
GenAI-ML HW5

Regression

TA: 徐啓惇, 李梁玉軒, 袁紹翔

ntu-gen-ai-ml-2025-fall-ta@googlegroups.com

Deadline: 2025/**11/14** 23:59:59 (UTC+8)

Outline

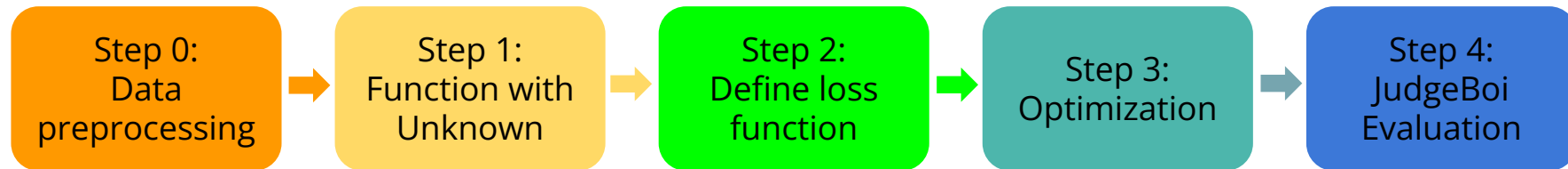
- Homework Overview
- Dataset
- Training
- Evaluation
- Baselines
 - Simple
 - Medium
 - Strong
 - Boss
- Submission & Deadline
- Grading Release Date
- Regulations

Links

- [GENAI-ML 2025 FALL](#)
- [NTU COOL](#)
- [JudgeBoi](#)
- [ML2025 Colab and Kaggle Tutorial](#)
- [HW5 Colab](#)
- [HW5 Kaggle](#)

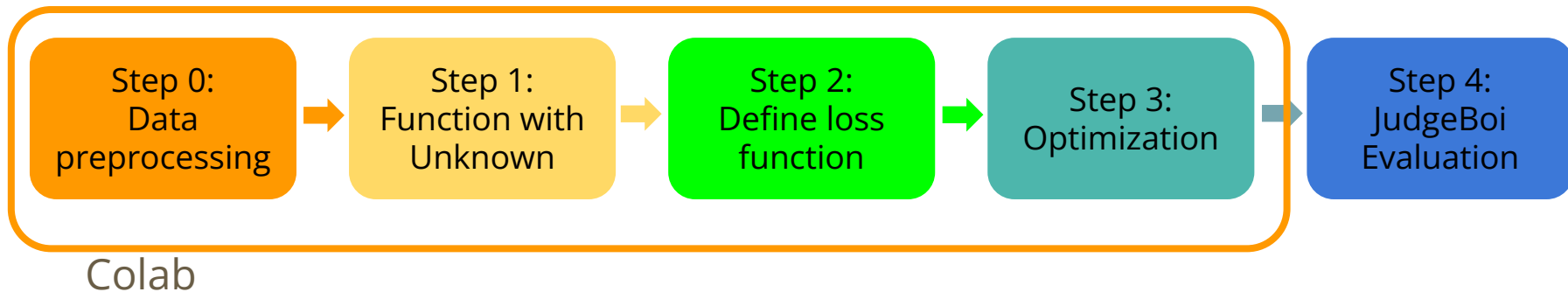
Homework Overview

- In this assignment, we will learn how to use a neural network (NN) to solve a regression task.
- We can decompose the whole assignment into 5 steps :



