Copyright Notice

C1W3 Slides

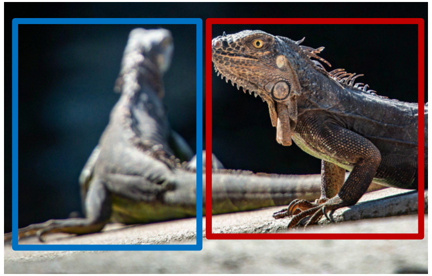
Define data and establish baseline

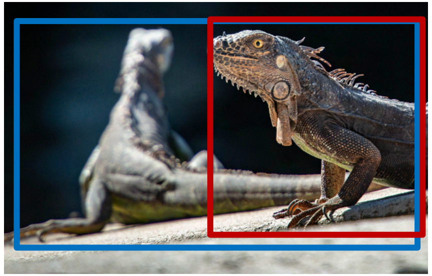
Why is data definition hard?

DeepLearning.AI

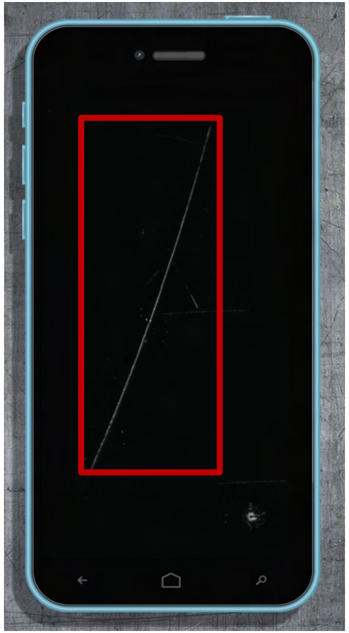
Iguana detection example

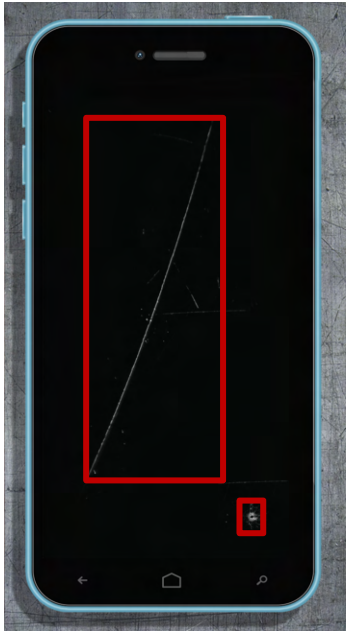
Labeling instructions: "Use bounding boxes to indicate the position of iguanas"

