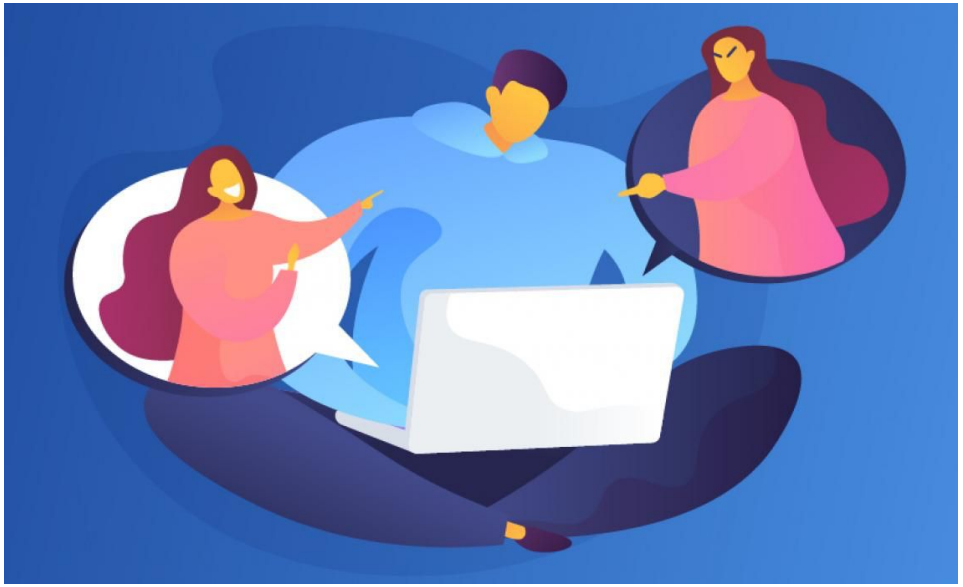
Cyberbullying Detection Using AI

<Instructors>

Outline

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

What is Cyberbullying?



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 repeated behavior, aimed at scaring, angering or shaming those 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 …**

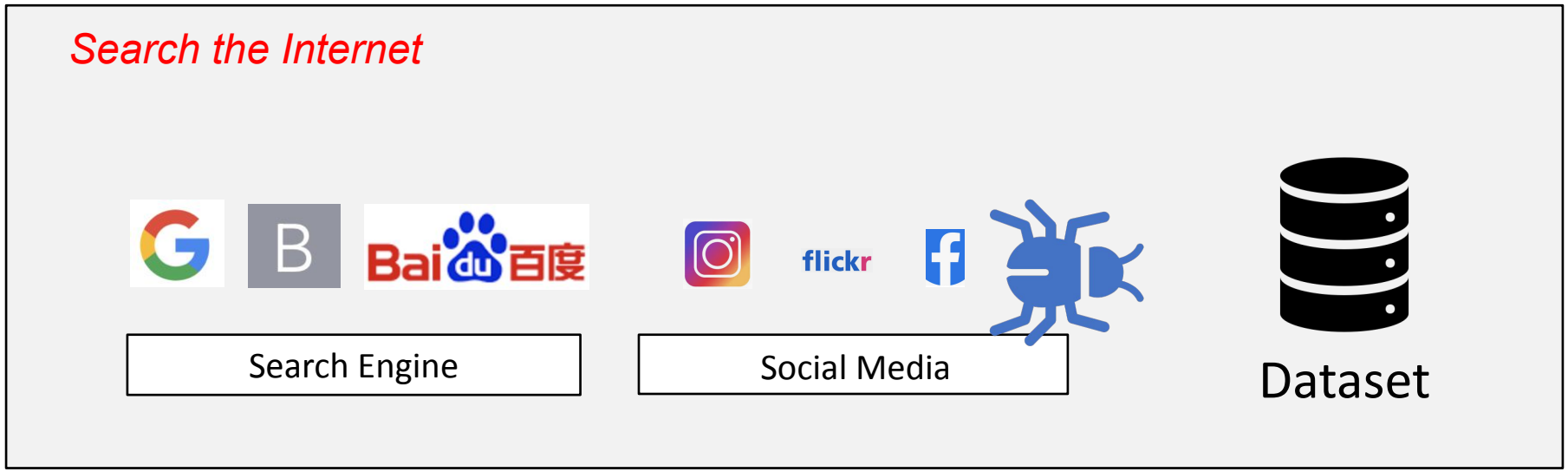
State-of-the-art detectors Google clarifai Amazon Rekognition