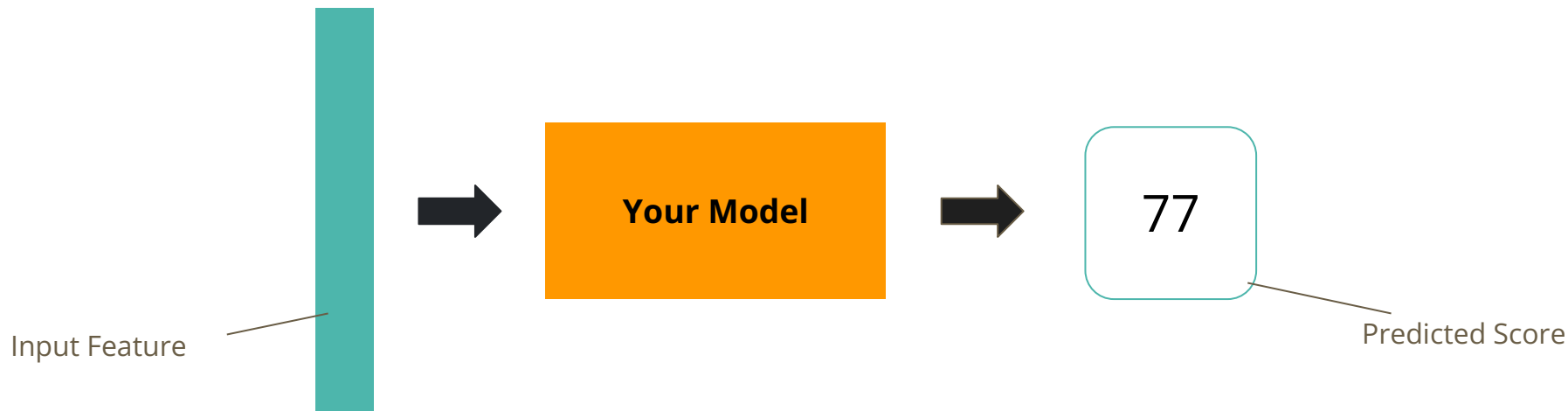
Homework Overview

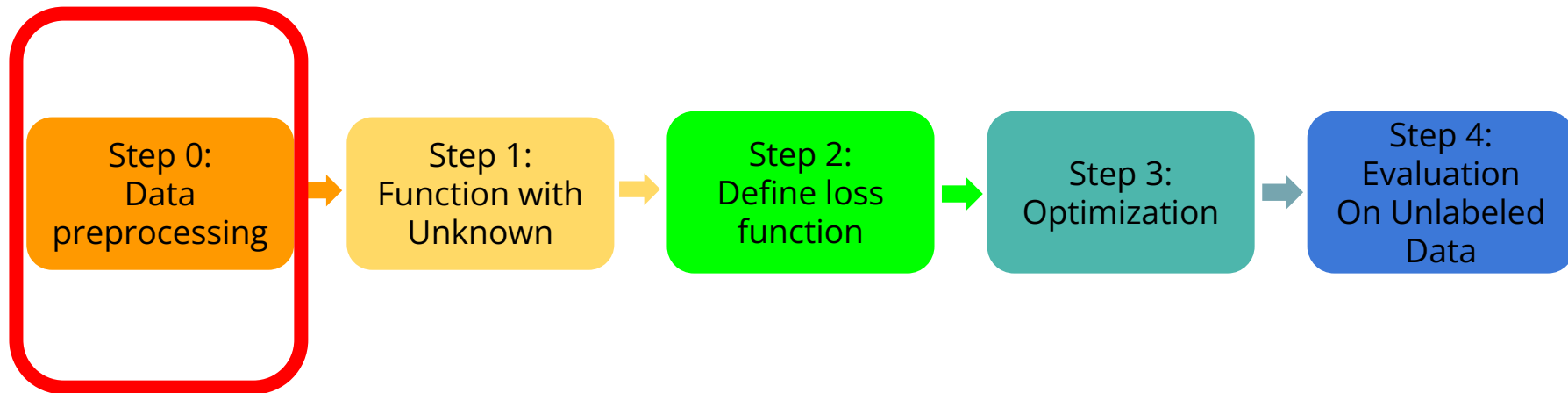
- We expect you to know a bit of python & PyTorch in this assignment.
- You can check the PyTorch tutorial on the [course website](#).
- If you're not familiar with programming, don't worry — we've already implemented the most complex parts for you.
- You can get at least **4 points** just by running the code, and **6 points** by simply adjusting some numbers.



What is Regression ?

- Regression is tasks that predicts a continuous numerical value based on input features.
- In this assignment : Game feature → Score





Dataset Description

- Dataset: Steam Games Dataset 2025
 - Created with Steam store page scraping / Steam Games Scraper
 - Up to date as of March 2025
- In this assignment, you are going to train a model to predict the **“metacritic score”** of games on Steam.
- Metacritic score reflects overall preference from gamers



Dataset (1/5)

