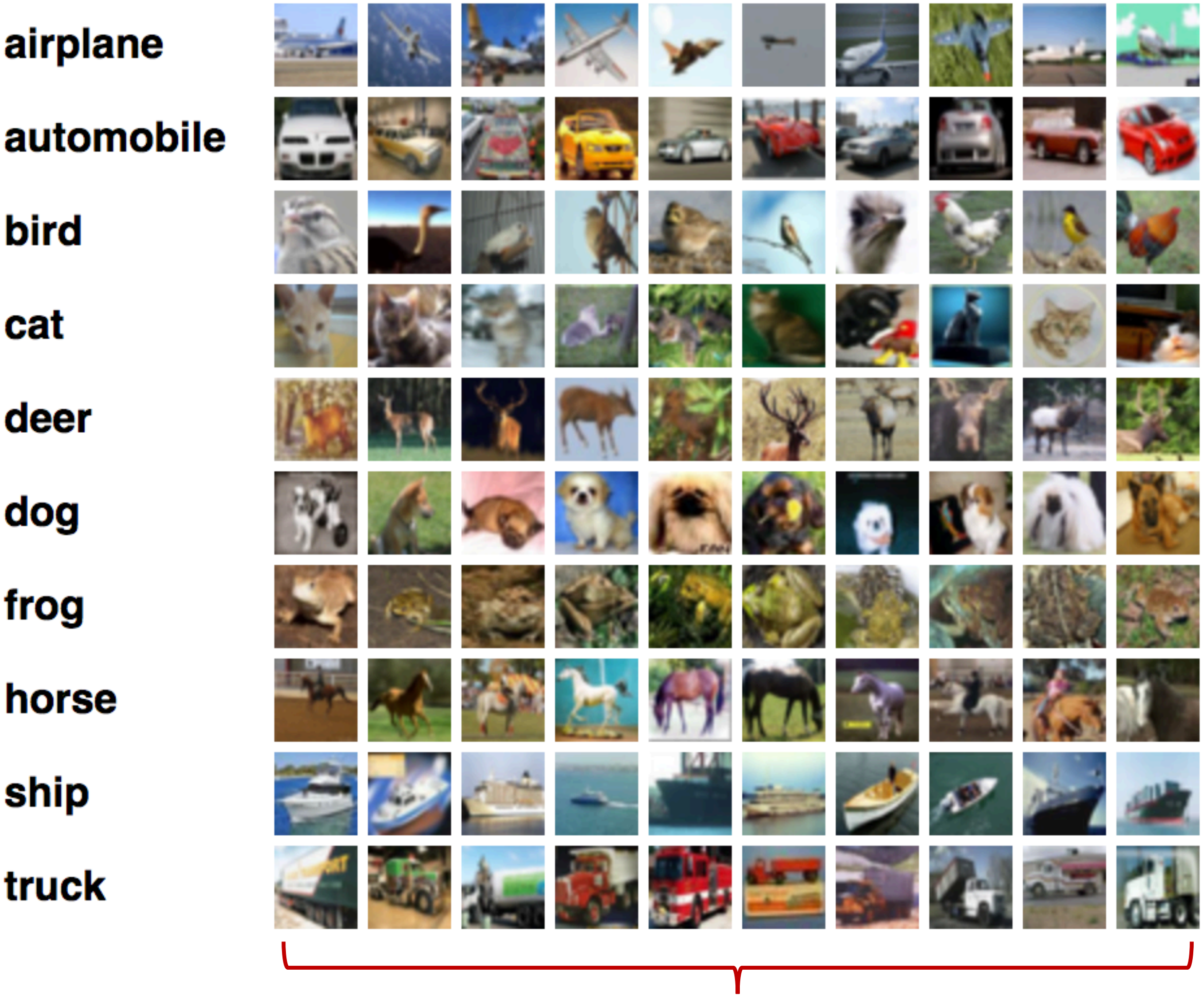
Data Collection

INFO-4604, Applied Machine Learning

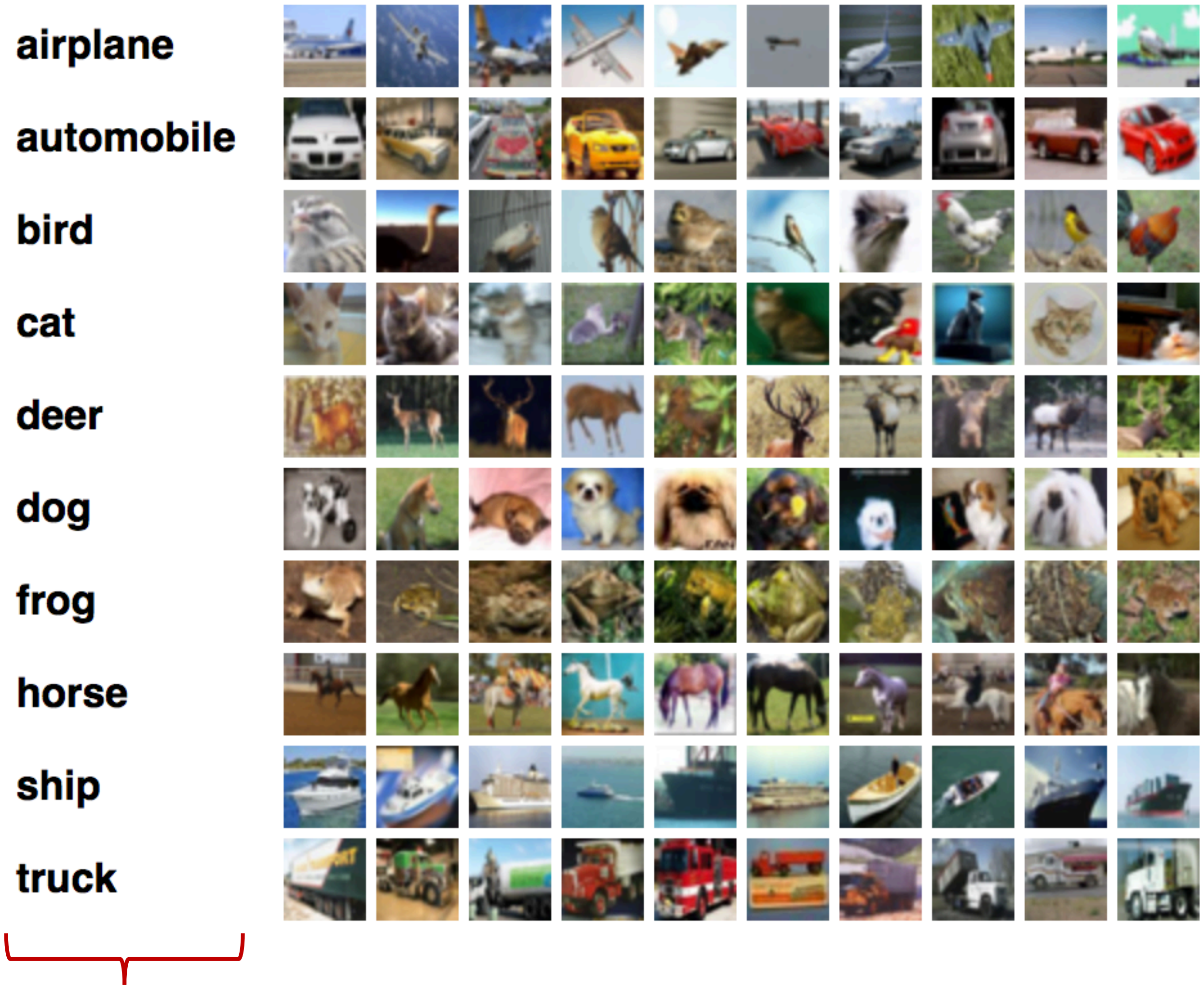
University of Colorado Boulder

October 19, 2017

Prof. Michael Paul



Where did these images come from?



Where did these labels come from?

