利用model output 偷取 Machine Learning-as-a-Service 平台 model

網路多媒體實驗

組員：林承德、趙冠豪、李羚毓

指導助教：曾煒傑
指導老師：林宗男教授

實驗目的

- 1. 實作論文“ Stealing Machine Learning Models via Prediction APIs” neural network 的方法，偷取ML model
- 2. 利用 neural network 的方式實際嘗試偷取各種不同Model

實驗動機

論文：Stealing Machine Learning Models via Prediction APIs

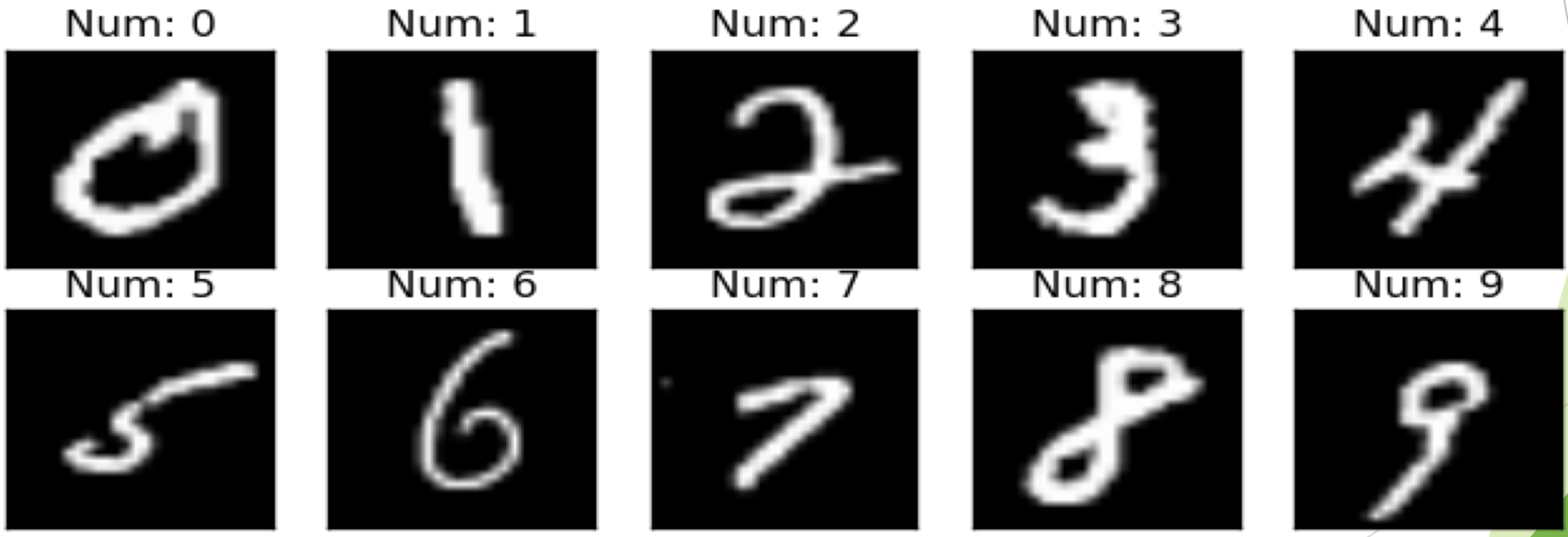
1. Linear Regression, Decision Tree 的Model
   ➔ 同一種方式偷取

2. Neural Network 論文篇幅少，回去看 code
   ➔ 利用 API

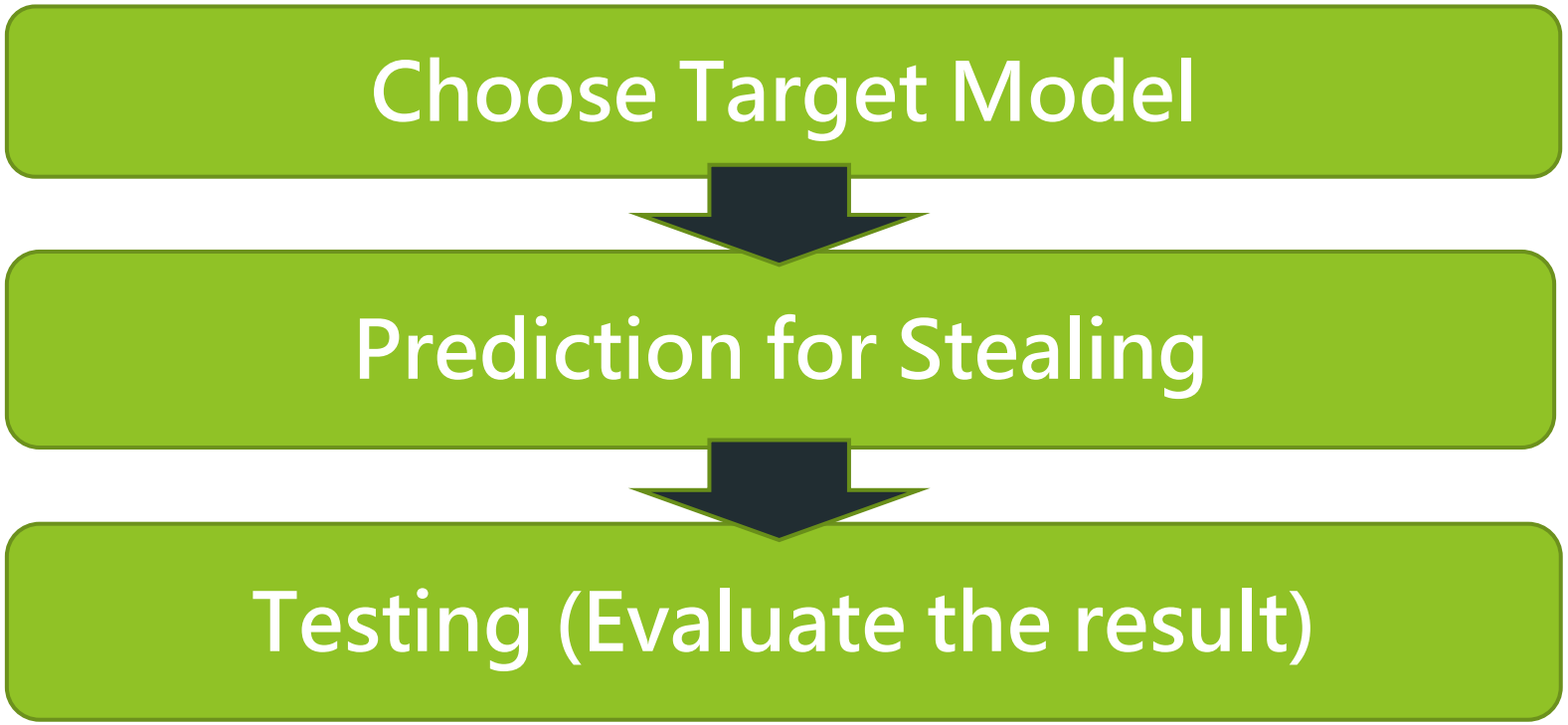
classes	0	1	2	3	4	5	6	7	8	9
label	0	1	0	0	0	0	0	0	0	0
probability	0.005	0.8	0.005	0.005	0.005	0.005	0.005	0.16	0.005	0.005