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

Process of AI Development

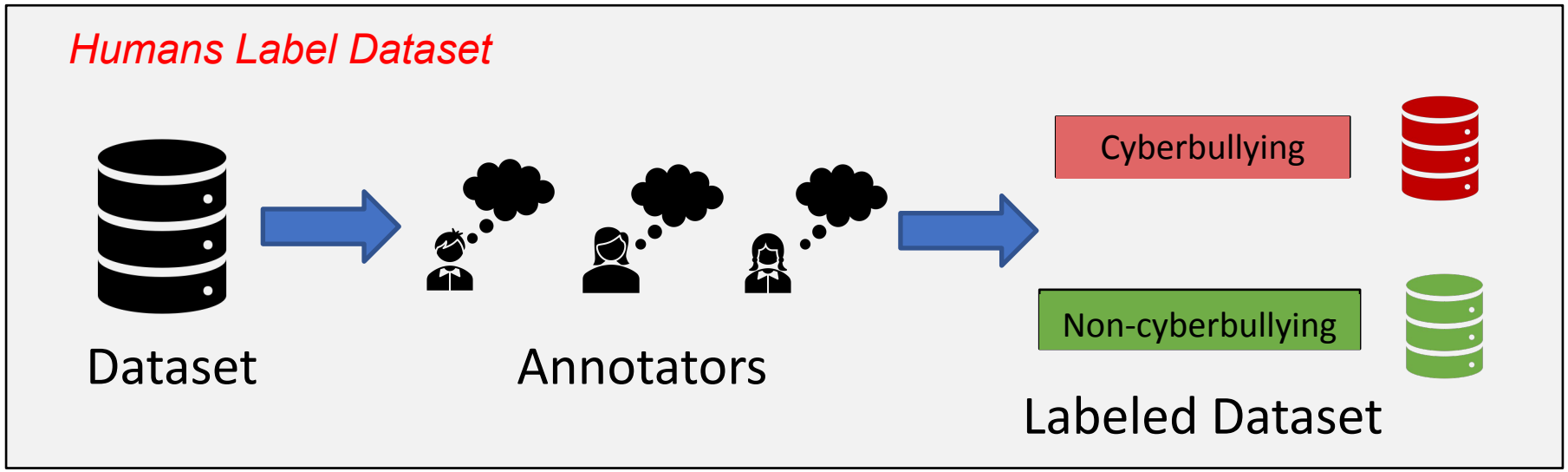
- Data Collection
  - We need dataset for training AI
- Annotation
  - We need to label dataset
- Training
  - AI training process
- Evaluation
  - How good are we doing?
- Deployment
  - Detect on real-world samples

Dataset Collection



Cyberbullying dataset is from Formspring

Annotation


Humans Label Dataset
Cyberbullying
Non-cyberbullying
Dataset
Annotators
Labeled Dataset

