Spurious Correlations in Deep Learning Models

Kevin Quinn

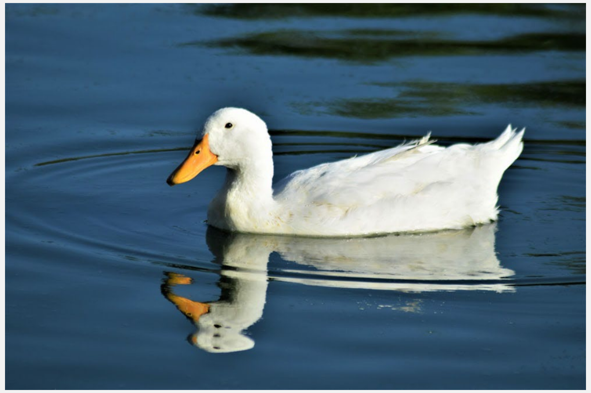
Motivation

- **Spuriously Correlated** data is correlated with, but not actually predictive of a given target variable

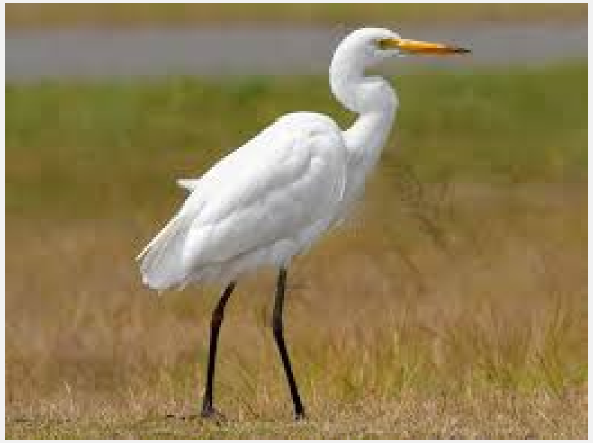
Training:

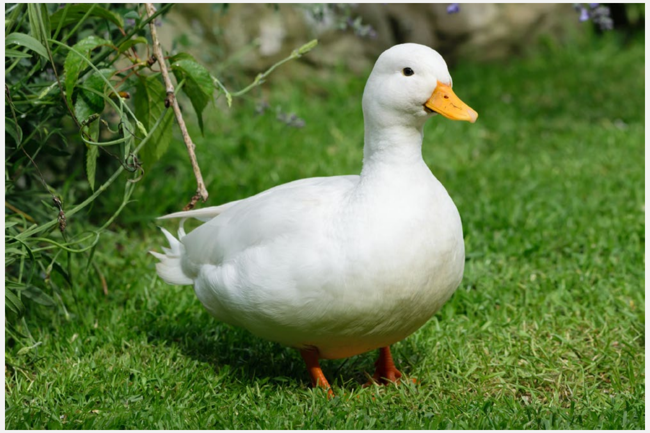
Testing:

Duck



Egret





Egret

Recent Work

- A lot of really interesting recent work addresses this issue!