資料介紹

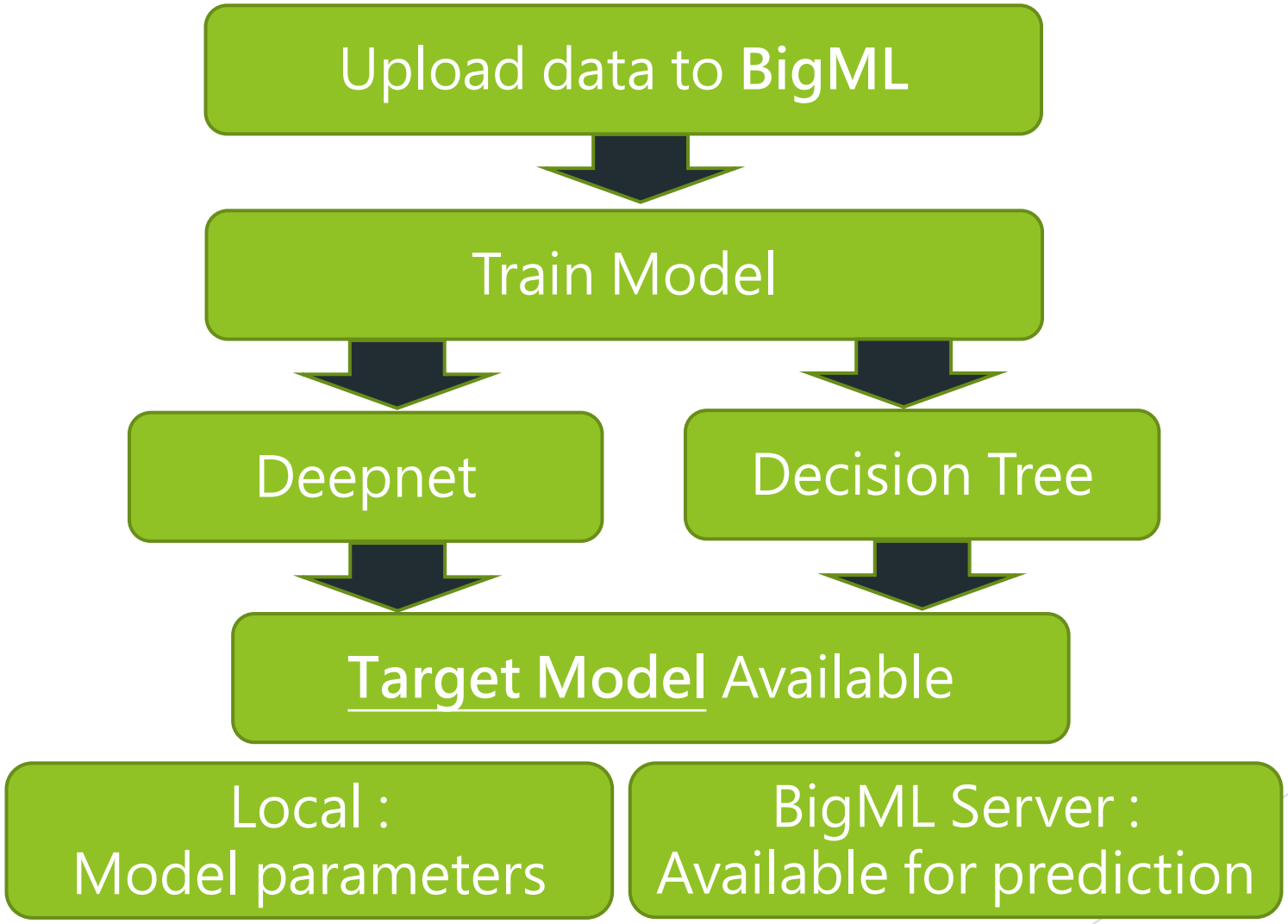
- MNIST 資料集



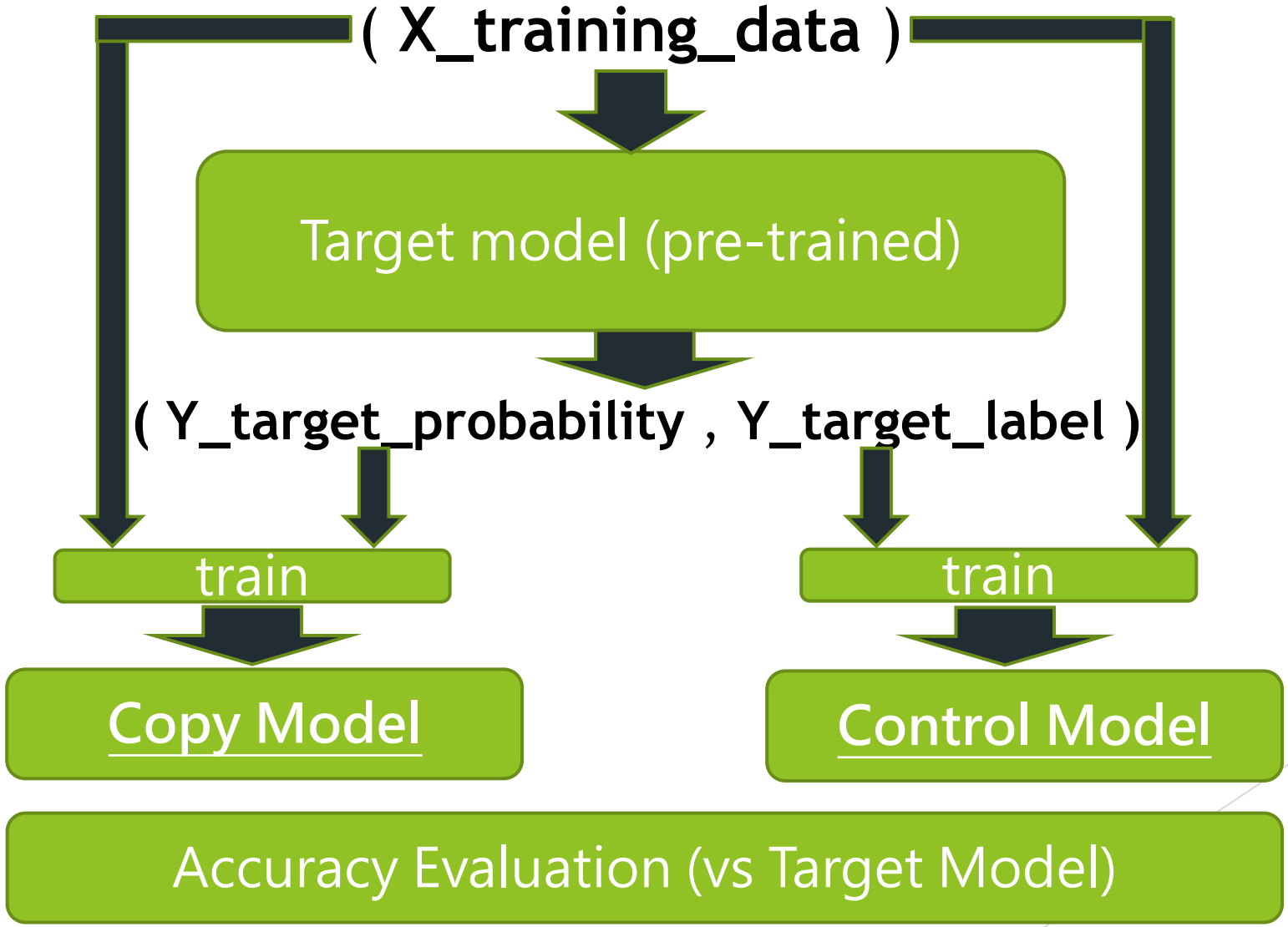
實驗步驟


Choose Target Model
Prediction for Stealing
Testing (Evaluate the result)