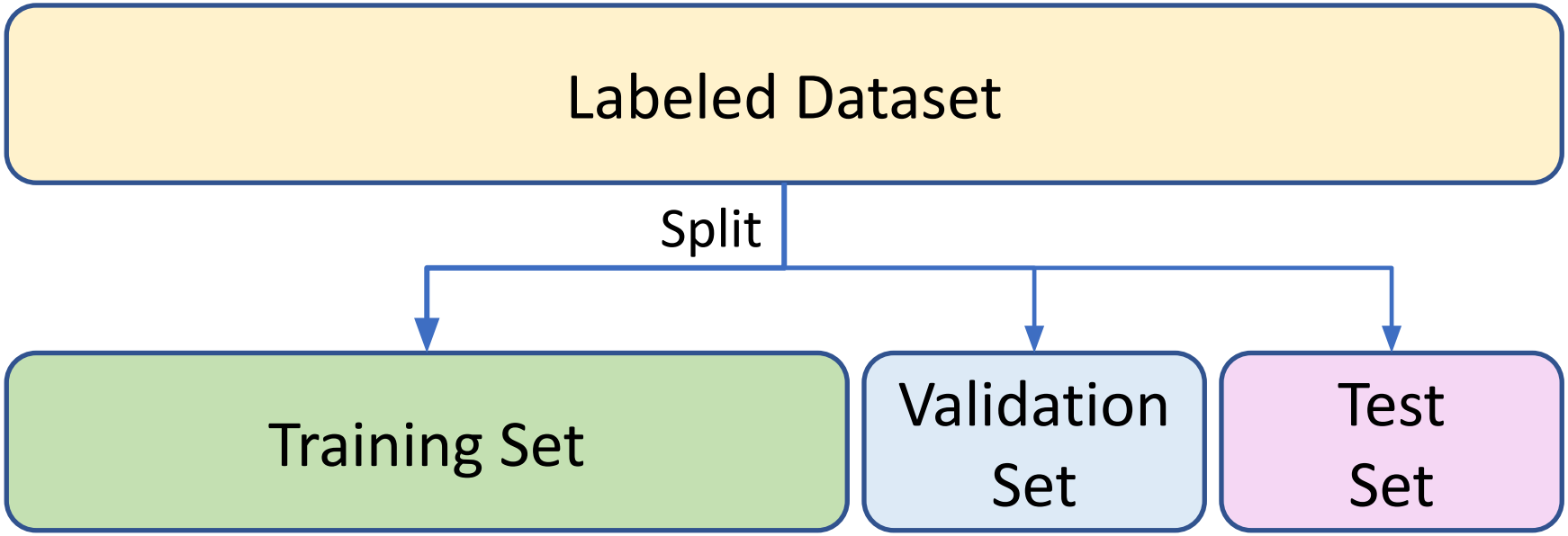
Training, Evaluation, Deployment

Let’s get our hands dirty!

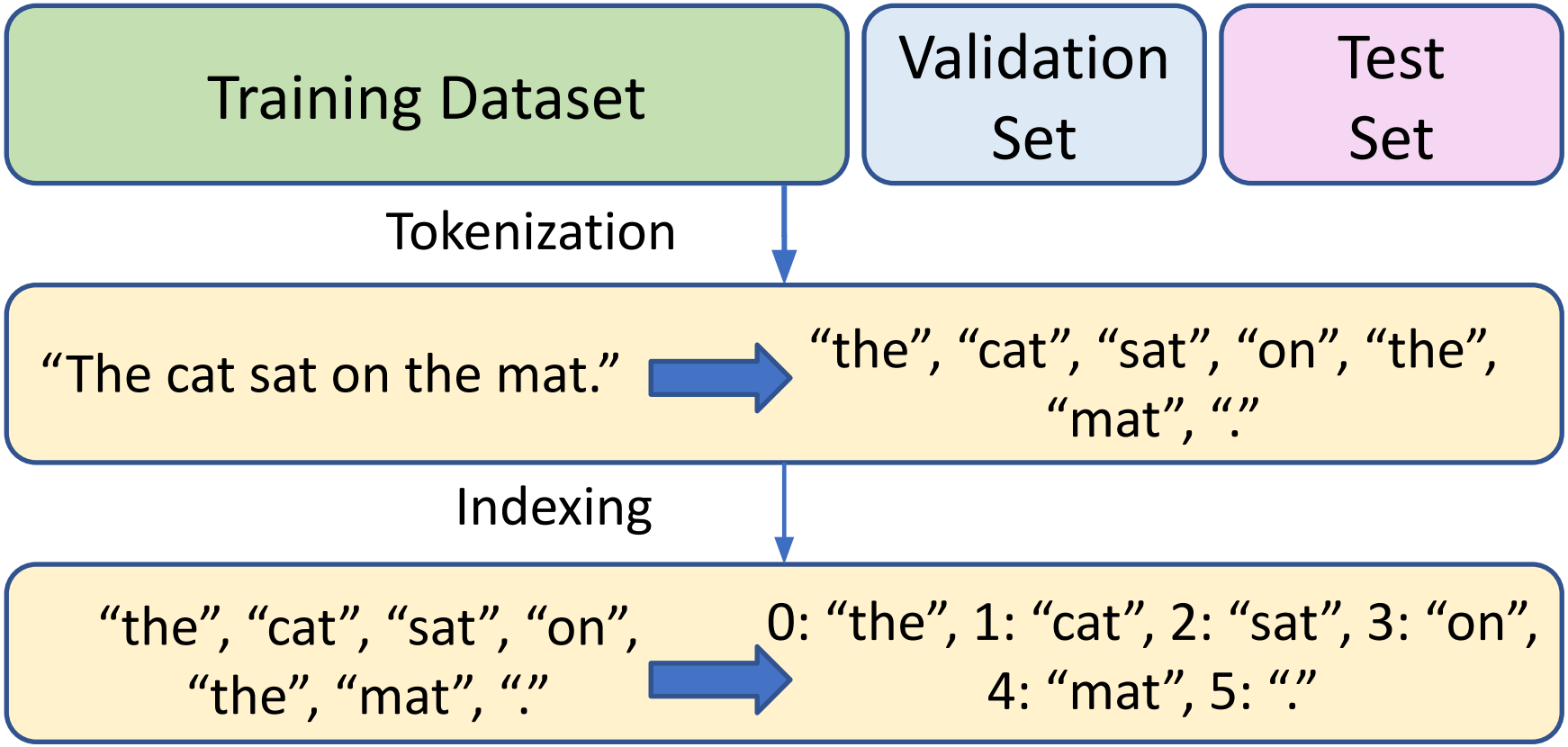
Open the manual and jump to our hands-on lab:

