Cyberbullying Detection Using AI

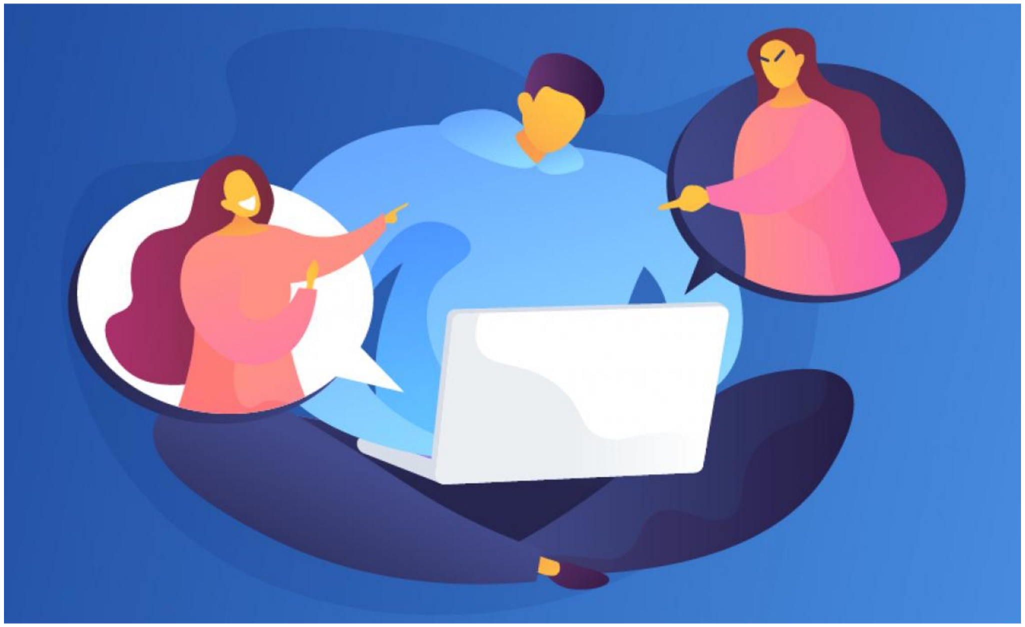
<Instructor>

Outline

- The Critical Problem of Cyberbullying
- Processes in AI Development
- Data Collection
- Annotation
- Training
- Evaluation
- Deployment
- Q&A

What is Cyberbullying?

- Different from bullying, CYBERBULLYING is bullying with the use of **Digital Technologies**.



What is Cyberbullying?

- Cyberbullying is bullying with the use of **Digital Technologies**
  - Social media
  - Messaging platforms
  - Gaming platforms
  - Mobile phones

What is Cyberbullying?

- Cyberbullying is bullying with the use of **Digital Technologies**

- It is repetitive behavior, aimed at scaring, angering or shaming people who are targeted.

Common types of cyberbullying

- **Spreading** lies about or posting embarrassing photos or videos of someone

- **Sending** hurtful, abusive or threatening messages, images or videos to someone

- **Impersonating** someone and sending mean messages to others

The Critical Problem of Cyberbullying

I hate you! I dislike these people because… You are an idiot, ... Get out from my house! You are not my friend anymore… 😠

