README

Result (empirical accuracy)

Flow Chart

Technical details

Conclusion & Discussion

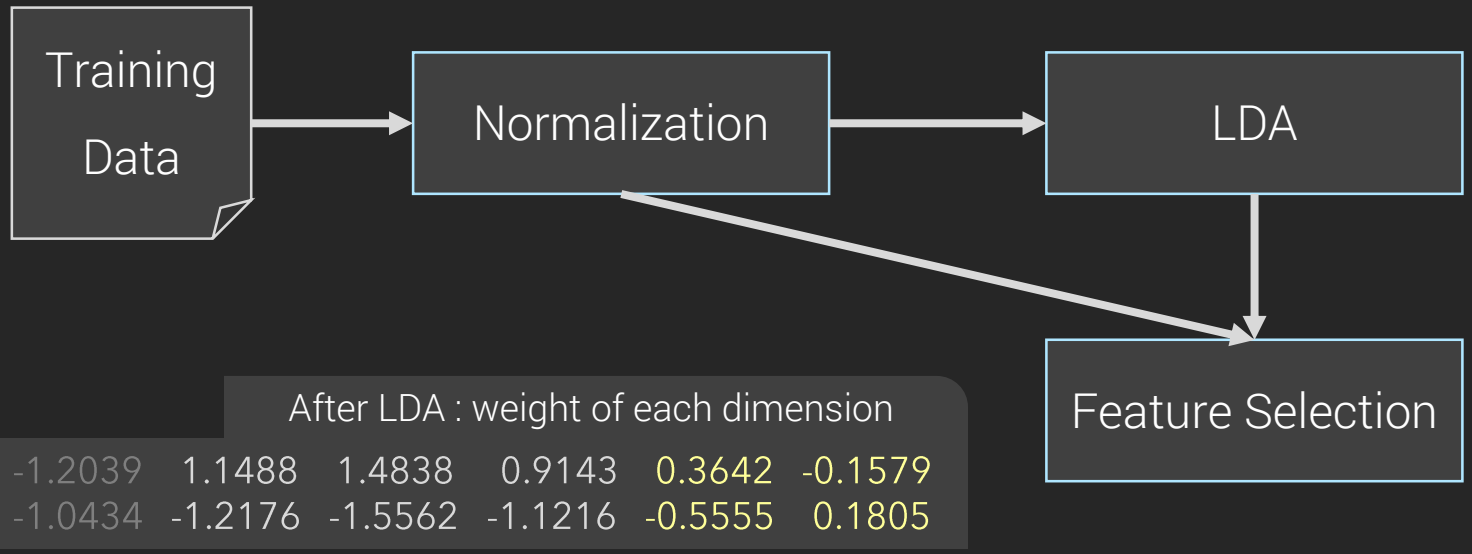
103062528 林玉山

103062518 李元正

Training Accuracy	100%
Training Error	0%

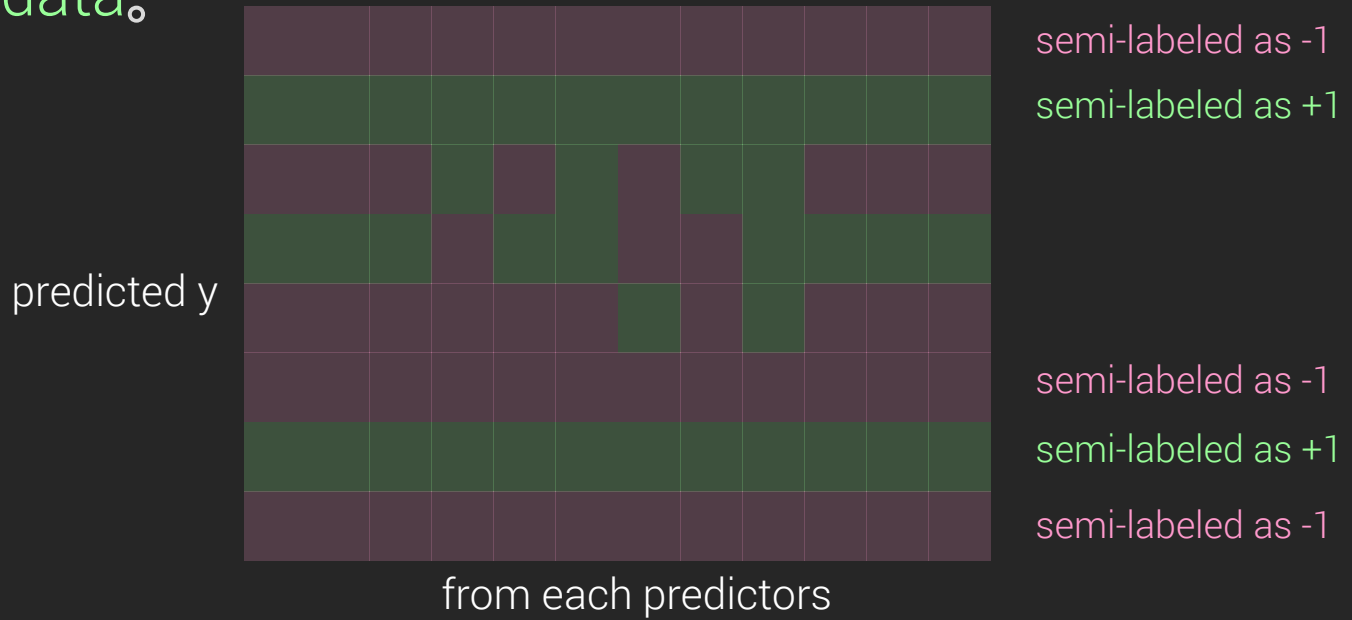
Testing Accuracy	96.00%
Testing Error	4.00%

Flow Chart : Feature Selection