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

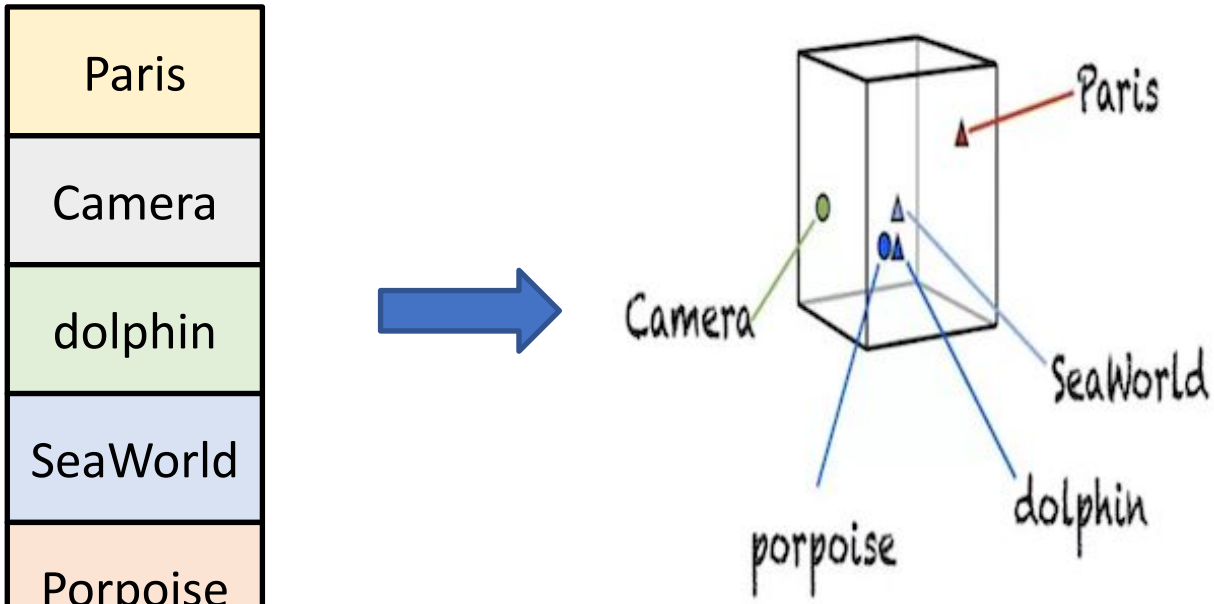
Training


Labeled Dataset
Split
Training Set
Validation Set
Test Set

Training Cont…