Based on words and phrases **hate, dislike, ugly, get out …**

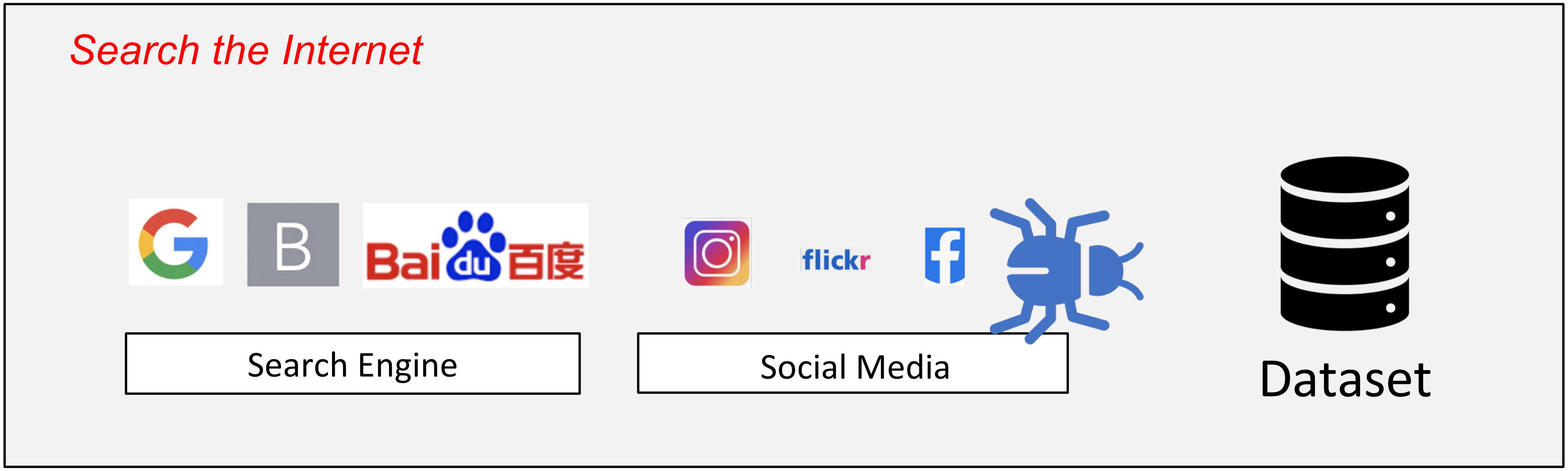
State-of-the-art detectors Google clarifai Amazon Rekognition

Actively researched problem **Psychology** **Sociology** **Computer science**

Process of AI Development

- Data Collection
  - We need dataset for training AI
- Annotation
  - We need to label dataset
- Training
  - AI training process
- Evaluation
  - How good are we doing?
- Deployment
  - Detect on real-world samples