Collecting a Dataset

- What are you trying to do?
- Where will the instances come from?
- What should your labels be?
- Where can you get labels?
    - They might already exist
    - You might be able to approximate them from something that exists
    - You might have to manually label them

Define the Task

What is the prediction task? Might seem obvious, but important things to think about:

- Output discrete or continuous?
    - Some tasks could be either, and you have to decide
    - e.g., movie recommendation:
        - how much will you like a movie on a scale from 1-5?
        - will you like a movie, yes or no?
- How fine-grained do you need to be?
    - Image search: probably want to distinguish between “deer” and “dog”
    - Self-driving car: know that an animal walked in front of the car; not so important what kind of animal

Get the Data

Won’t go too much into obtaining data in this class.

Sometimes need to do some steps to get data out of files:

- Convert PDF to text
- Convert plots to numbers
- Parse HTML to get information

Label the Data

Training data needs to be labeled! How do we get labels?

One of the most important parts of data collection in machine learning

- Bad labels → bad classifiers
- Bad labels → misleading evaluation

Good labels can be hard to obtain – needs thought

Label the Data

Some data comes with labels

- Or information that can be used as approximate labels

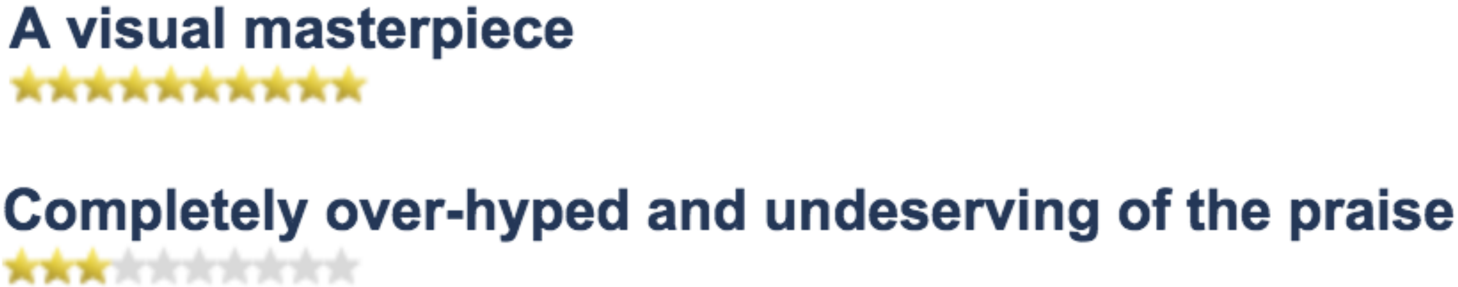
Sometimes you need to create labels

- Data **annotation**

Label the Data

Example: sentiment analysis

IMDb



A visual masterpiece

★★★★★★★★★★

Completely over-hyped and undeserving of the praise

★★★☆☆☆☆☆☆☆

The reviews already comes with scores that indicate the reviewers' sentiment

Label the Data

Example: sentiment analysis

Positive	I really enjoyed Blade Runner 2049 & more than any movie in a long time I recommend you see it on a huge, loud screen.
Negative	gosling was the best part of blade runner 2049, but even then it’s his worst role in his career. his character is bland as a piece of bread.

These tweets don’t come with a rating, but a person could read these and determine the sentiment

Label the Data

Example: sentiment analysis

Thought: could you train a sentiment classifier on IMDB data and apply it to Twitter data?

- Answer is: maybe
- Sentiment will be similar – but there will also be differences in the text in the two sources
- We'll revisit this later in the semester (if time)

Labeled Data

Let’s start by considering cases where we don’t have to create labels from scratch.