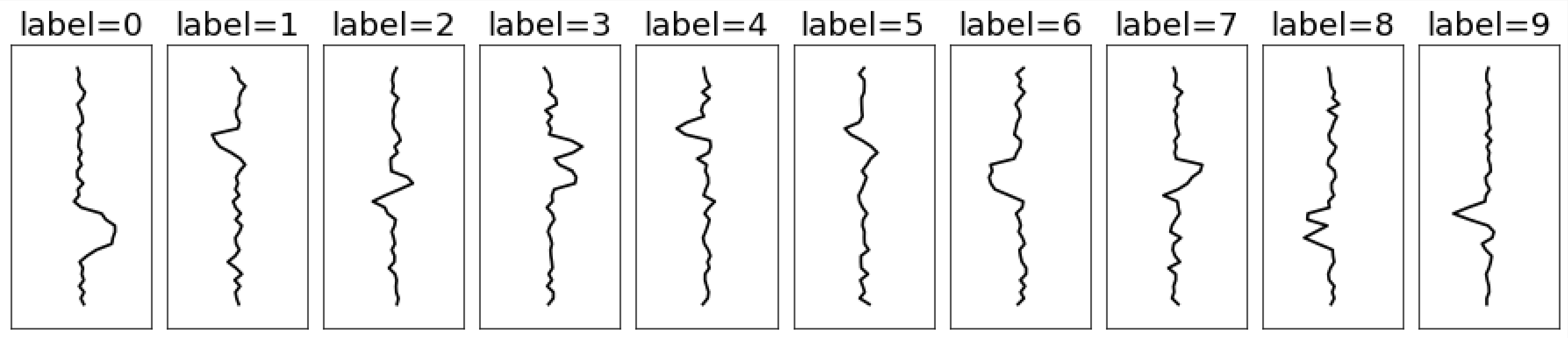
1. *[Last Layer Re-training Is Sufficient For Robustness To Spurious Correlations,* Kirichenko, Izmailov, Gordon Wilson ICLR 2023]
   - **Retrain** the final linear classification layer of a model on new data where spurious correlations aren't present

2. [*Simple and Fast Group Robustness by Automatic Feature Reweighting* Qiu, Potapczynski, Izmailov, Gordon Wilson ICML 2023]
   - **Re-weight** the final linear classification layer of a model by giving more importance to training samples in the minority

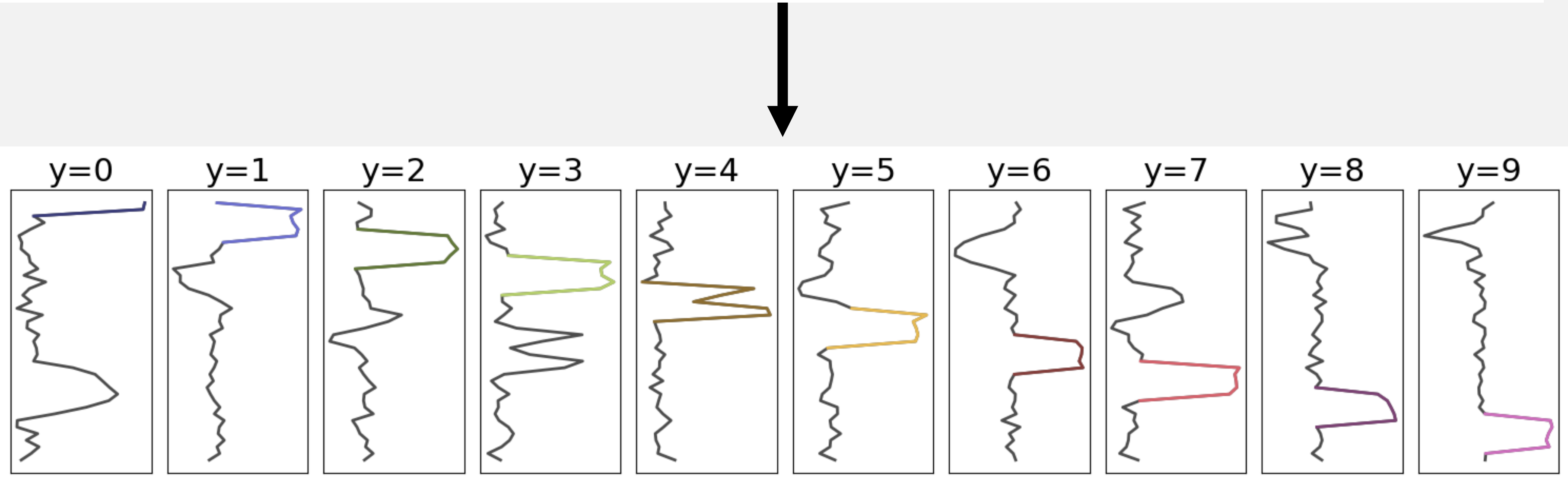
Dataset

- What can we learn from a simplified setting of the problem?