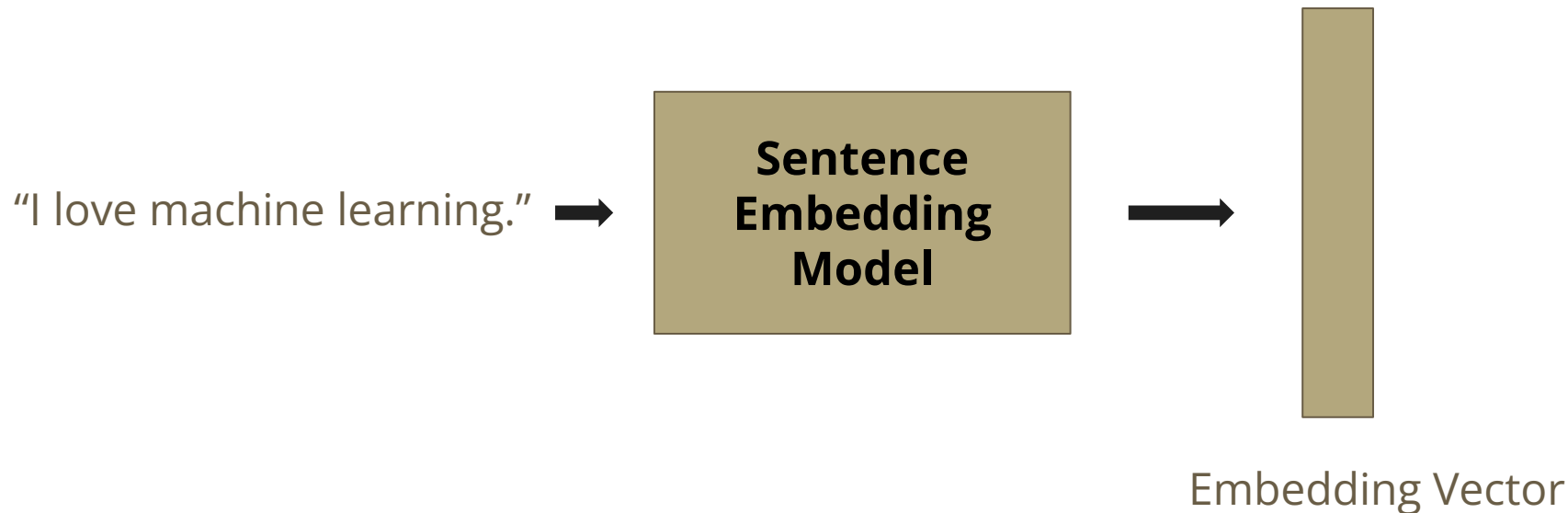
- 45 feature + 1 label
- Label : metacritic score
 - Integer within 0 - 100
- Numeric Feature (41)
 - Basic Information : appid, release_date, required_age, price ...
 - Stats : average_playtime, median_playtime_2weeks ...
 - Binary Features : linux , has_website ...
 - To see more, inspect the [csv](#) ...

Dataset(2/5)

- Text Feature (4)
 - name : Game name
 - detailed_description : Detailed description of the game
 - short_description : Brief description of the game.
 - Reviews : User reviews or feedback
- But, basically, the model used in this assignment need to take vector as input
- **Utilize pretrained sentence embedder !**

Dataset(3/5)

- Sentence Embedding



Dataset(4/5)

- train.csv (3200 samples)

Features

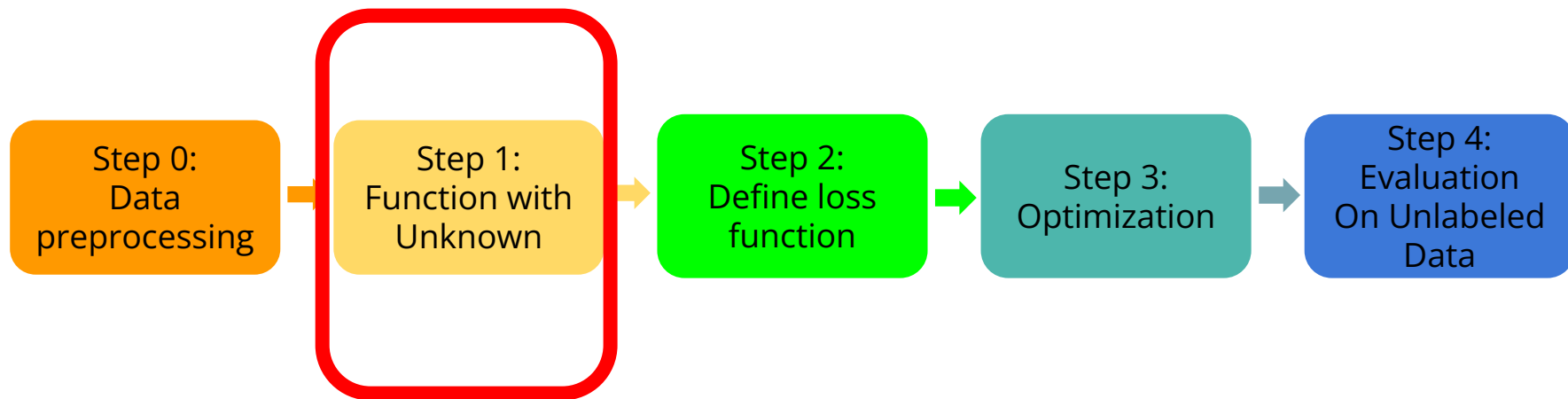
Metacritic Score

[illegible]

- test.csv (800 samples)