Dataset Collection



Search the Internet

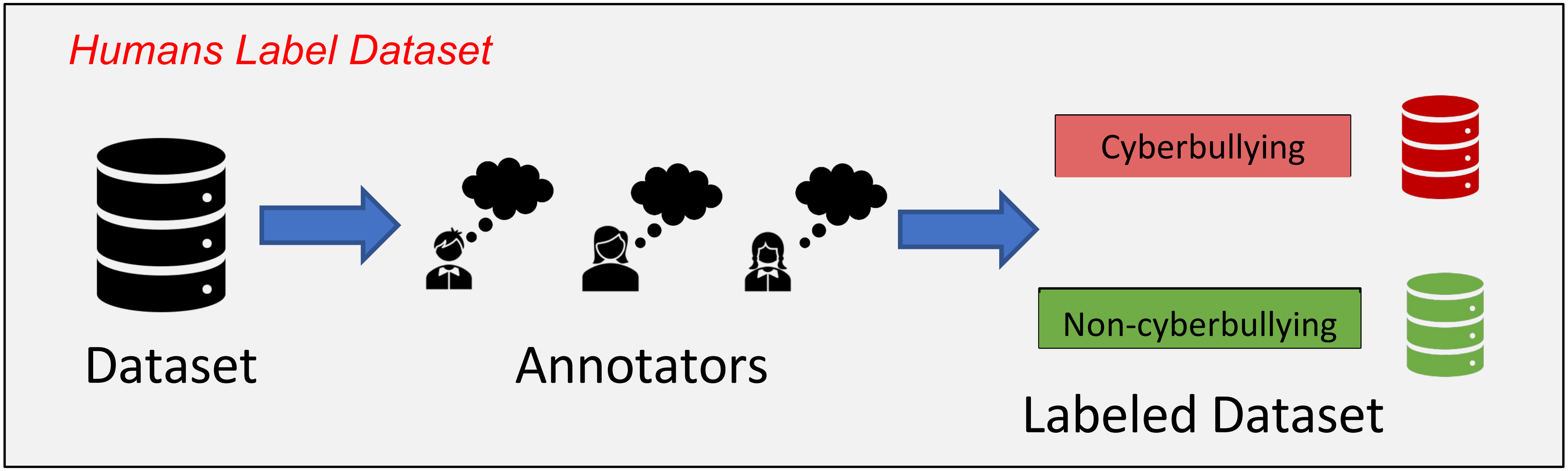
Search Engine

Social Media

Dataset

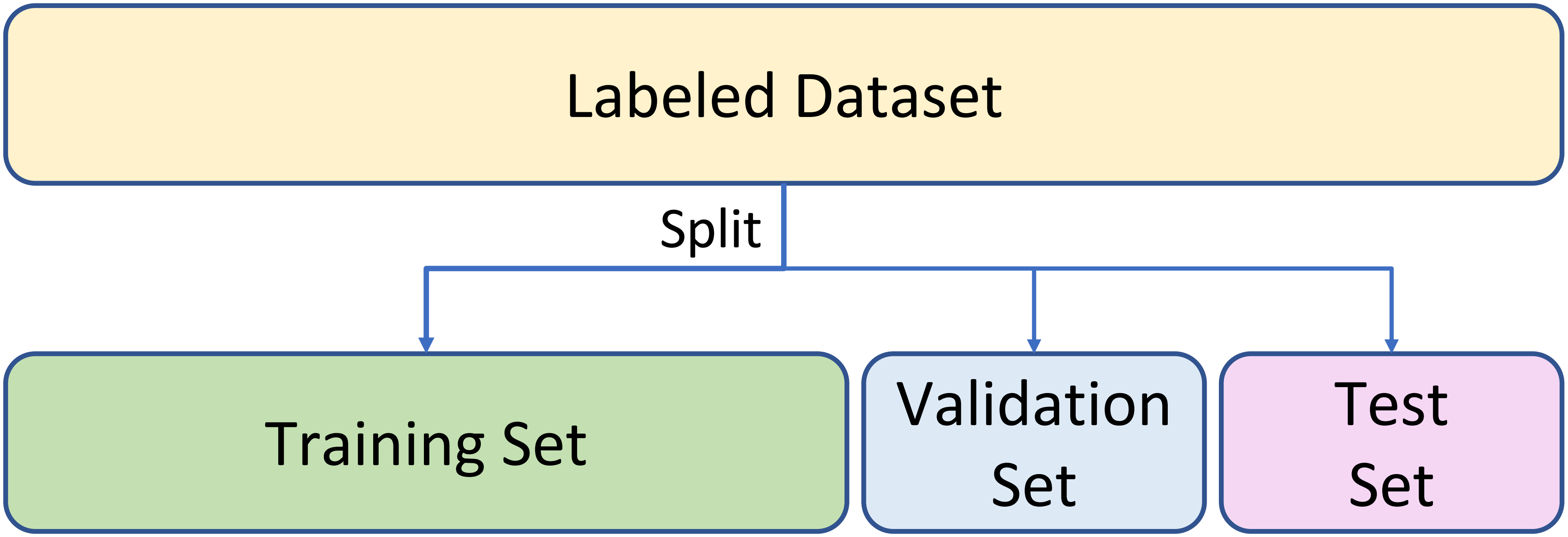
Cyberbullying dataset is from Formspring