Labeled Data

A lot of *user-generated content* is labeled in some way by the user

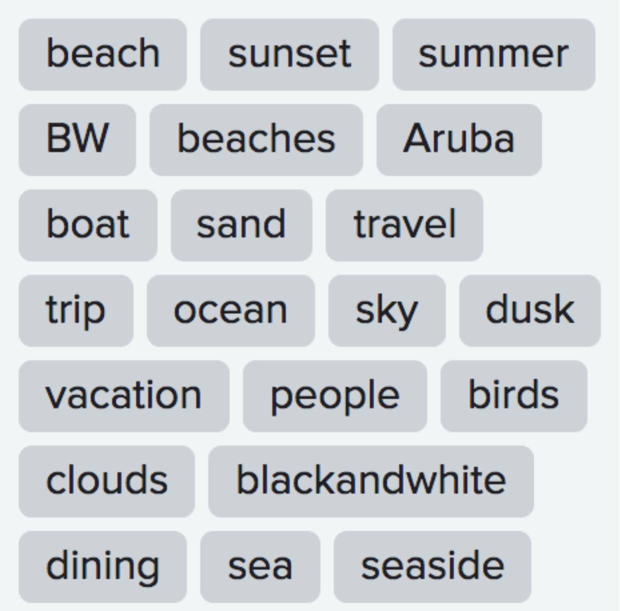
- Ratings in reviews
- Tags of posts/images

Usually correct, but:

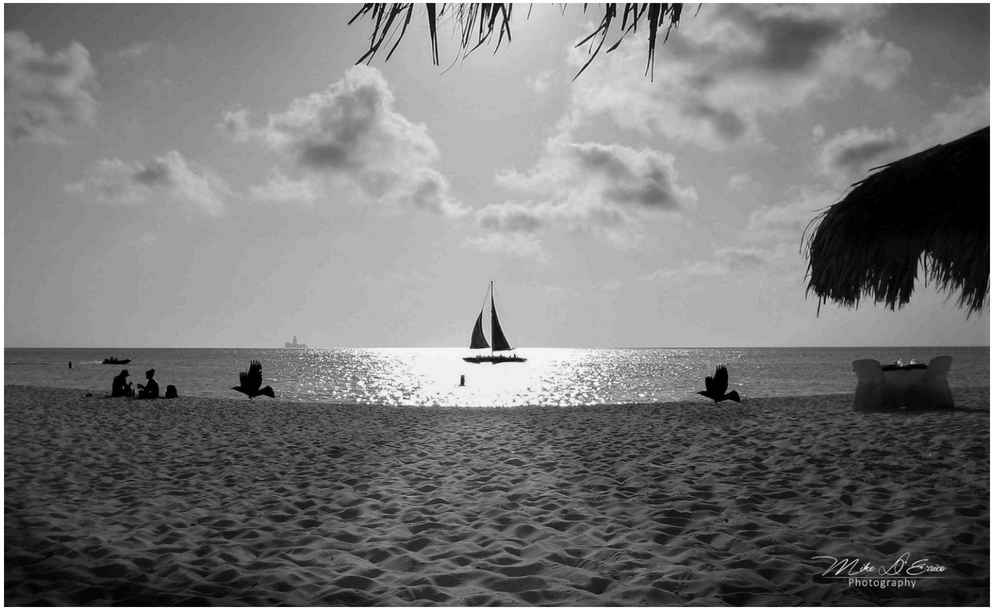
- Sometimes not what you think
- May be incomplete
- Variation in how different users rate/tag things

Labeled Data

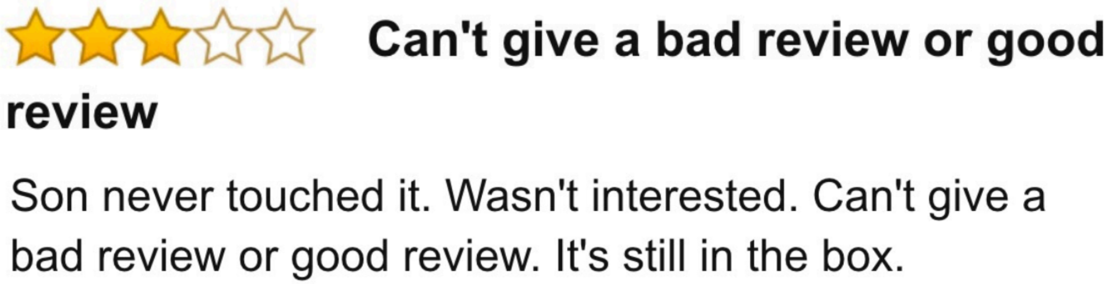
flickr



beach sunset summer
BW beaches Aruba
boat sand travel
trip ocean sky dusk
vacation people birds
clouds blackandwhite
dining sea seaside