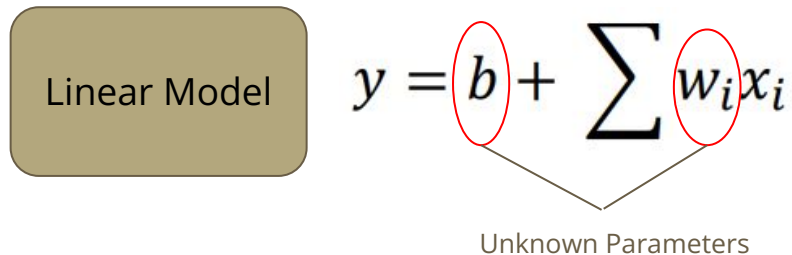
Only features, no label (metacritic score)

12



Model - “Function with Unknown”

- Machine Learning $\approx$ Looking for Function
- A neural network is essentially a function with unknown parameters
- By using training data, we aim to find a set of parameters that make this function perform well

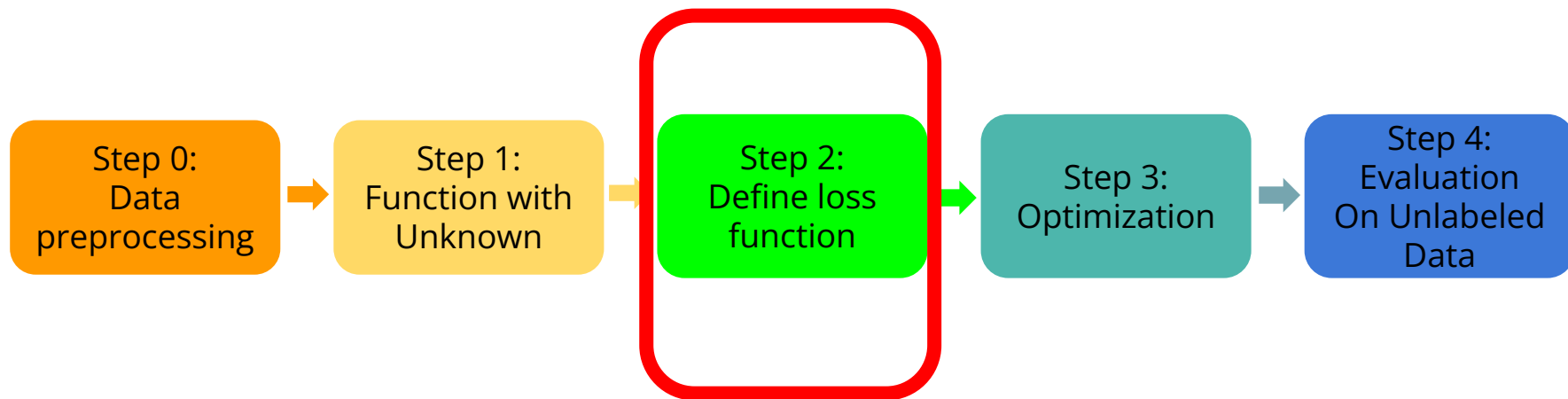


The diagram shows a rounded rectangular box on the left containing the text "Linear Model". To its right is the equation $y = b + \sum w_i x_i$. In this equation, the terms b and w_i are each enclosed in a red oval. A bracket is drawn below these two ovals, pointing down to the text "Unknown Parameters".