Annotation



Humans Label Dataset

Training


Labeled Dataset

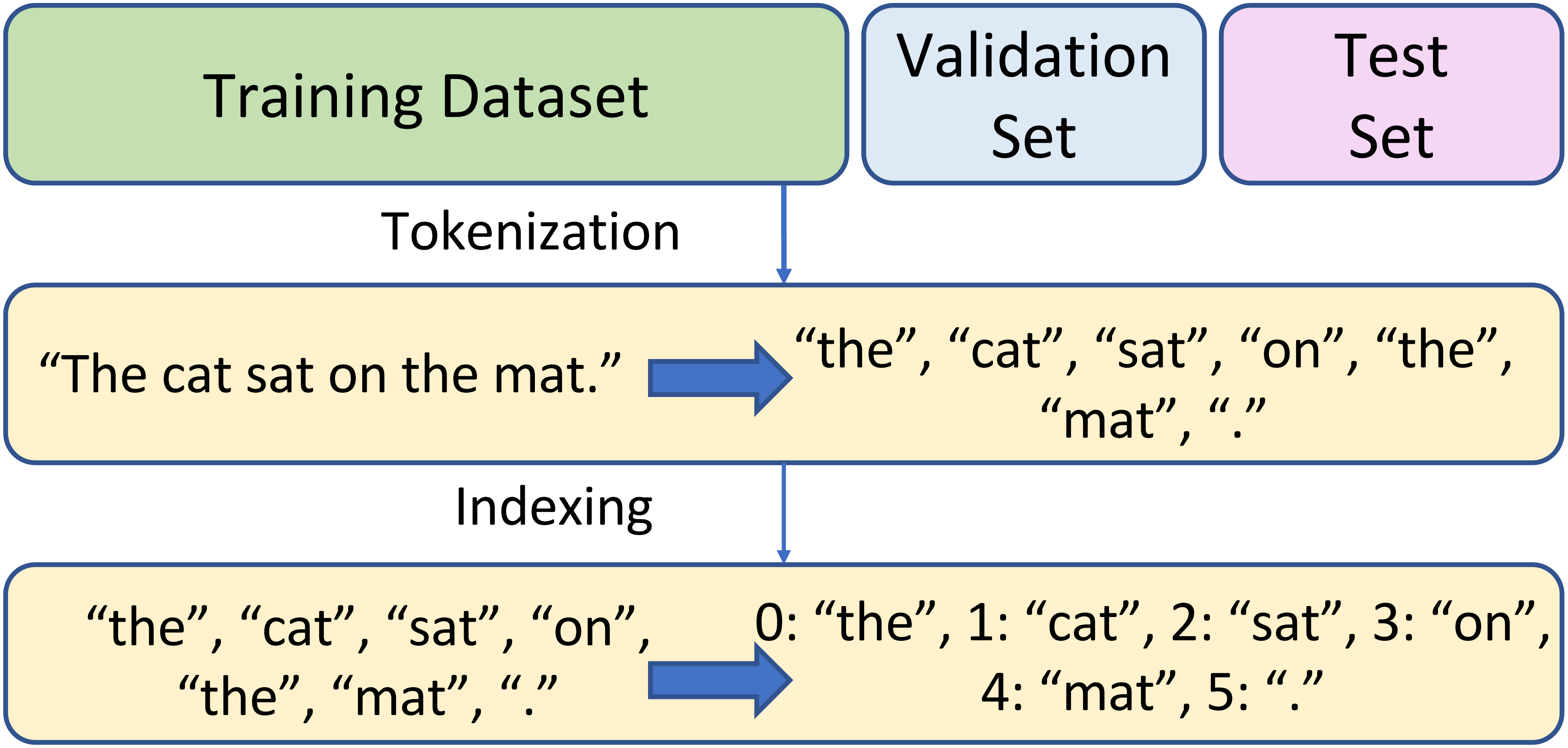
Split

Training Set

Validation Set

Test Set

Training Cont…



Training Dataset

Validation Set

Test Set

Tokenization

“The cat sat on the mat.”

“the”, “cat”, “sat”, “on”, “the”, “mat”, “.”