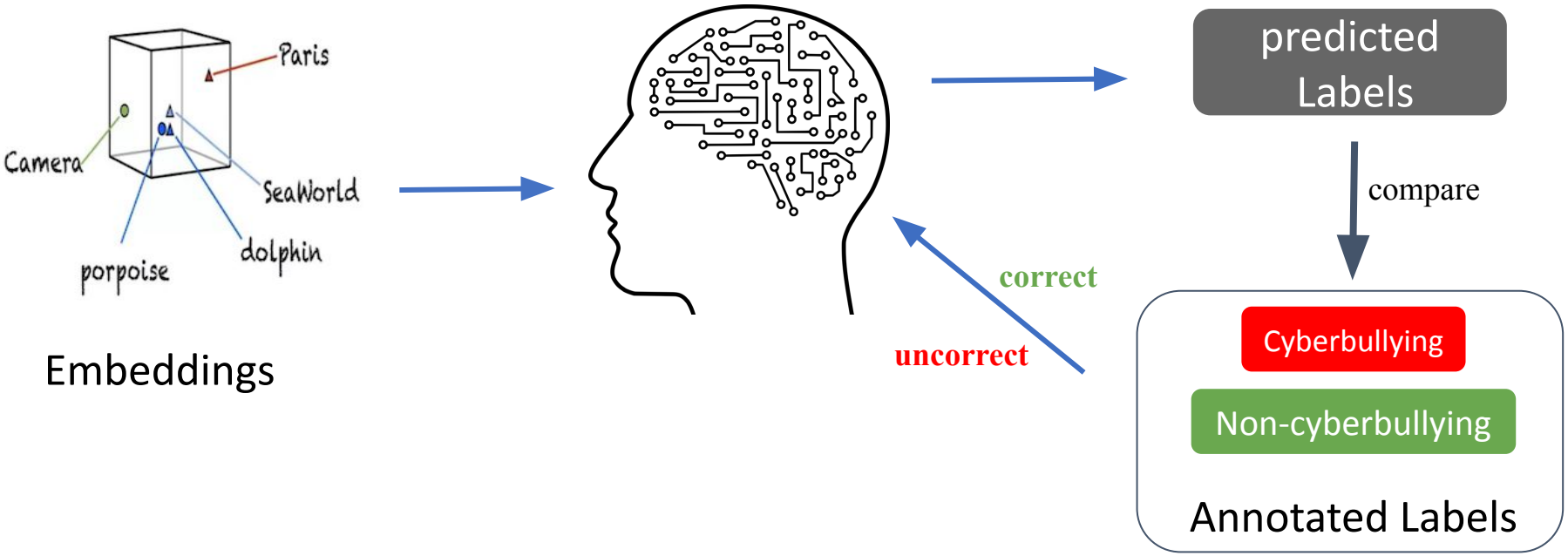
Training Cont…



Tokens

Embeddings

Training Cont…



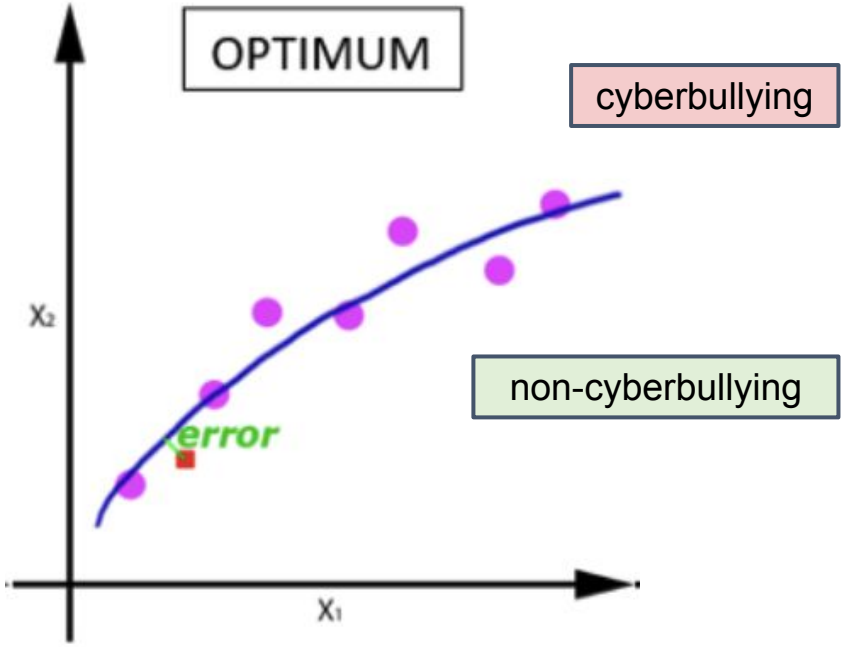
AI is Learning …

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?