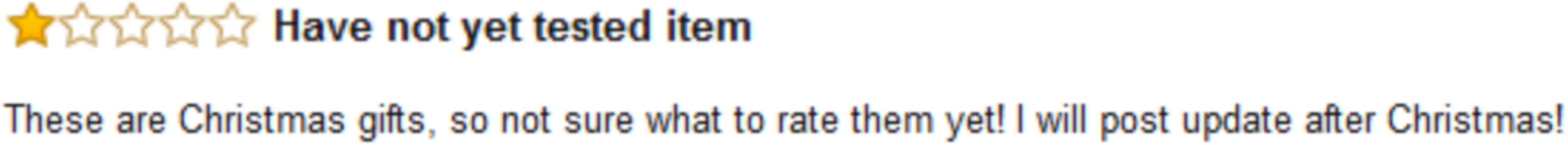


Labeled Data



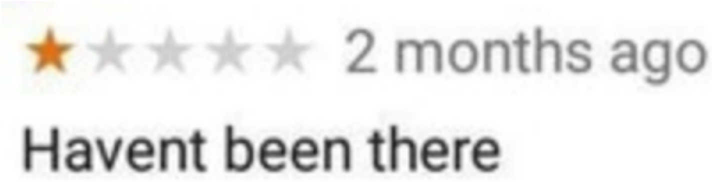
★★★☆☆ **Can't give a bad review or good review**

Son never touched it. Wasn't interested. Can't give a bad review or good review. It's still in the box.



★☆☆☆☆ **Have not yet tested item**

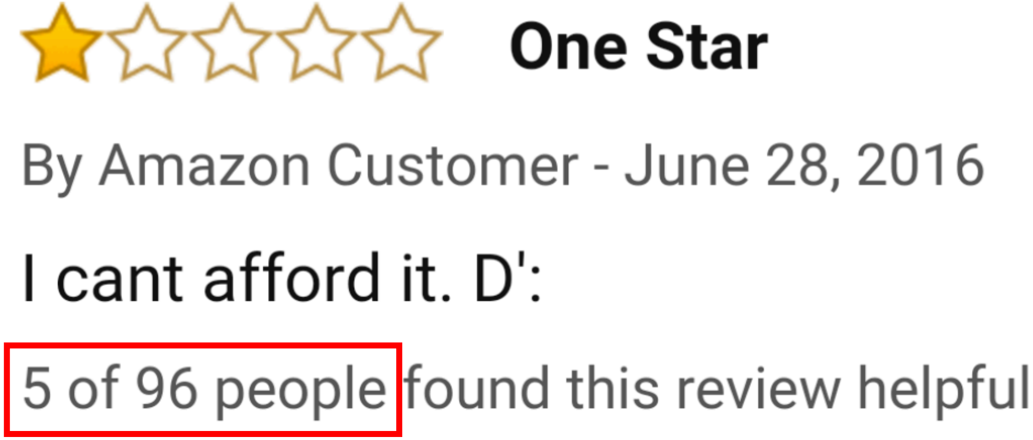
These are Christmas gifts, so not sure what to rate them yet! I will post update after Christmas!