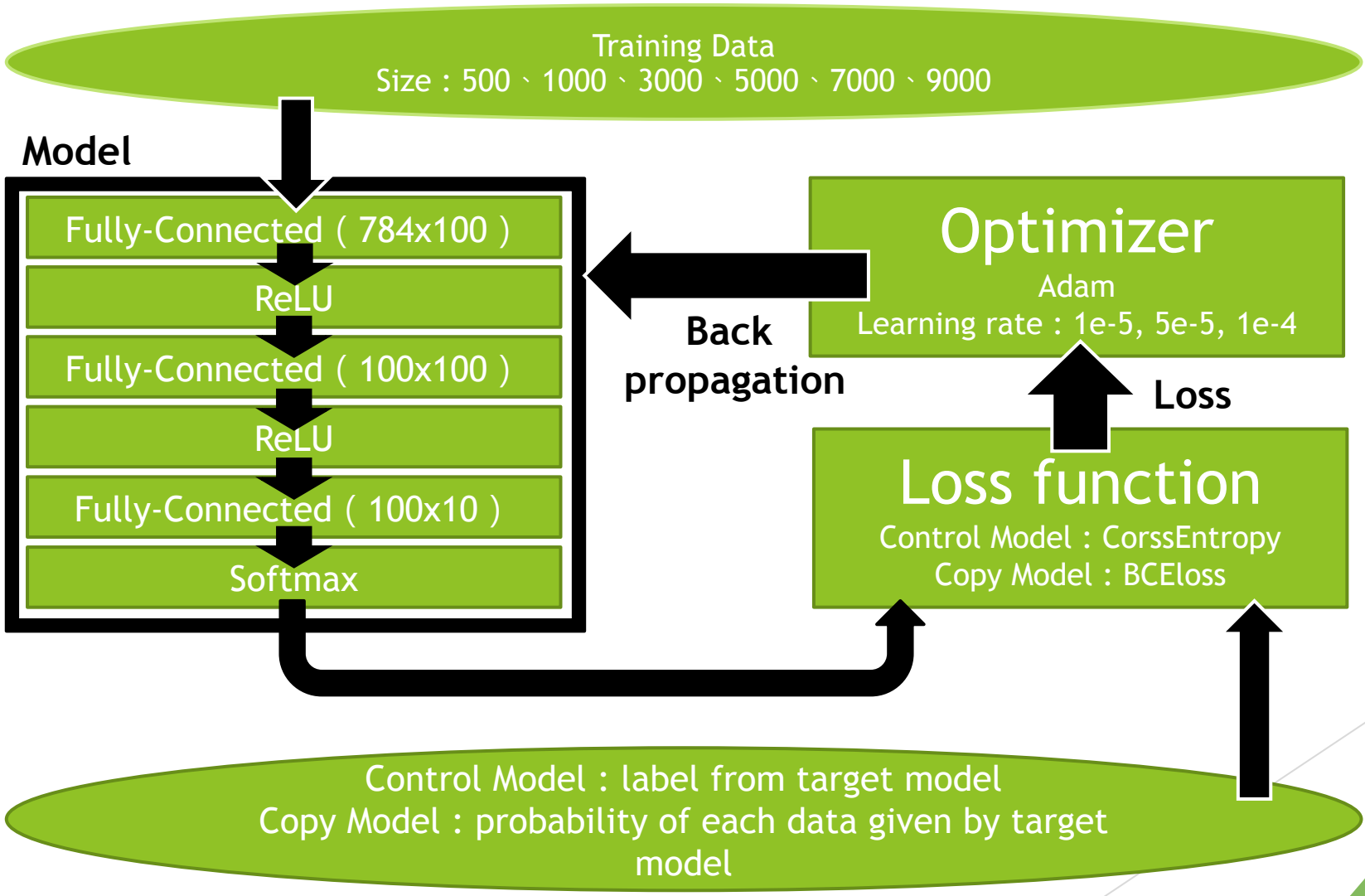
Target Model 建立



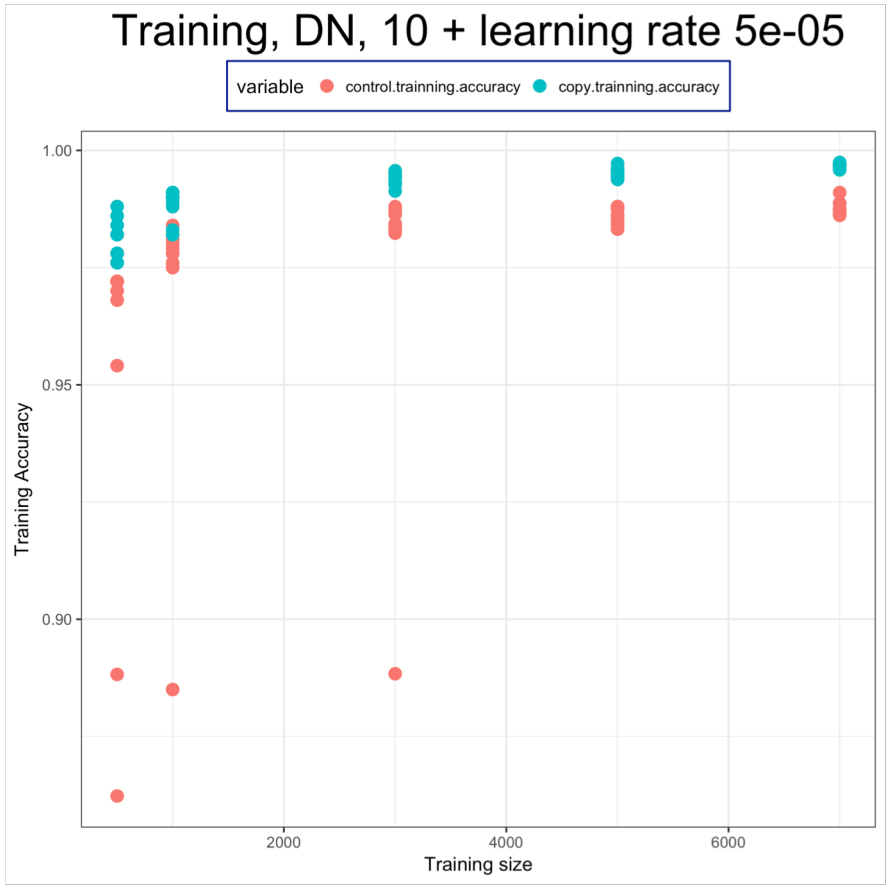
Copy Model VS Control Model



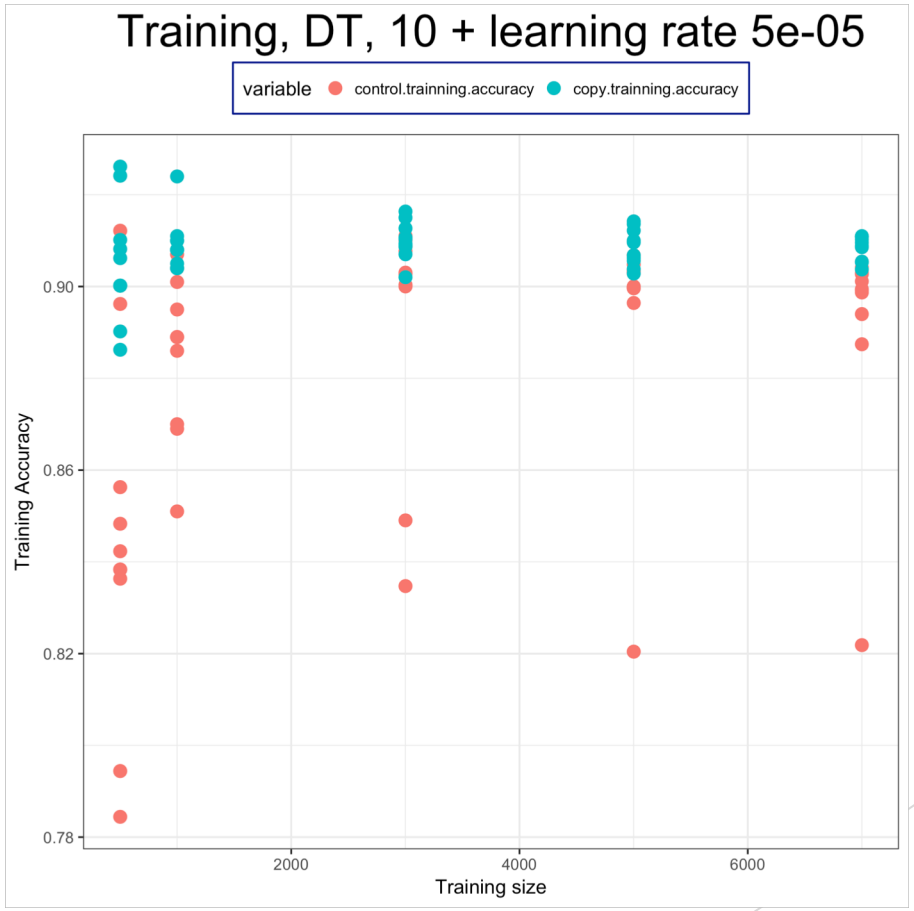