Indexing

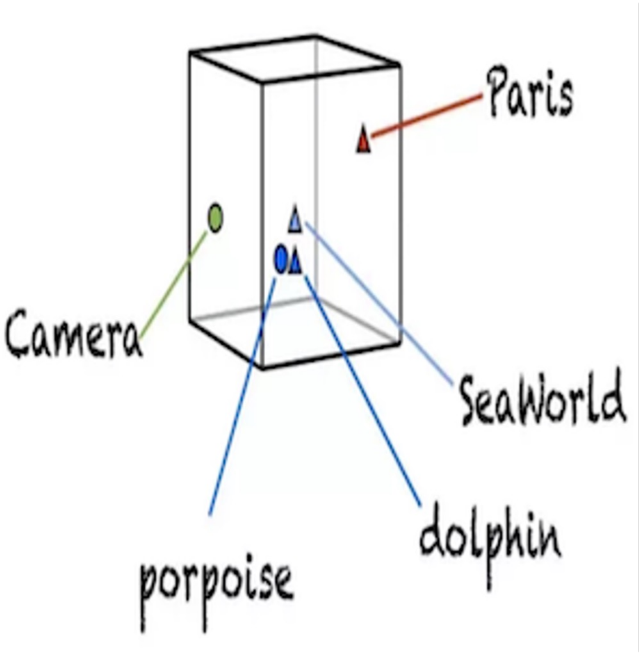
“the”, “cat”, “sat”, “on”, “the”, “mat”, “.”

0: “the”, 1: “cat”, 2: “sat”, 3: “on”, 4: “mat”, 5: “.”

Training Cont...

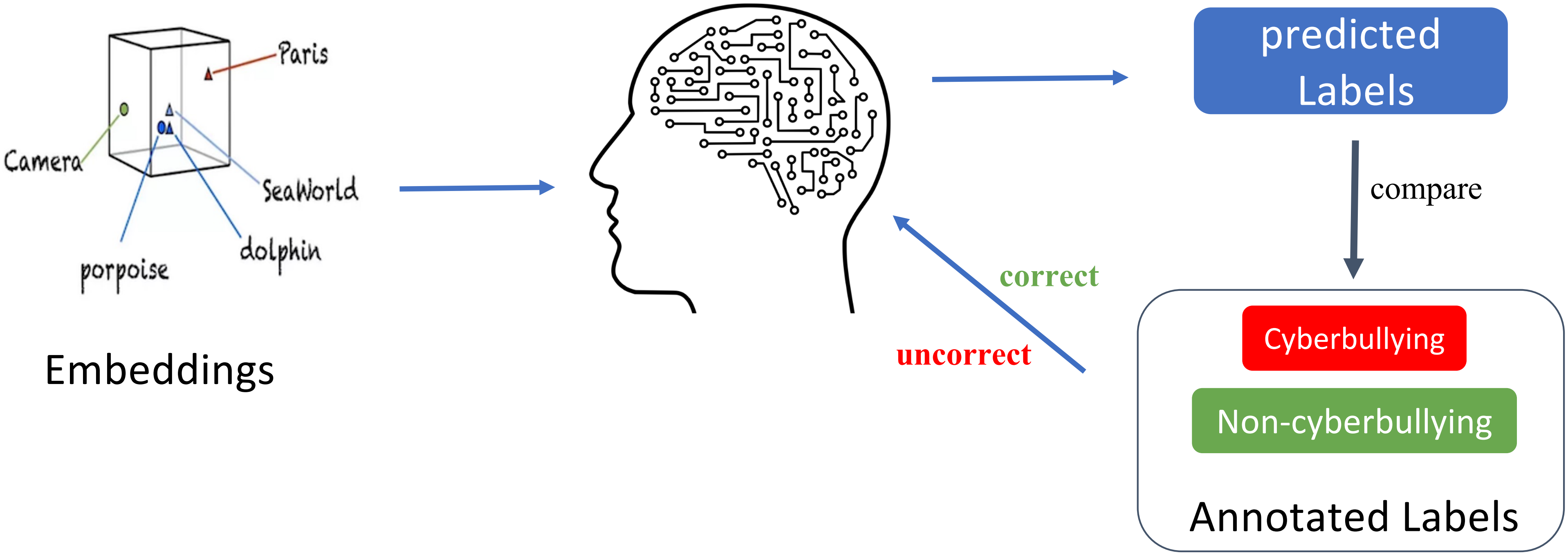
- Paris
- Camera
- dolphin
- SeaWorld
- Porpoise