★★★★★ 2 months ago

Havent been there

Labeled Data



★☆☆☆☆ **One Star**

By Amazon Customer - June 28, 2016

I cant afford it. D':

5 of 96 people found this review helpful

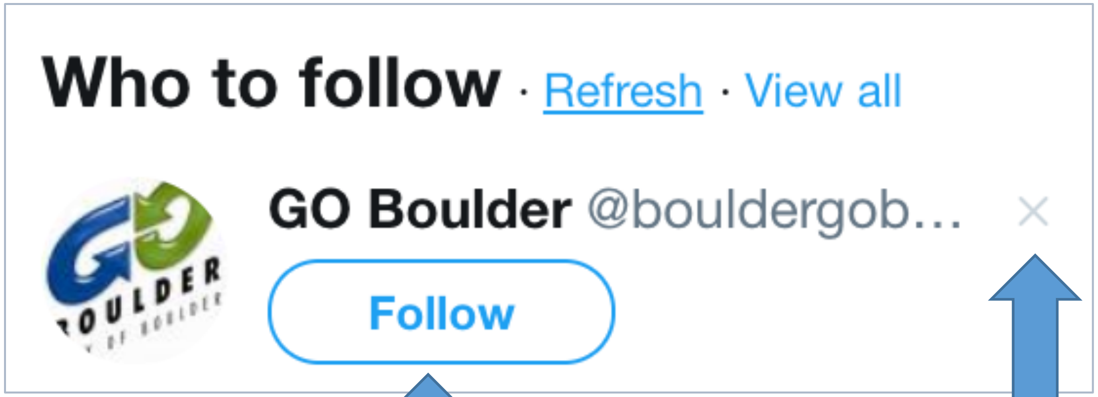
Some data also comes with ratings of the quality of content, which you could leverage to identify low-quality instances

Implicit Labels

Many companies can use user *engagement* (e.g., clicks, “likes”) with a product as a type of label

Often this type of feedback is only a proxy for what you actually want, but it is useful because it can be obtained without effort

Implicit Labels



Who to follow · Refresh · View all

GO Boulder @bouldergob…

Clicking this signals that this was a good recommendation

Clicking this signals that this was a bad recommendation

Not clicking doesn’t signal anything either way

Implicit Labels

Reasons you might “like” a post:

- You liked the content
- You want to show the poster that you liked it (even if you didn’t)
- You want to make it easier to find later (maybe because you hated it)

Might be wrong to assume “likes” would be good training data for predicting posts you will like

- But maybe a good enough approximation

Implicit Labels

Summary:

- Clicking might not mean what you think
- Not clicking might not mean anything
  - Might be reasonable to use clicks to get ‘positive’ labels, but a lack of click shouldn’t count as a ‘negative’ label

Annotation

Annotation (sometimes called *coding* or *labeling*) is the process of having people assign labels to data by hand.

Annotation can yield high-quality labels since they have been verified by a person.

But can also give low-quality labels if done wrong.

A person doing annotation is called an **annotator**.

Annotation

Sometimes seemingly straightforward:

> “does this image contain a truck?”

Though often annotation becomes less straightforward once you start doing it…

- Is a truck different from a car?
- Is a pickup truck different from a semi-truck?
- Is an SUV a truck?

The answers to these questions will depend on your task (as discussed at the start of this lecture)

Annotation

Need to decide on the set of possible labels and what they mean.

- Need clear guidelines for annotation, otherwise annotator(s) will make inconsistent decisions (e.g., what counts as a truck)

Often an iterative process is required to finalize the set of labels and guidelines.

- i.e., you’ll start with one idea, but after doing some annotations, you realize you need to refine some of your definitions

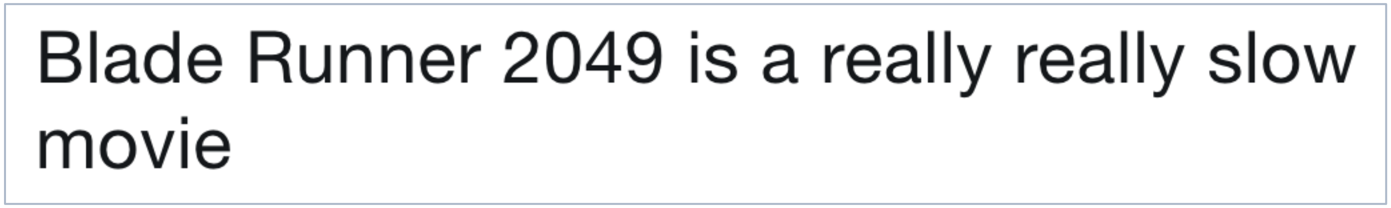
Annotation

Annotation can be hard.