Linear Model

$$y = b + \sum w_i x_i$$

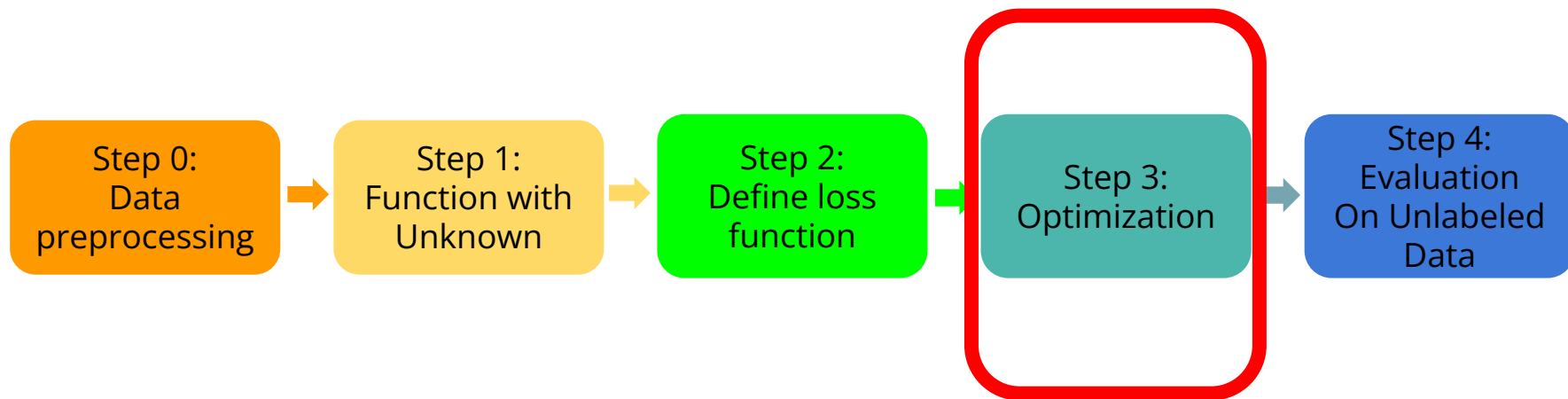
Unknown Parameters



Goodness of Function - Loss function

- **Loss function** quantifies how well or poorly the model fits the data
- It measures the distance between the model's prediction and the true value, guiding the learning process during training
- Regression : Mean Square Error

$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$$

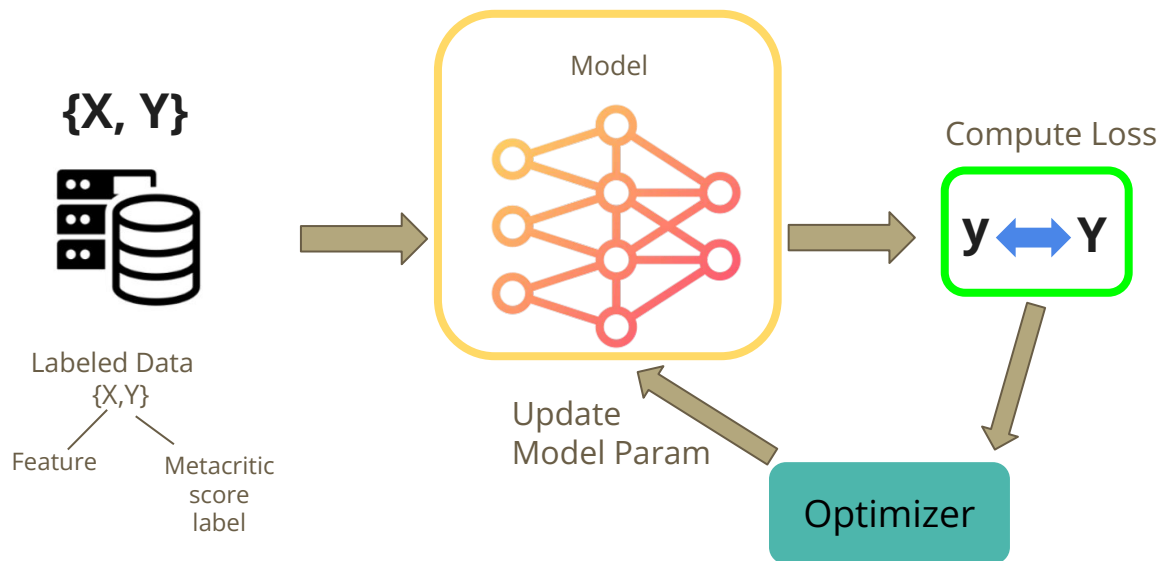


Optimization - Find the “Good” Model

- What happens in a training loop
 - **Forward Pass** : Feed the input into model and retrieve output
 - **Compute Loss** : calculate the loss according to ground truth answer
 - **Backpropagation** : Compute the gradient of the loss with respect to each model parameter
 - **Optimize** : Adjust model parameters in the direction that reduces the loss
 - Repeat until convergence ...
- Optimizer
 - PyTorch provides built-in optimization algorithms for parameter updates
 - Common choices
 - Stochastic Gradient Descent (SGD)
 - Adam

Training Loop

- Train



Hyperparameters

- Values that control the training process, not learned from data

Number of Epoch	How many times the model sees the full dataset
Learning Rate	Step size for updating parameters
Batch Size	Number of samples per gradient update
Optimizer	Algorithm that updates weights based on gradients (e.g. SGD, Adam)
Weight Decay	Prevents overfitting by penalizing model complexity

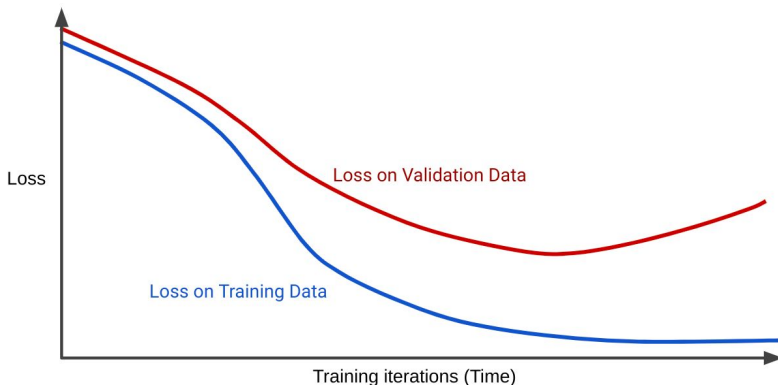
Issue You May Meet When Doing Optimization...