在LDA的結果中，我們發現資料的第4、第5維度的feature作用在線性判別的效益較低，因此我們執行feature selection，刪去資料的第4、第5維度的feature

在調整hyper-parameter 的過程中，有一些unlabeled data，無論我們如何調整參數，都得到相同的predicted value，我們稱這些穩定(invariant to hyper-parameter)的data為semi-labeled data。



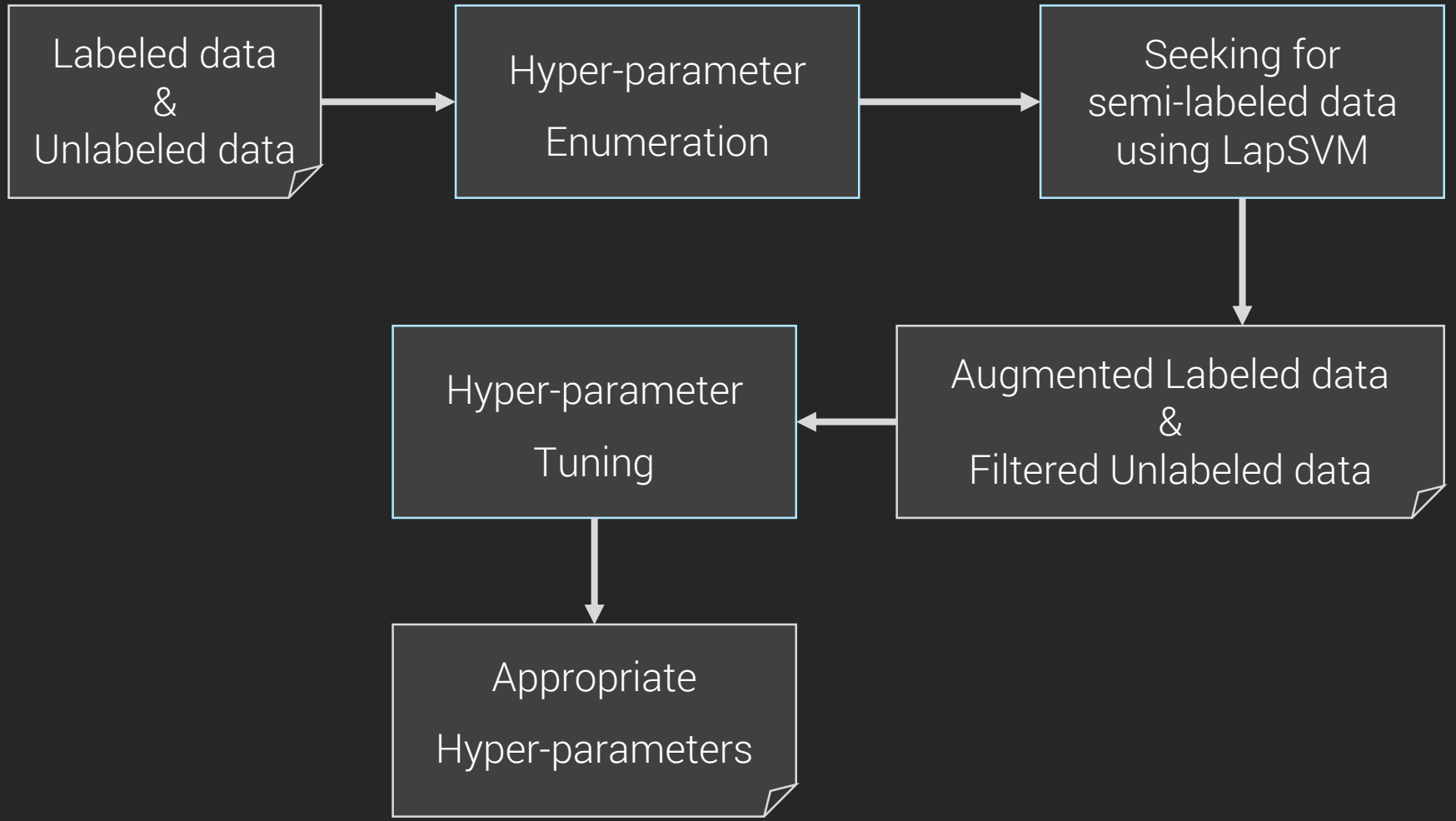
在我們調整hyper-parameter時，我們發現這種semi-labeled data甚至比labeled data更穩定。