Training of Control / Copy Model



結論



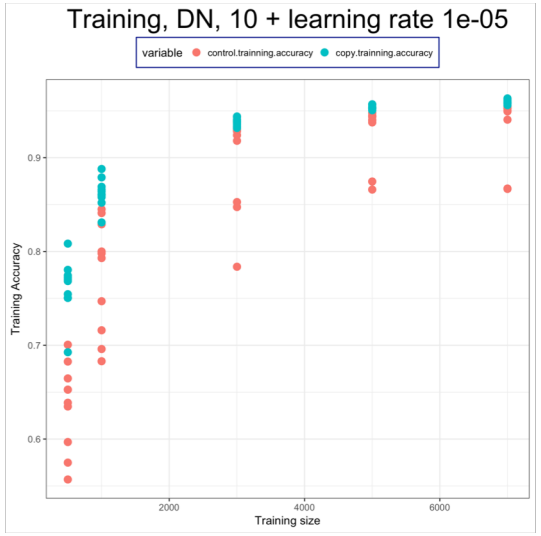
Deepnet

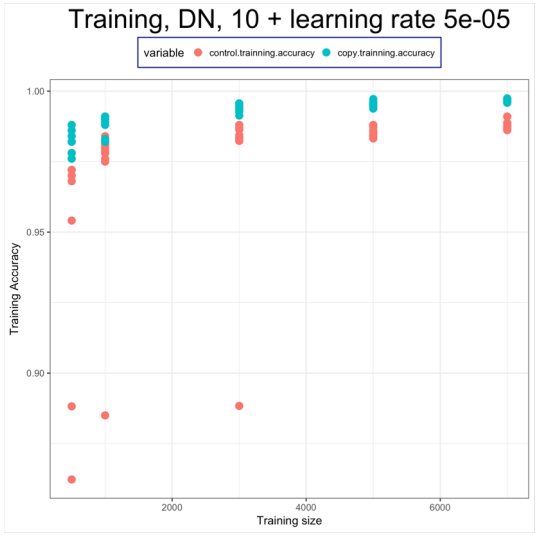


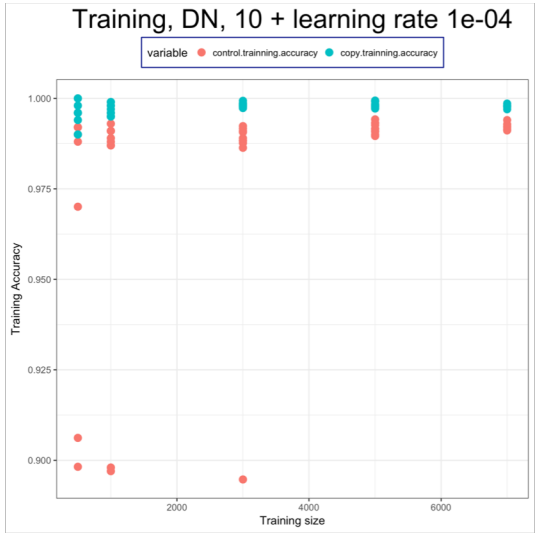