If an instance is hard to label, usually this is either because:

- the definition of the label is ambiguous
- the instance itself is ambiguous
  (maybe not enough information, or intentionally unclear like from sarcasm)

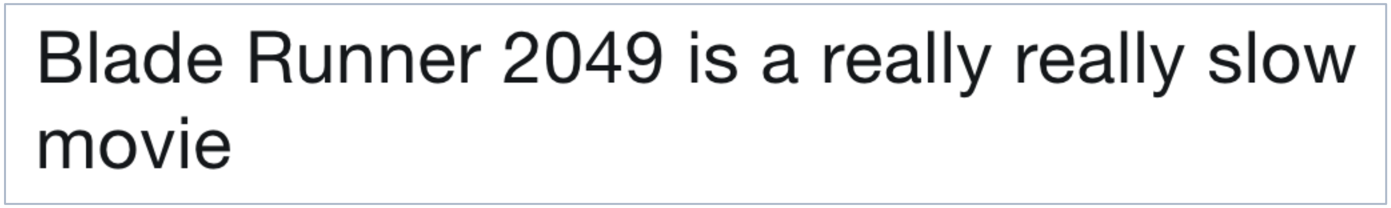
Annotation



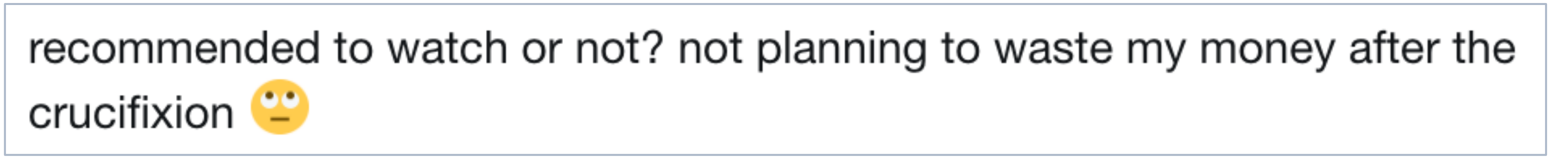
Blade Runner 2049 is a really really slow movie

Does this tweet express negative sentiment?
Maybe?

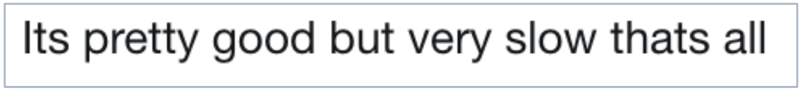
Annotation



Blade Runner 2049 is a really really slow movie



recommended to watch or not? not planning to waste my money after the crucifixion 🙄



Its pretty good but very slow thats all

The ability to annotate an instance depends on how much information is available.

Annotation

What if an instance is genuinely unclear and you don’t know how to assign a label?

One solution: just exclude from training data

- Then classifier won’t learn what to do when it encounters similar instances in the future
- But maybe that’s better than teaching the classifier something that’s wrong

Annotation

What if an instance is genuinely unclear and you don’t know how to assign a label?

Sometimes you might want a special class for ‘other’ / ‘unknown’ / ‘irrelevant’ instances

- For some tasks, there is a natural class for instances that do not clearly belong to another; e.g., sentiment classification should have a ‘neutral’ class

Crowdsourcing

Annotation can be slow. What if you want thousands or tens of thousands of labeled instances?