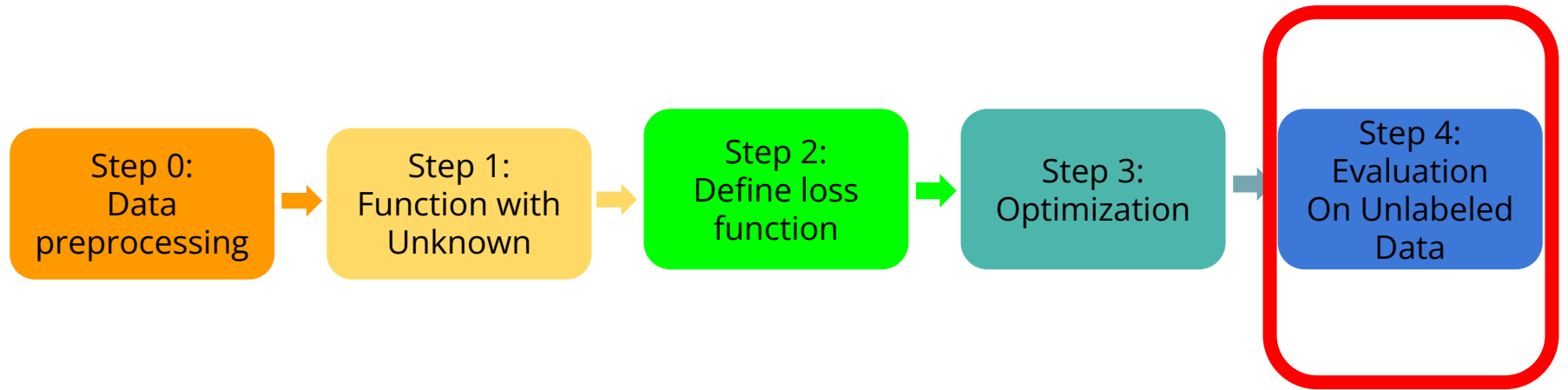
- Overfitting
 - Model performs well on training data but poorly on unseen data
 - Prevent overfitting : **use validation set** to monitor performance
 - Ways to solve : reduce model complexity , more training data **(Not available in HW5)**
- Validation Set
 - A separate dataset (not used in training) to evaluate model performance during training
 - In practice, we reserve a portion of labeled data as validation set

How to read ?

If both decrease → learning well

If they diverge → possible overfitting





Evaluation Metric (1/2)

- Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^N \|y(i) - \hat{y}(i)\|^2}{N}},$$

Diagram illustrating the RMSE formula with labels:

- ground truth y (points to $y(i)$)
- predicted y (points to $\hat{y}(i)$)
- N (points to the denominator, labeled "number of evaluated data")

Evaluation Metric (2/2)

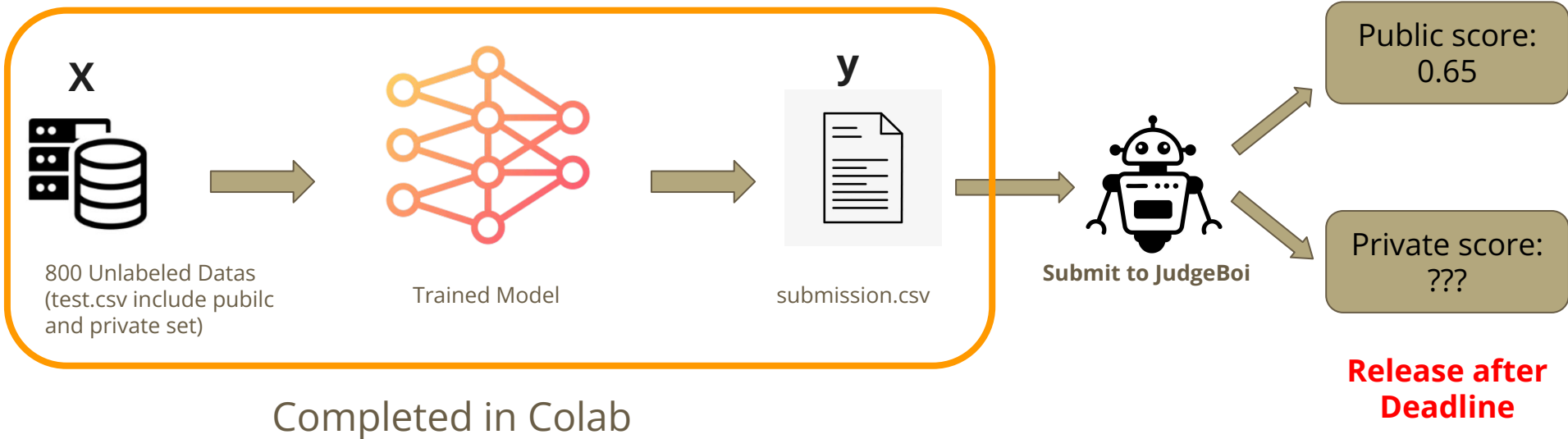
- An example ...