Tokens



Embeddings

Training Cont…



Embeddings

predicted Labels

compare

correct

uncorrect

Cyberbullying

Non-cyberbullying

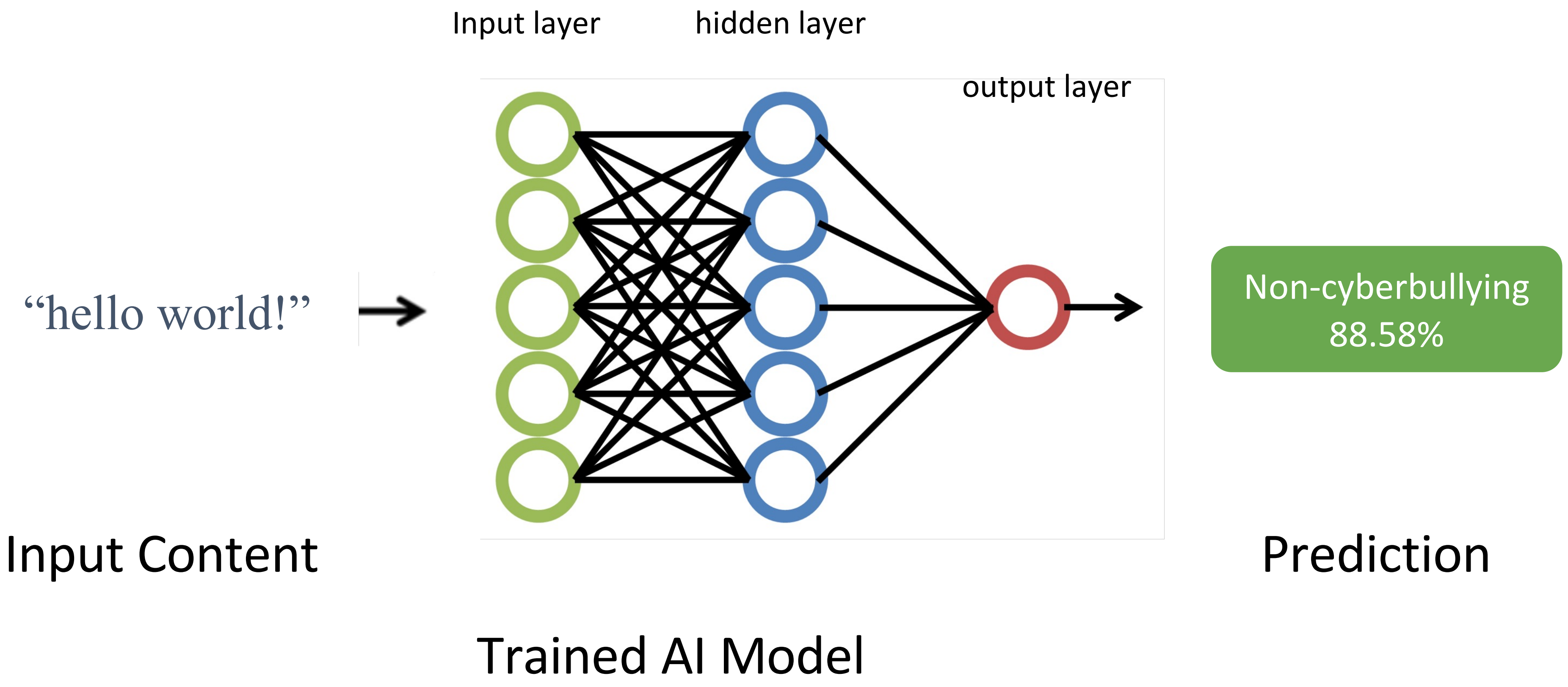
Annotated Labels

AI is Learning …

Training Cont…

- epoch
  - When all training data has been learned by the AI once, this process is called an **epoch**.
- How many epochs should we use?
  - **Unfortunately**, there is no correct answer to this question.
  - For different datasets, the answer is not the same. We still need to try different epochs number to figure it out.