Crowdsourcing platforms (e.g., Amazon Mechanical Turk, CrowdFlower) let you outsource the work to strangers on the internet

- Split the work across hundreds of annotators

Crowdsourcing

Harder to get accurate annotations for a variety of reasons:

- You don’t have the same people labeling every instance, so harder to ensure consistent rules
- Crowd workers might lack the necessary expertise
- Crowd workers might work too quickly to do well

Crowdsourcing

Usually want 3-5 annotators per instance so that if some of them are wrong, you have a better chance of getting the right label after going with the majority

You should also include tests to ensure competency

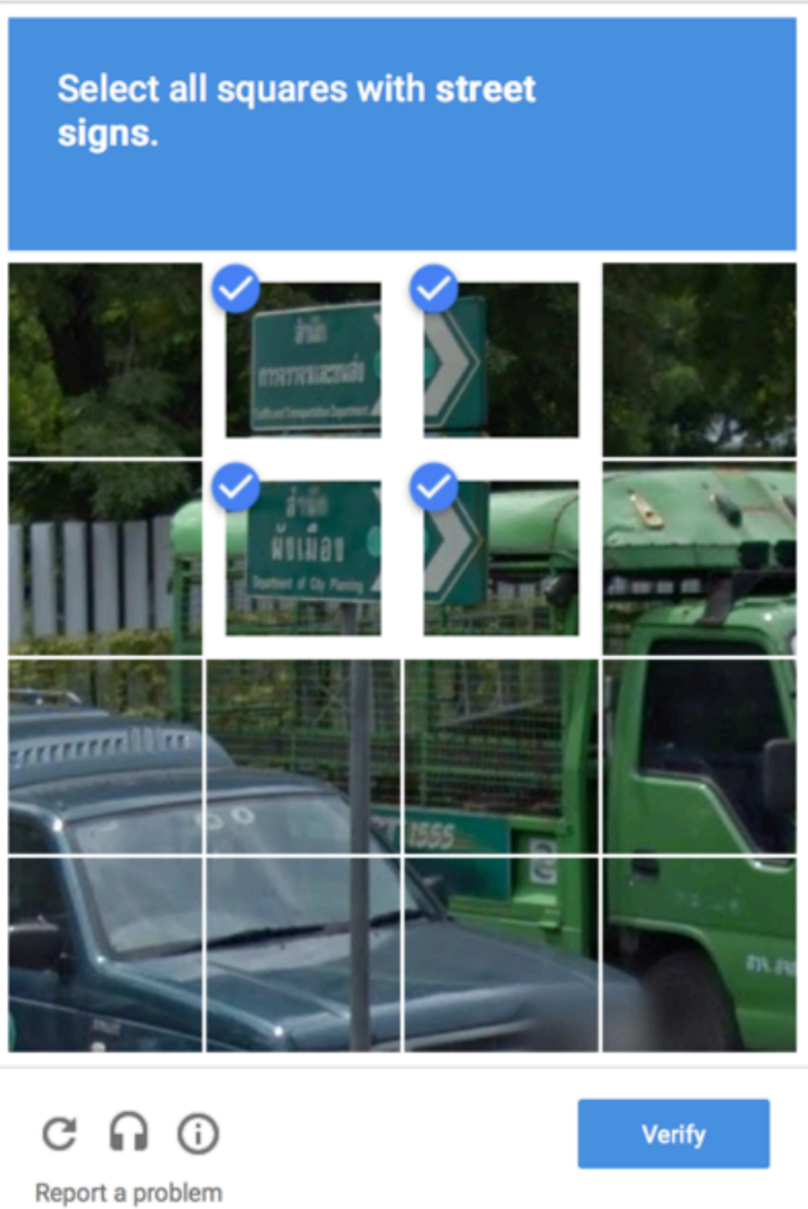
- Sometimes have an explicit test at the start, to check their expertise
- Can include “easy” examples mixed with the annotation tasks to see if they get them correct

Crowdsourcing

- There are entire courses on crowdsourcing; mostly beyond scope of this course
- But fairly common in machine learning
- Many large companies have internal crowdsourcing platforms
  - That way you can crowdsource data that can’t be shared outside the company
  - Maybe higher quality work, though safeguards still a good idea

Crowdsourcing

Other creative ways of getting people to give you labels exist…



Select all squares with street signs.