Predicted		Ground Truth	
ID	Score	ID	Score
123	92	123	87
246	84	246	80
989	47	989	55

$$\Rightarrow \text{RMSE} = \sqrt{1/3 * (25+16+64)} \doteq 5.91$$

Public/ Private set

- Evaluate



Grading

There are **10 points** in total for HW5:

- Successful submission to JudgeBoi +2 pt
- Each public baseline worth +1 pt
 - Simple : $RMSE < 8.00$
 - Medium : $RMSE < 6.30$
 - Strong : $RMSE < 5.30$
 - Boss : $RMSE < 3.70$
- Each private baseline worth +1pt
 - There are also 4 baselines for private baseline
 - **We would reveal the result after the deadline**

Submission(1/2)

- Submit your prediction file to JudgeBoi
- You have 5 submission quota per day, reset at 23:59
- Deadline: 2025/11/14 23:59:59 (UTC+8)
- No late submission is allowed, please finish your homework as soon as possible

Submission(2/2)

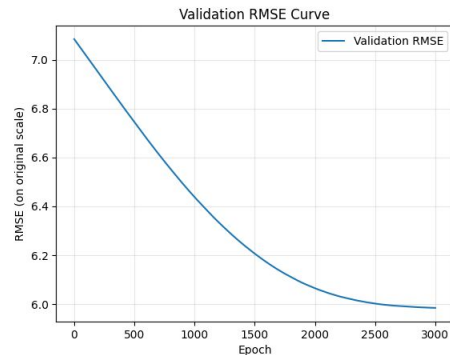
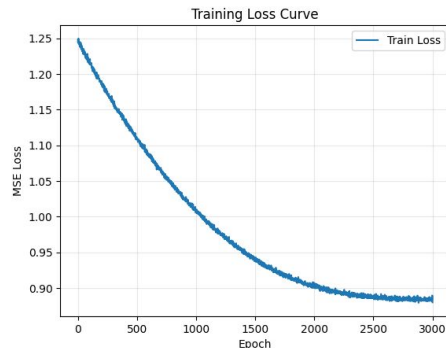
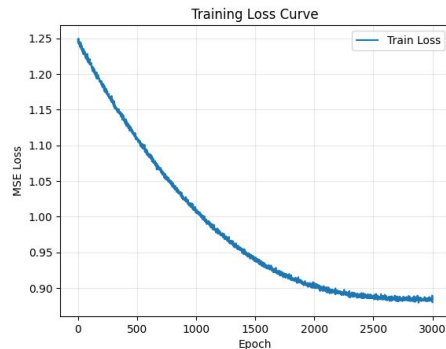
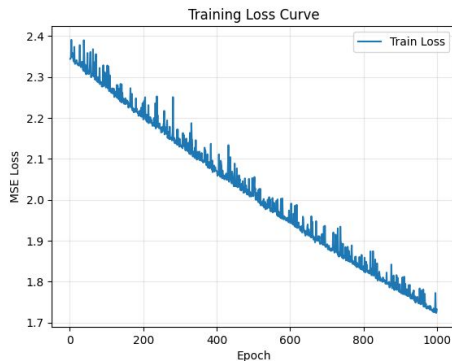
- Reminder : You can select up to 2 submissions on JudgeBot (If not selected, top 2 public score will be chosen by default)
- Beware of overfitting, high public score doesn't assure high private score
 - Refer to your validation score during training

Hints For Passing the Baseline

JudgeBoi Submission (+2 pt)	Submit predicted results to JudgeBoi	Estimated Training Time	Estimated Inference Time
Public Simple Baseline (+1 pt)	Just run the colab and submit the predicted "submission.csv" to JudgeBoi	2 minutes	1 minutes
Private Simple Baseline(+1 pt)			
Public Medium Baseline (+1 pt)	Hyperparameter tuning	5 minutes	1 minutes
Private Medium Baseline(+1 pt)			
Public Strong Baseline (+1 pt)	Change model to MLP	5 minutes	1 minutes
Private Strong Baseline(+1 pt)			
Public Boss Baseline (+1 pt)	Feature selection	5-10 minutes	1 minutes
Private Boss Baseline(+1 pt)			