在cross-validation調整hyper-parameter的階段，我們還不能偷看testing data，而training data中的labeled data的數量又太少(僅50筆)，造成cross-validation的代表性不足。

因此我們將semi-labeled data也視為labeled data，藉此擴充semi-labeled data。

在Cross-Validation時，labeled data有50筆，
semi-labeled data有163筆(213/400)

Flow Chart : Cross-validation (Hyper-parameter tuning)



Semi-labeled data : in training (1/2)

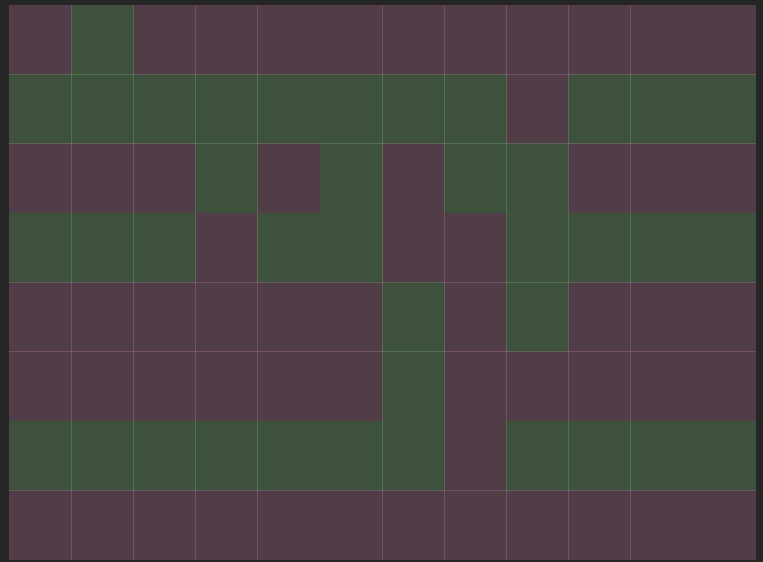
在training的過程中，我們也使用semi-labeled data的概念，不過這次定義有所修改。

在cross-validation中，我們調整參數的步進為10倍，例如 $\gamma_k = \{ 5, 50, 500 \}$，如果我們發現$\gamma_k$=50運作良好，那麼或許其實$\gamma_k$=25和$\gamma_k$=100也運作地大致良好。

因此在training的過程中，我們也調整hyper-parameter，只不過調整的步進為2倍。如此，對於同樣的一份training data，我們可以產生許多份predicted values。