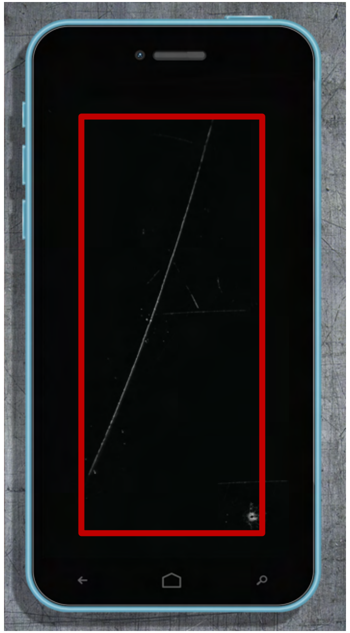




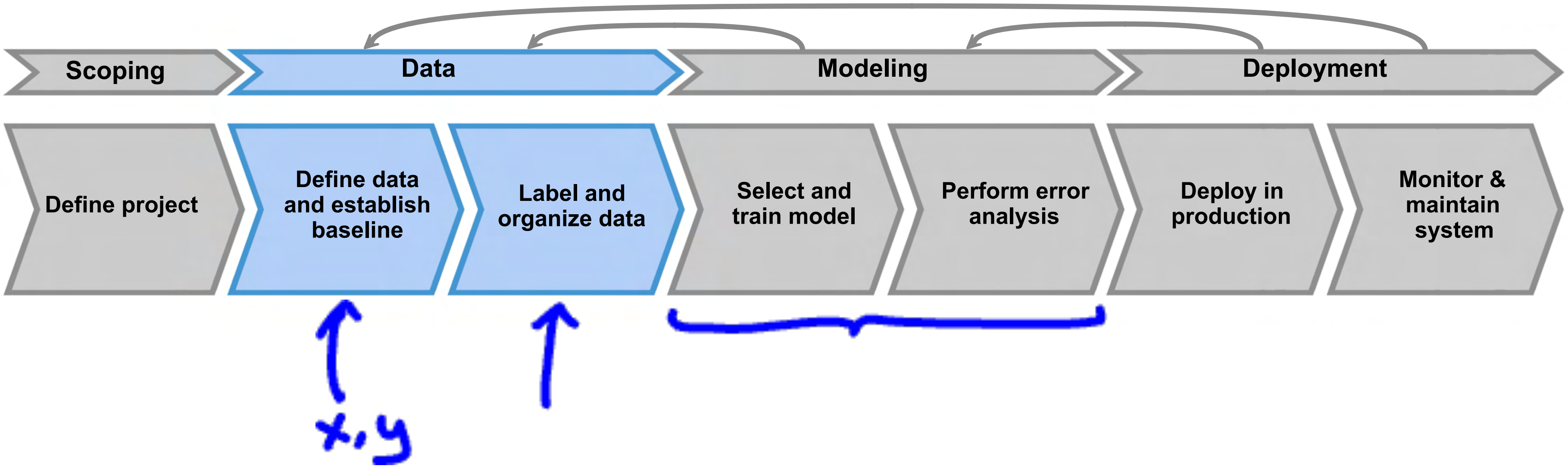
Phone defect detection







Data stage


Scoping
Data
Modeling
Deployment
Define project
Define data and establish baseline
Label and organize data
Select and train model
Perform error analysis
Deploy in production
Monitor & maintain system
x, y

Define data and establish baseline

More label ambiguity examples

Speech recognition example

"Um, nearest gas station"

"Umm, nearest gas station"

”Nearest gas station [unintelligible]"

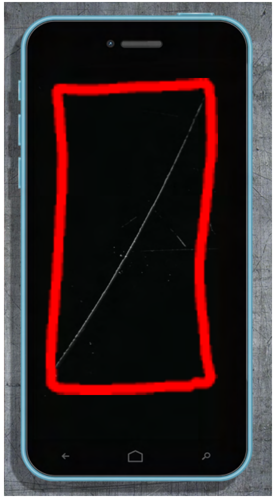
User ID merge example

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

	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

$\begin{cases} \text{1 if same} \\ \text{0 if different} \end{cases}$

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?

DeepLearning.AI

Define data and establish baseline

Major types of data problems

Major types of data problems

	Unstructured	Structured		
Small data	Manufacturing visual inspection from 100 training examples	Housing price prediction based on square footage, etc. from 50 training examples	≤10,000	Clean labels are critical.
Big data	Speech recognition from 50 million training examples	Online shopping recommendations for 1 million users	>10,000	Emphasis on data process.

Humans can label data.

Data augmentation.

Harder to obtain more data.

Unstructured vs. structured data

Unstructured data

- May or may not have huge collection of unlabeled examples x.
- Humans can label more data.
- 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).

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.

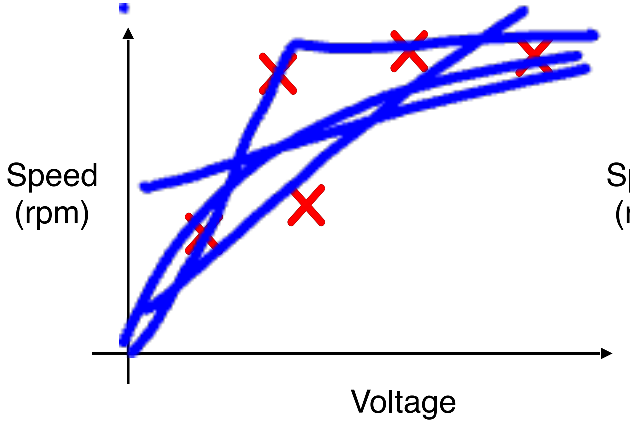
Big data

- Emphasis data process.