Medium Baseline : Observe Training Curves

- Converged or not ?
 - learning rate
 - Add # of epochs
- Overfit ?
 - Decrease # of epochs
 - Add weight_decay



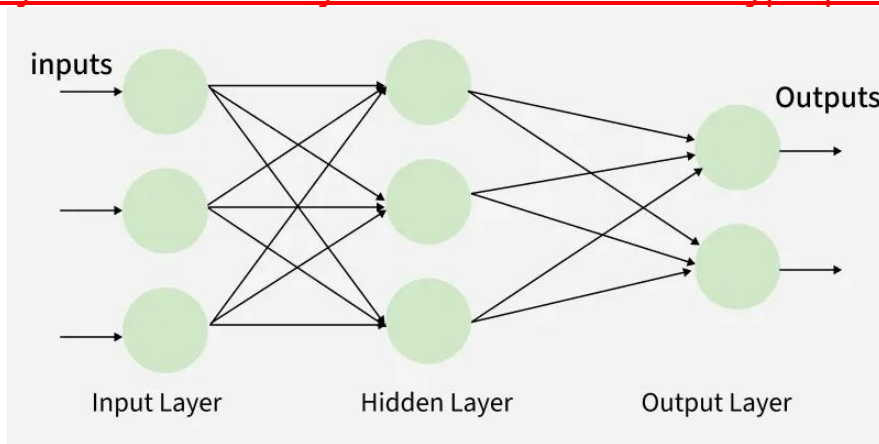
Medium Baseline : Tune Hyper-parameter

- Experiment on these hyperparameter

Number of Epoch	1000 ~ 5000
Learning Rate	1e-2 ~ 1e-4
Batch Size	[8, 16, 32, 64, 128]
Optimizer	Adam, SGD
Weight Decay	1e-3 ~ 1e-5

Strong Baseline: Modify the Model

- Multilayer Perceptron (MLP)
 - Input Layer : takes input data
 - Hidden Layer : linear layer + activation function
 - Output Layer : gives the final prediction (in this assignment, a scalar value)
 - # of Hidden Layer and hidden layer dimension are also hyperparameter



Boss Baseline : Feature Selection

- Choosing the most relevant input variables that contribute significantly to a model's prediction
- It improves model performance and reduces overfitting by removing noisy or redundant features

pct_pos_total	0.708	Metacritic Score	
positive	0.273	Metacritic Score	
is_top_20	0.0369	Metacritic Score	

Feature Name Correlation coefficient

Grading Release Date

- The grading of the homework will be released before 2025/11/20 23:59:59 (UTC+8)

Useful Links

- Hung-yi Lee, Regression & Gradient Descent (Mandarin)
 - ([link1](#), [link2](#), [link3](#), [link4](#), [link5](#), [link6](#))
- Hung-yi Lee, Tips for Training Deep Networks (Mandarin)
 - ([link1](#), [link2](#))
- <https://pytorch.org/docs/stable/index.html>
- Ask ChatGpt, Google, Stackoverflow may also help ...

Regulations

- You should NOT plagiarize
- You should NOT modify your prediction files manually
- Do NOT share codes or prediction files with any living creatures
- Do NOT use any approaches to submit your results more than 5 times a day
- **Do not search or use additional data or pre-trained models.**
- Please protect your own work and ensure that your answers are not accessible to others. If your work is found to have been copied by others, you will be subject to the same penalties
- Your final grade $\times 0.9$ and get a score 0 for that homework if you violate any of the above rules first time (within a semester)
- You will get F for the final grade if you violate any of the above rules multiple times (within a semester)
- Prof. Lee & TAs preserve the rights to change the rules & grades

Contact us if you have problems ...

- NTU Cool Homework 5 Discussion
 - 如果同學的問題不涉及作業答案或隱私, 請一律使用 NTU Cool 討論區
 - 助教們會優先回答 NTU Cool 討論區上的問題
- Email : ntu-gen-ai-ml-2025-fall-ta@googlegroups.com
 - Title should start with [GenAI-ML 2025 Fall HW5]
 - Email with the wrong title will be moved to trash automatically
- TA-Hour
 - Time:
 - 10/27 11/3 11/10 Monday 20:00~22:00
 - 10/24 10/31 11/7 11/14 Friday 17:30~19:30
 - Location: [Google Meet](#)