

History of project



DeepLearning.AI

Scoping (optional)

Diligence on value

Diligence on value

Have technical and business teams try to agree on metrics that both are comfortable with.

Fermi estimates

Ethical considerations

- Is this project creating net positive societal value?
- Is this project reasonably fair and free from bias?
- Have any ethical concerns been openly aired and debated?

DeepLearning.AI

Scoping (optional)

Milestones and resourcing

Milestones

Key specifications:

- ML metrics (accuracy, precision/recall, etc.)
- Software metrics (latency, throughput, etc. given compute resources)
- Business metrics (revenue, etc.)
- Resources needed (data, personnel, help from other teams)

Timeline

If unsure, consider benchmarking to other projects, or building a POC (Proof of Concept) first.

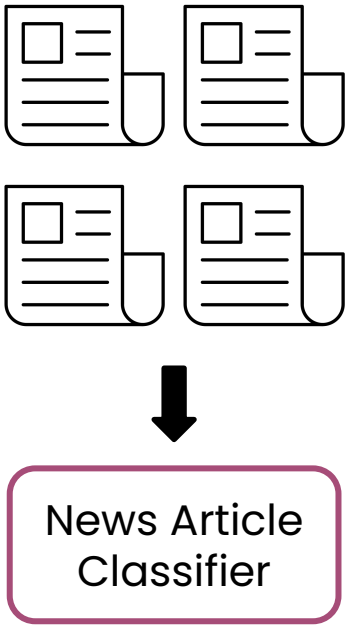
DeepLearning.AI

Final project

Final project overview

Project overview

- Classify news articles
- Start from an existing prototype
- Iteratively improve the performance of the system



Techniques

- Establish a baseline
- Balanced train/dev/test split
- Error analysis
- Track experiments
- Deploy using Tensorflow Serving

DeepLearning.AI Community
community.deeplearning.ai/c/ai-projects/