Semi-labeled data : in training (2/2)

既然有了許多份predicted values，我們就可以為training的過程重新定義semi-labeled data：凡是這些predicted values給出高度一致性(高達8成)的結果的data，為semi-labeled data。



現在，我們有了良好的hyper-parameters，以及semi-labeled data。

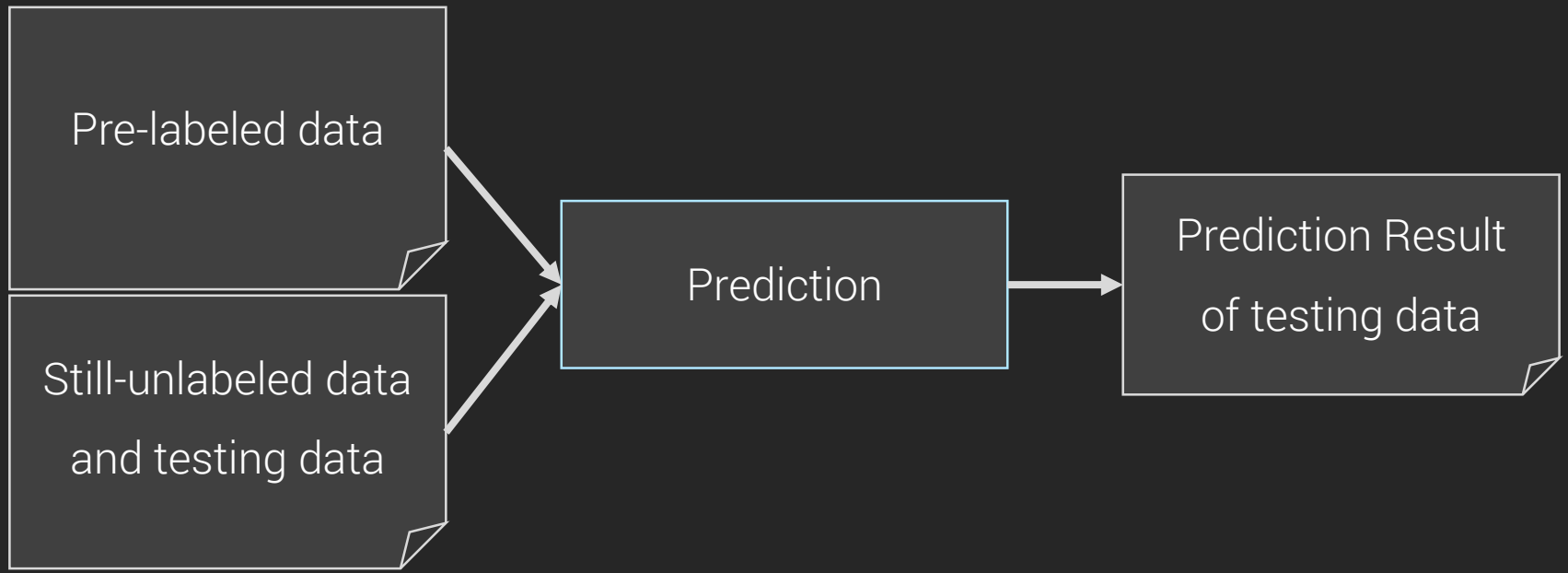
為了避免在training的過程中，我們強制將不確定的data賦予label，導致testing時誤差放大並傳遞，因此，我們僅將labeled data和semi-labeled data賦予label，而剩餘的(不確定的)data在prediction 時併入testing data。

在training時，labeled data有50筆，
semi-labeled data有198筆(248/400)

Flow Chart : Training

礙於篇幅，省略normalization
和feature selection等過程

Flow Chart : Testing



透過LDA(或者QDA)，我們可以輕易地分辨feature的那些維度對於整體classification更有幫助。

semi-labeled的構想是我們本次project的核心價值所在，我們應該只相信那些對自己有信心的data，而對於那些對自己的predicted value沒什麼信心的data，我們則僅參考它的feature，而不參考它在training時所得到的結果，如此我們便避免了誤差傳遞造成不可回復的錯誤結果，從而提升整體辨識準確度。