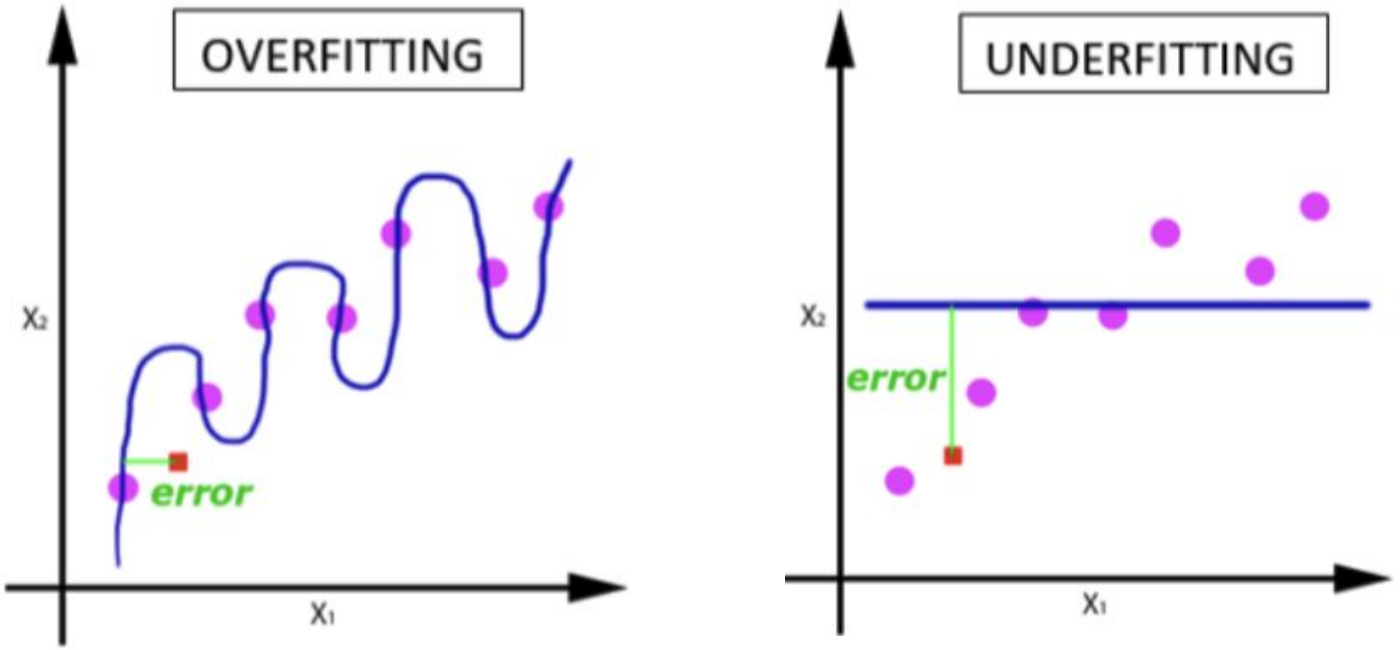
Training Cont…

- How many epochs should we use?
  - optimum epochs

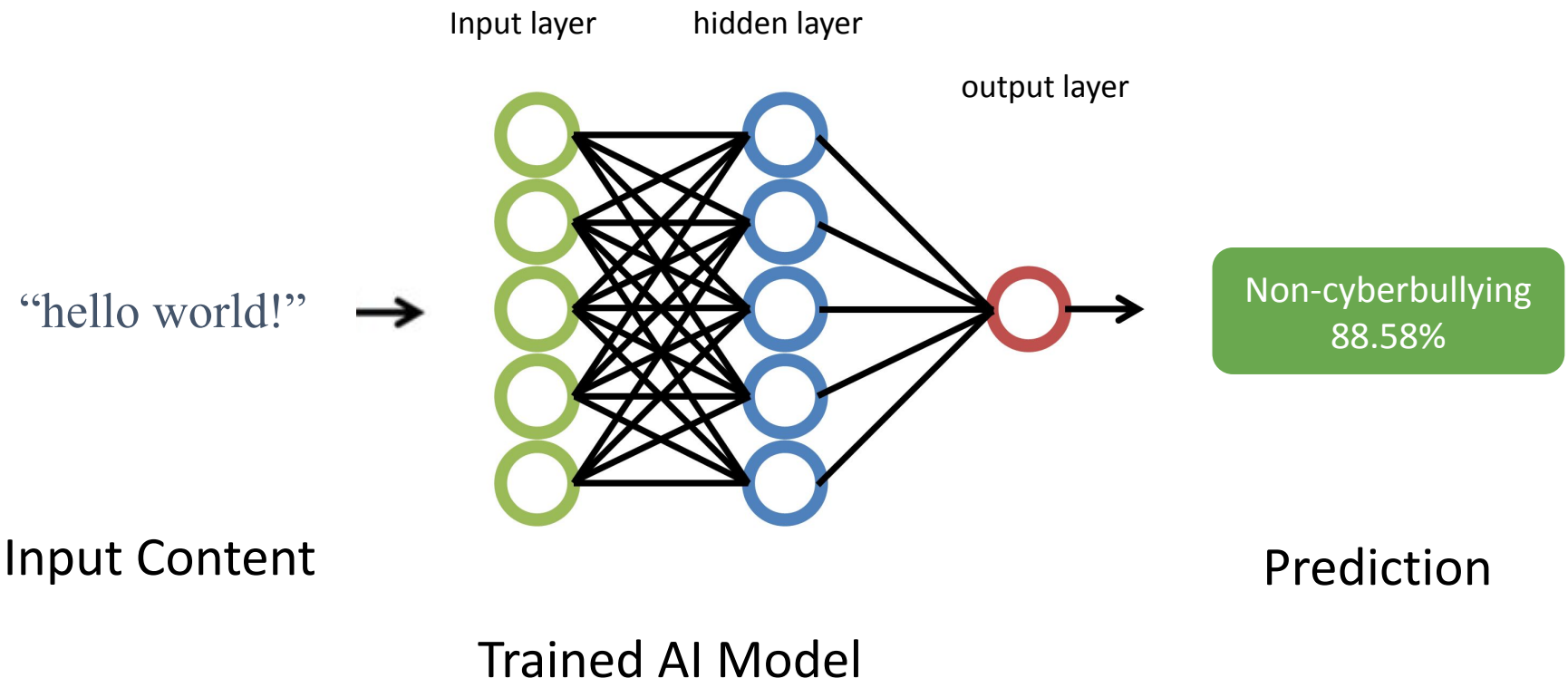


Training Cont…

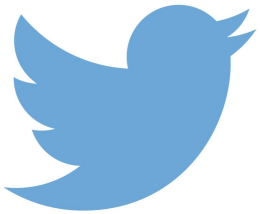
- How many epochs should we use?
  - too many or too few epochs

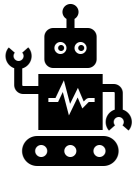


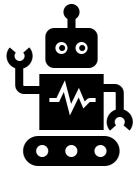
Evaluation

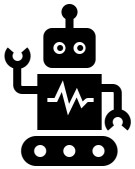

Input layer
hidden layer
output layer
“hello world!”
Non-cyberbullying
88.58%
Input Content
Prediction
Trained AI Model

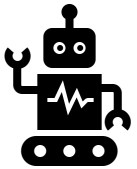
Deployment