Decision Tree

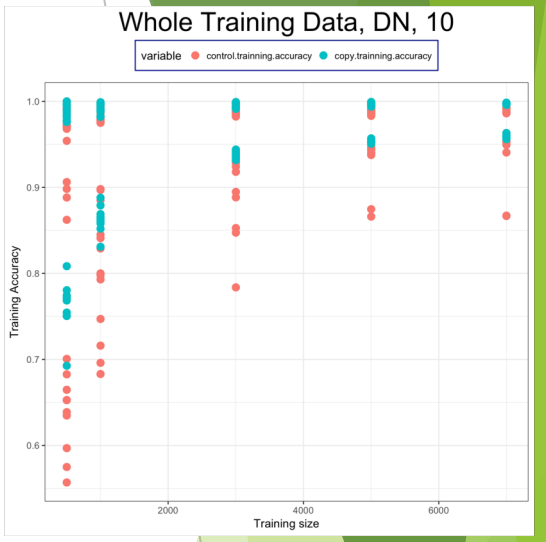
Thanks for your listening

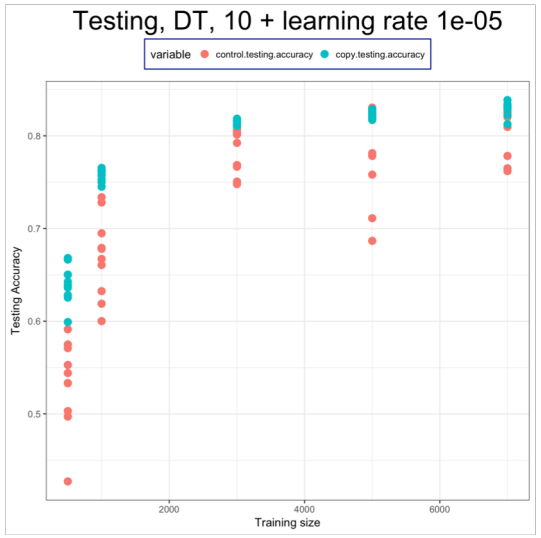
Training accuracy (not average)