Evaluation



Input layer

hidden layer

output layer

“hello world!”

Non-cyberbullying 88.58%

Input Content

Prediction

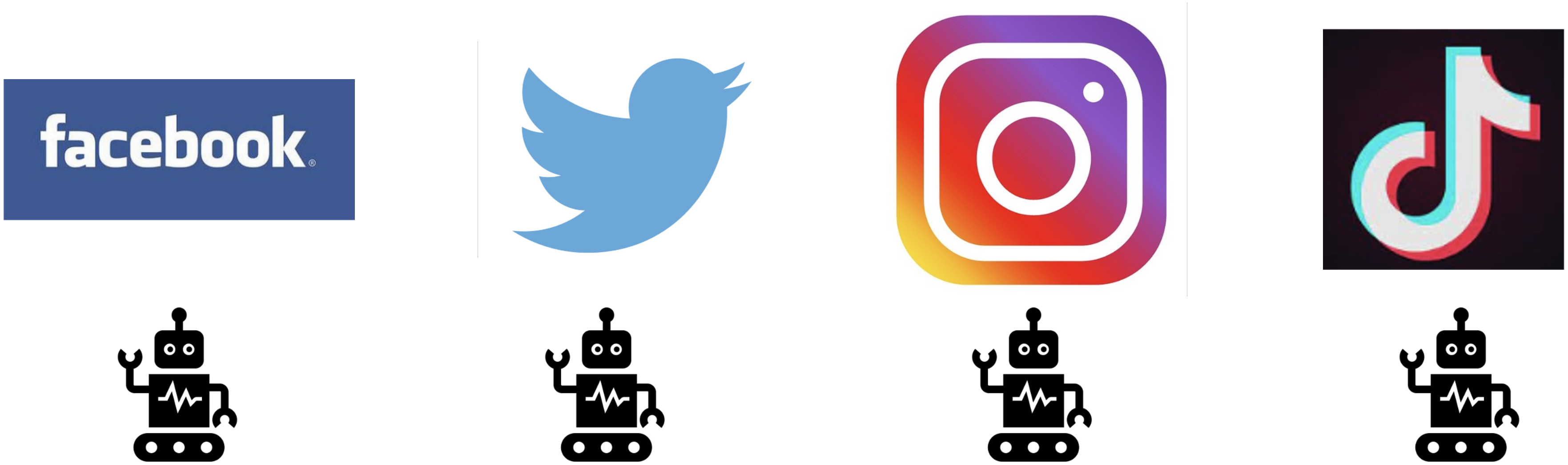
Trained AI Model

Evaluation Cont…

- Accuracy
  - One simple way we measure an AI model’s performance with test dataset
  - Accuracy = the percentage of the correct predictions