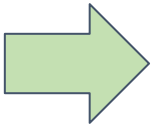




Even More Labs!

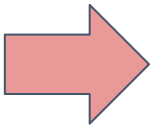
Go check Link: https://cuadvancelab.github.io/labs.html

- Cyberbullying Detection on Images



Cyberbullying detected!

- Adversarial attack on Cyberbully detection models

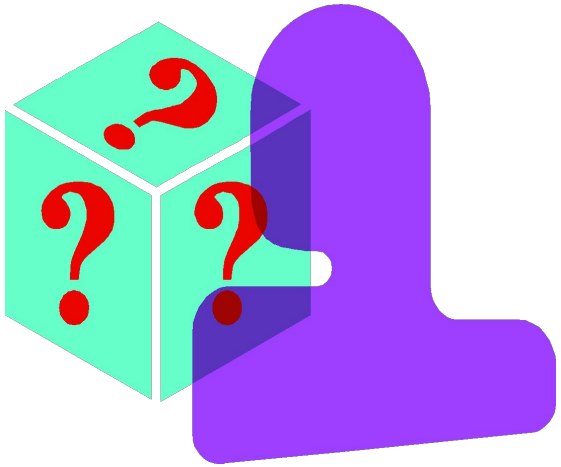


Cyberbullying not detected.

Questions

- Answer the following question in the chat
  - A labeled dataset is split into three sets. Name them
  - Which of these is done first, dataset collection or dataset annotation?

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/computer-science/lab1_interactive.ipynb