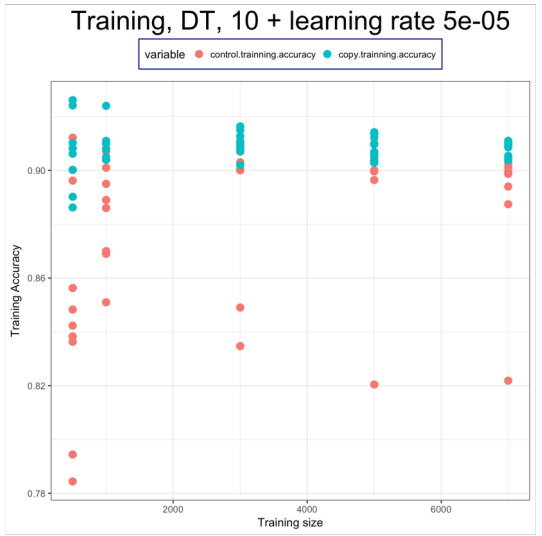


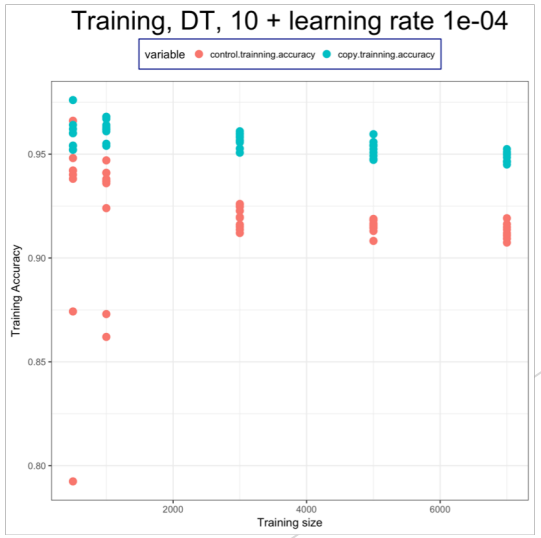




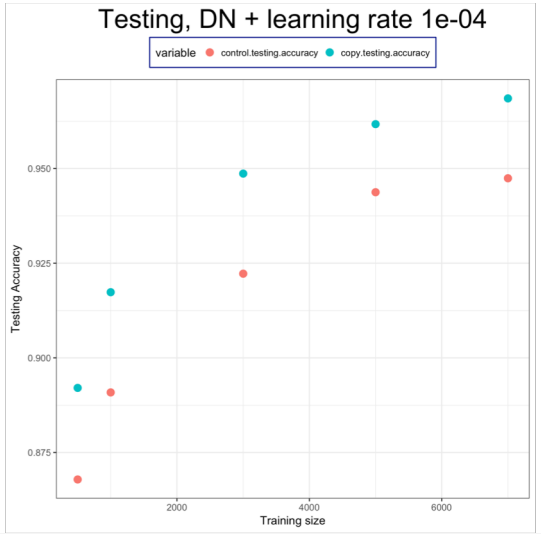


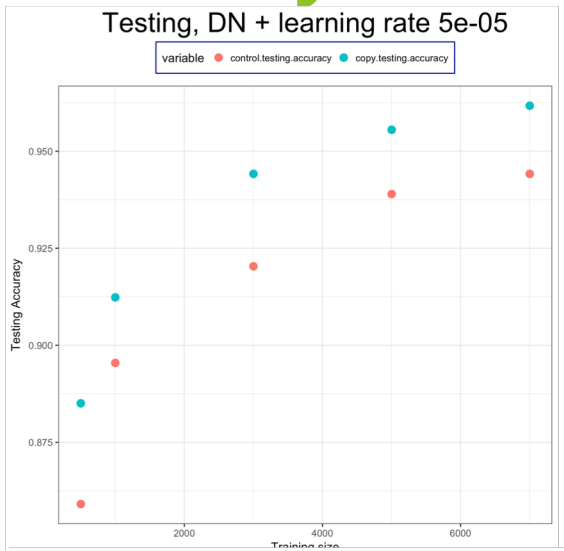


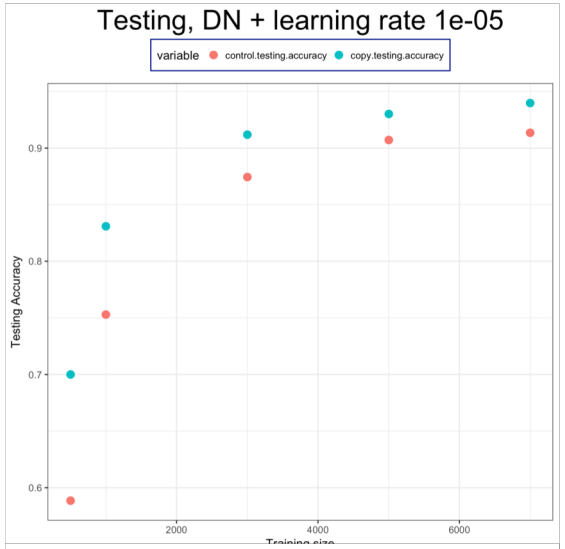


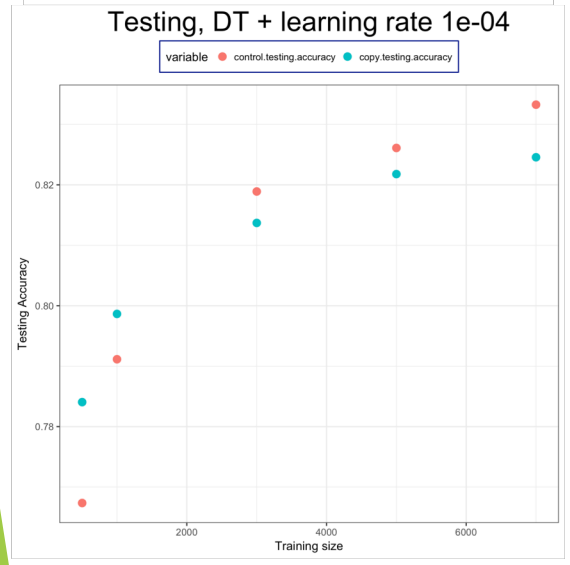


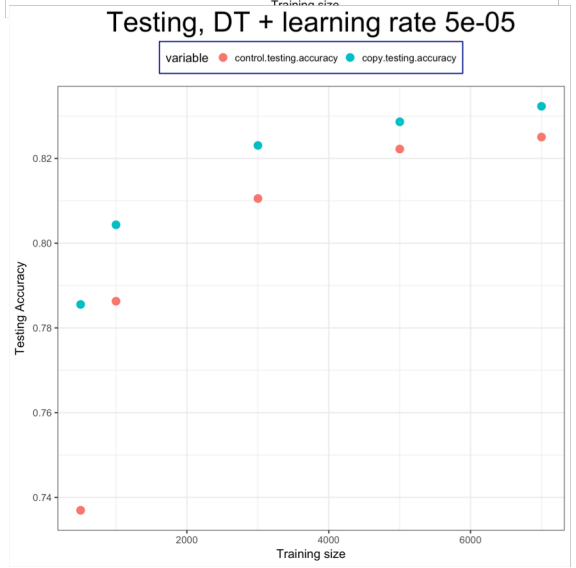
Training accuracy (average)