$$Accuracy = \frac{\#\ of\ correct\ predictions}{\#\ of\ all\ predictions}$$

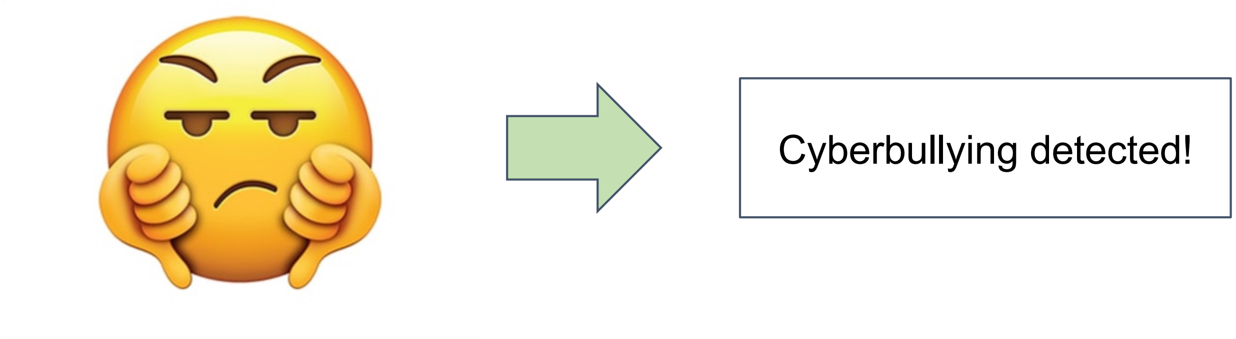
Deployment



Even More Labs!

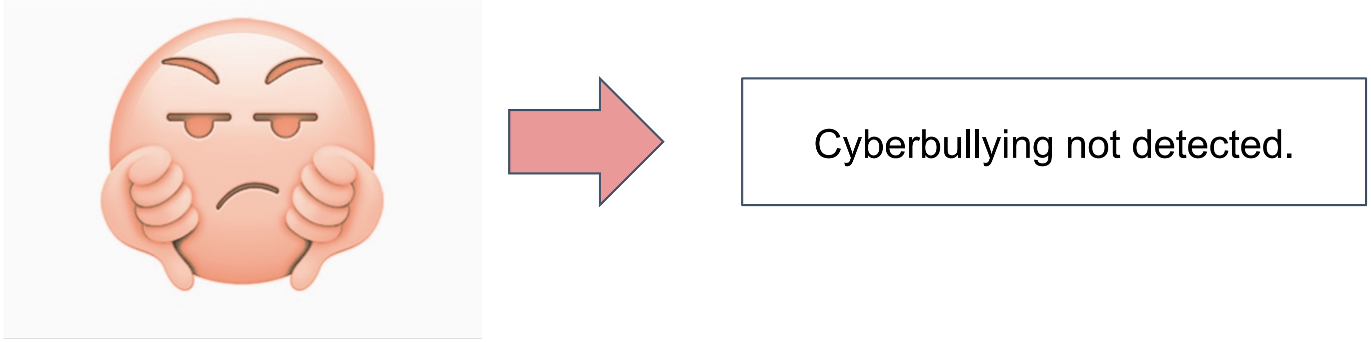
Go check Link: https://cuadvancelab.github.io/labs.html

- Cyberbullying Detection on Images



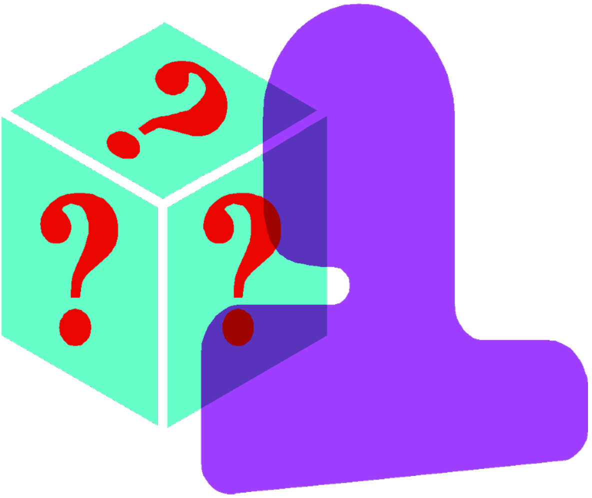
Cyberbullying detected!

- Adversarial attack on Cyberbully detection models



Cyberbullying not detected.

Q & A



Jump to the Lab

Let’s get our hands dirty!

Link for our lab:

https://colab.research.google.com/github/cuadvancelab/cuadvancelab.github.io/blob/main/instructions/lab1/social-science/lab1_interactive.ipynb