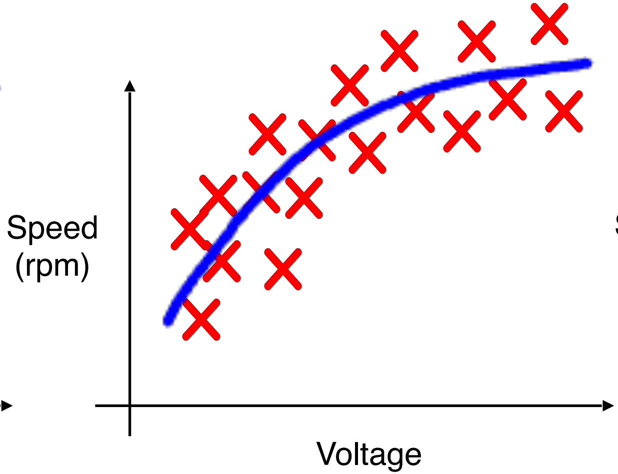
Define data and establish baseline

Small data and label consistency

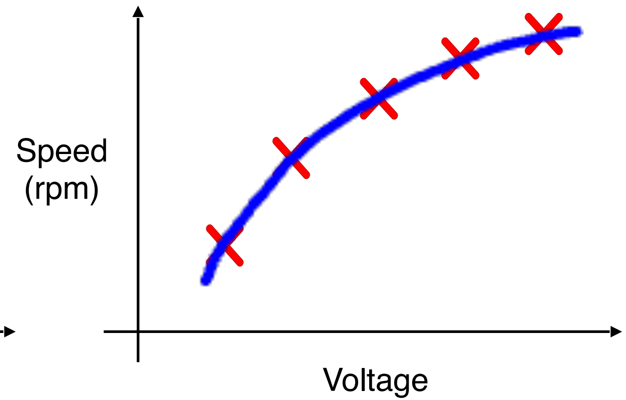
Why label consistency is important



- Small data
- Noisy labels

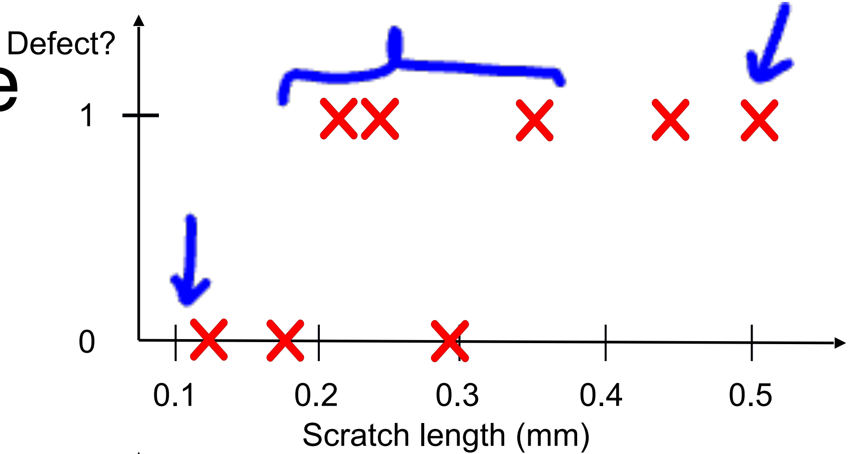


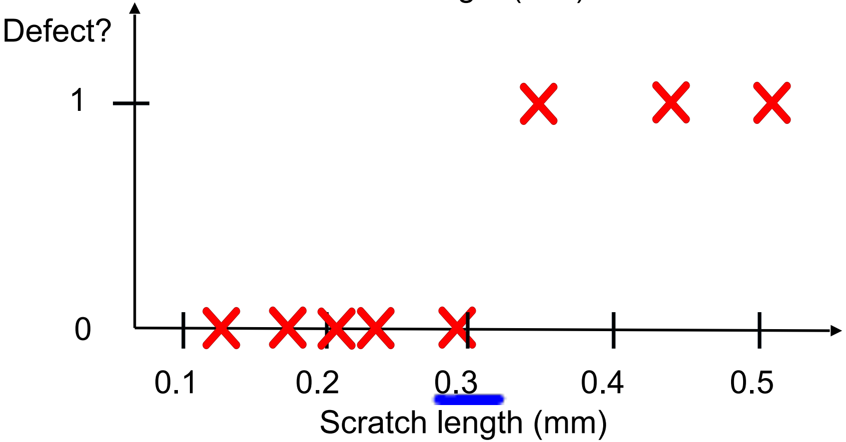
- Big data
- Noisy labels



- Small data
- Clean (consistent) labels

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

Define data and establish baseline

Improving label consistency

Improving label consistency

- 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.
- Iterate until it is hard to significantly increase agreement.

Examples

- Standardize labels

"Um, nearest gas station"

"Umm, nearest gas station" ➡ "Um, nearest gas station"

”Nearest gas station [unintelligible]”

- Merge classes

Deep scratch

Shallow scratch ➡ Scratch

Have a class/label to capture uncertainty

- Defect: 0 or 1

Alternative: 0, Borderline, 1

- Unintelligible audio

“nearest go”

“nearest grocery”

“nearest [unintelligible]”

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.