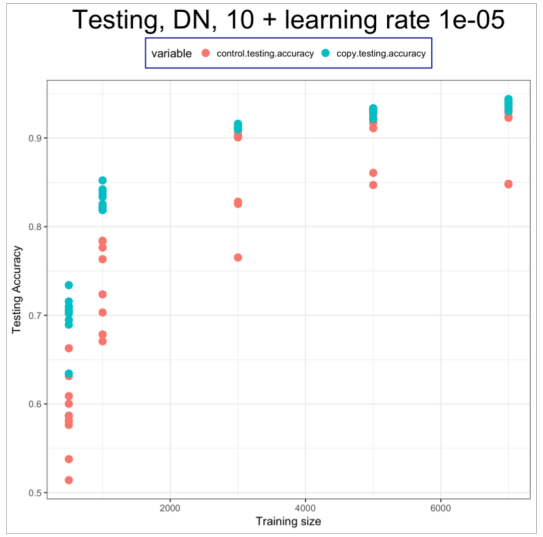


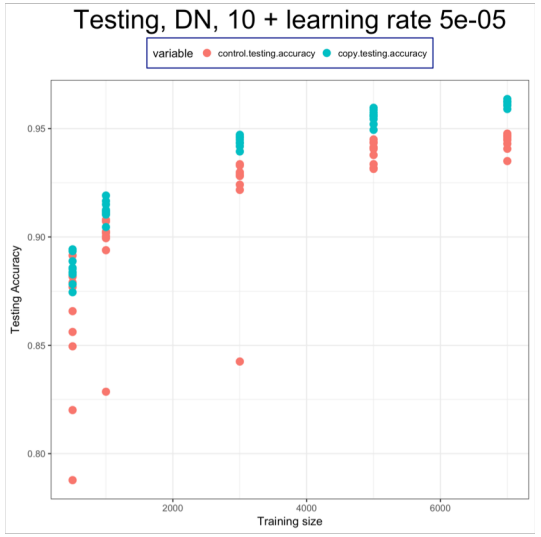


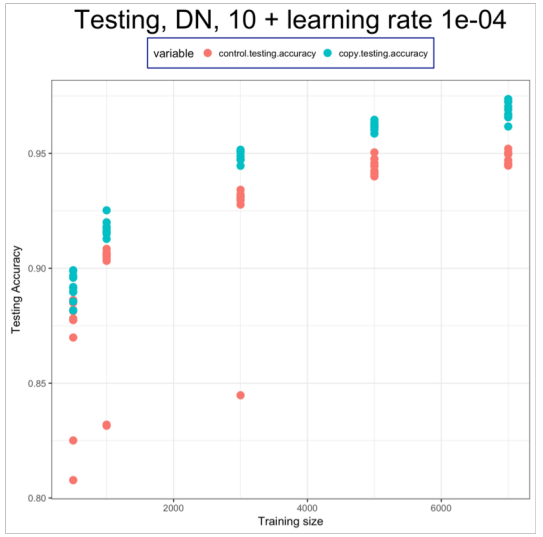


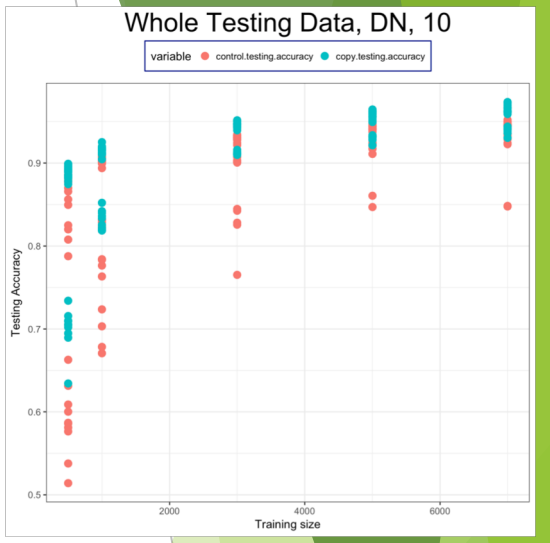


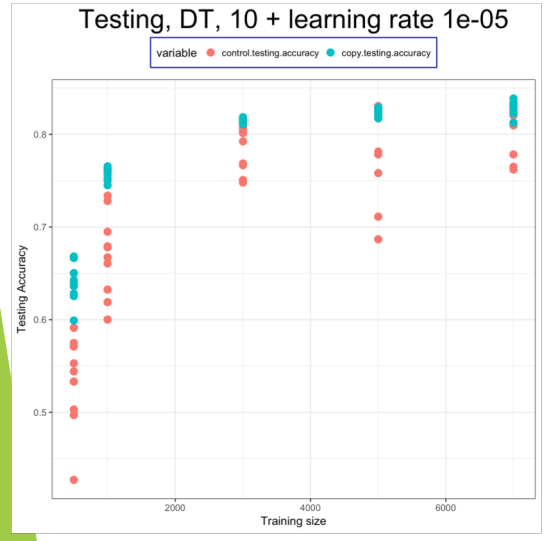
Testing accuracy (not average)

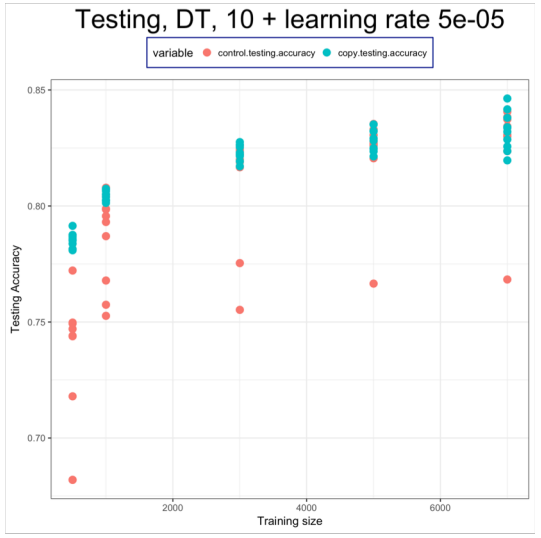


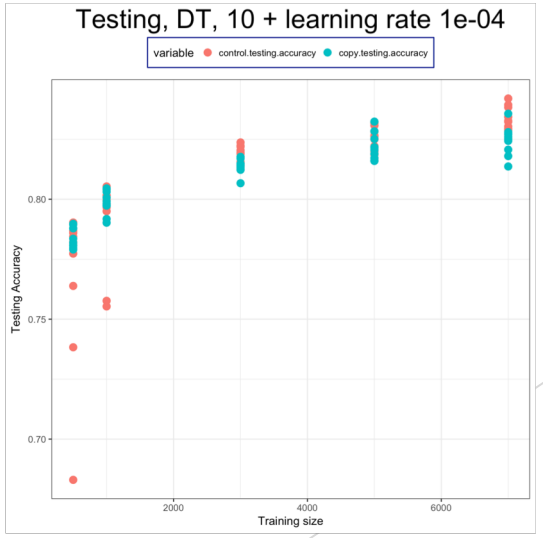


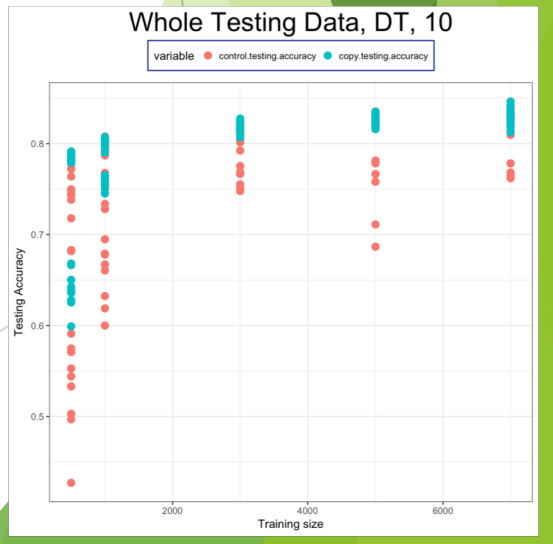












Testing accuracy (average)