Define data and establish baseline

Human level performance (HLP)

DeepLearning.AI

Why measure HLP?

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

99%

Ground Truth Label	Inspector
1	1 ✓
1	0 ✗
1	1 ✓
0	0 ✓
0	0 ✓
0	1 ✗

Human?

66.7% accuracy

Other uses of HLP

- In academia, establish and beat a respectable benchmark to support publication.
- 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.

× Use with caution

The problem with beating HLP as a “proof” of ML “superiority”

ML

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

"Um, nearest gas station" ← 30%

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

ML agrees with humans: 0.70

+12%

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

Define data and establish baseline

Raising HLP

DeepLearning.AI

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.

Ground Truth Label	Inspector
1	1
~~1~~ 0	0
1	1
0	0
0	0
0	~~1~~ 0

0.3mm

66.7%

↓

100%

Raising HLP

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

- Improving label consistency will raise HLP.

- 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.

HLP on structured data

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

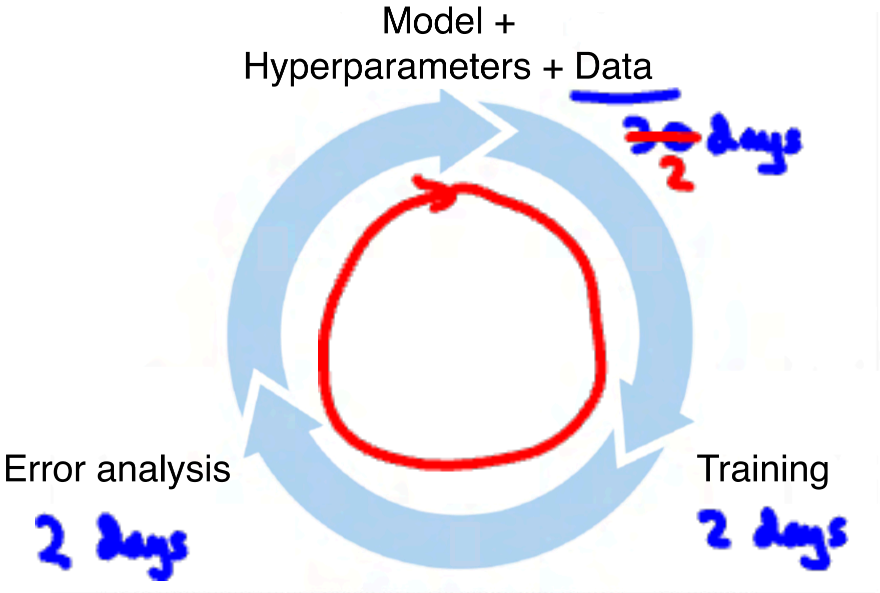
Some exceptions:

- User ID merging: Same person?
- 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?

Label and organize data

Obtaining data

How long 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
Owned	100h	$0
Crowdsourced – Reading	1000h	$10000
Pay for labels	100h	$6000
Purchase data	1000h	$10000

Other factors: Data quality, privacy, regulatory constraints

Labeling data

- 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

Label and organize data

Data pipeline

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

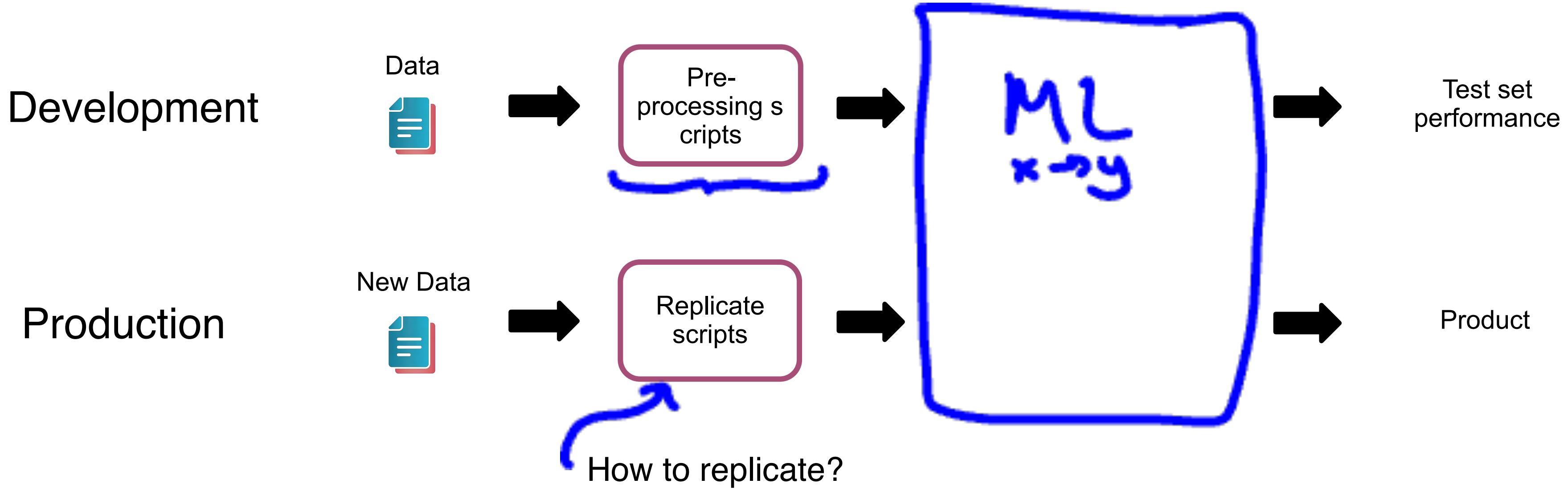
y = looking for job

Raw data → Data cleaning (spam cleanup → user ID merge) → ML

scripts

predict y

Data pipeline example



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.

Production phase:

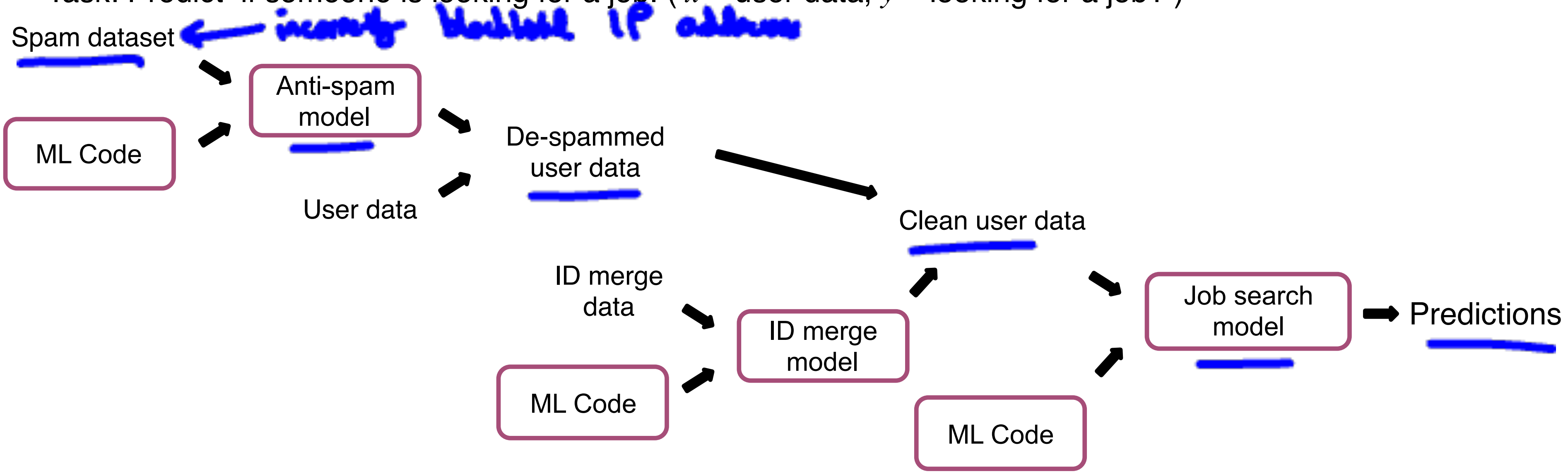
- After project utility is established, use more sophisticated tools to make sure the data pipeline is replicable.
- E.g., TensorFlow Transform, Apache Beam, Airflow,….

Label and organize data

Meta-data, data provenance and lineage

Data pipeline example

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



Keep track of data provenance and lineage

where it comes from

sequence of steps

Meta-data

line 17, factory 2

Examples:

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

xyz

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

Useful for:

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

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

35% 10% 35%

Want: 18/6/6

30%/30%/30% } balanced split

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