MNIST-1d

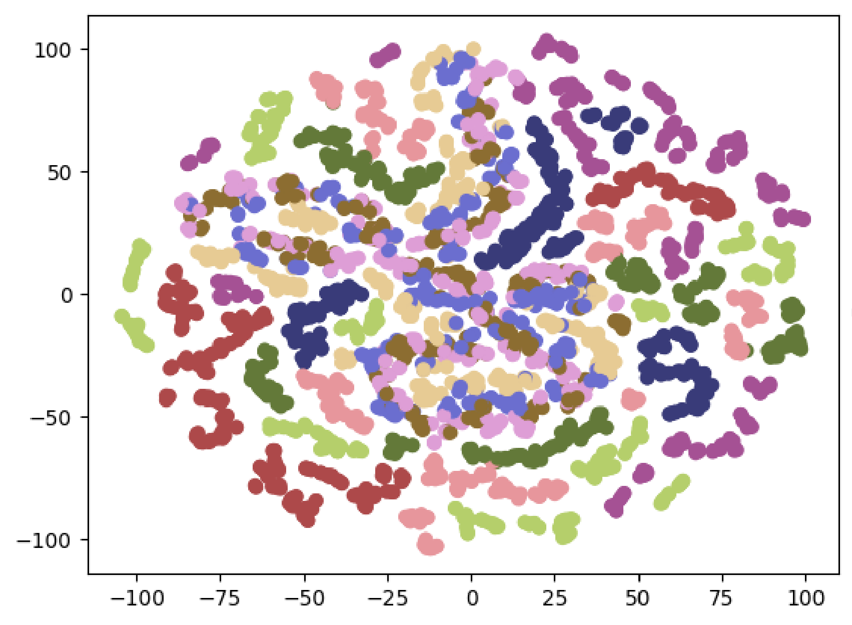


MNIST-1d with spurious correlations

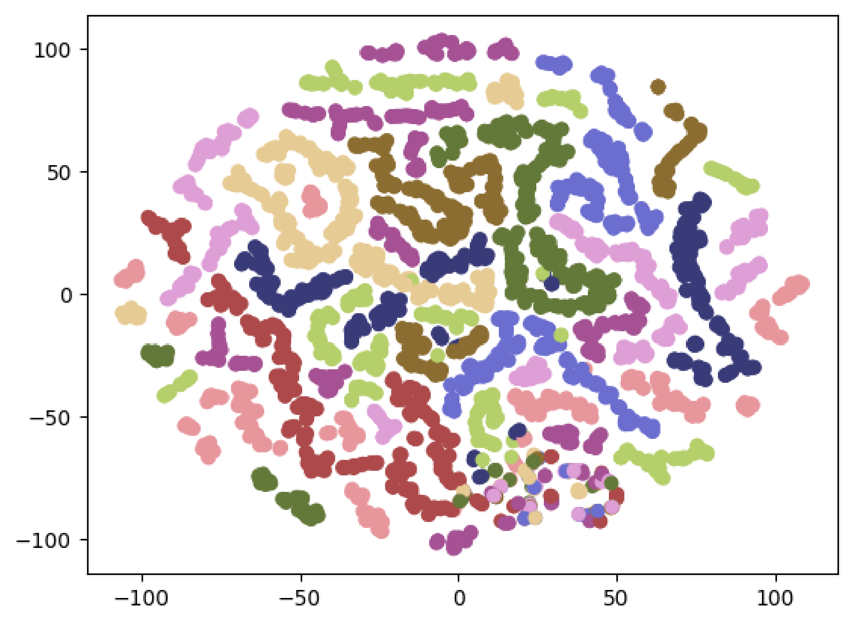


Dataset

- 2d embeddings with t-SNE show how the training data becomes ‘artificially’ easier to predict



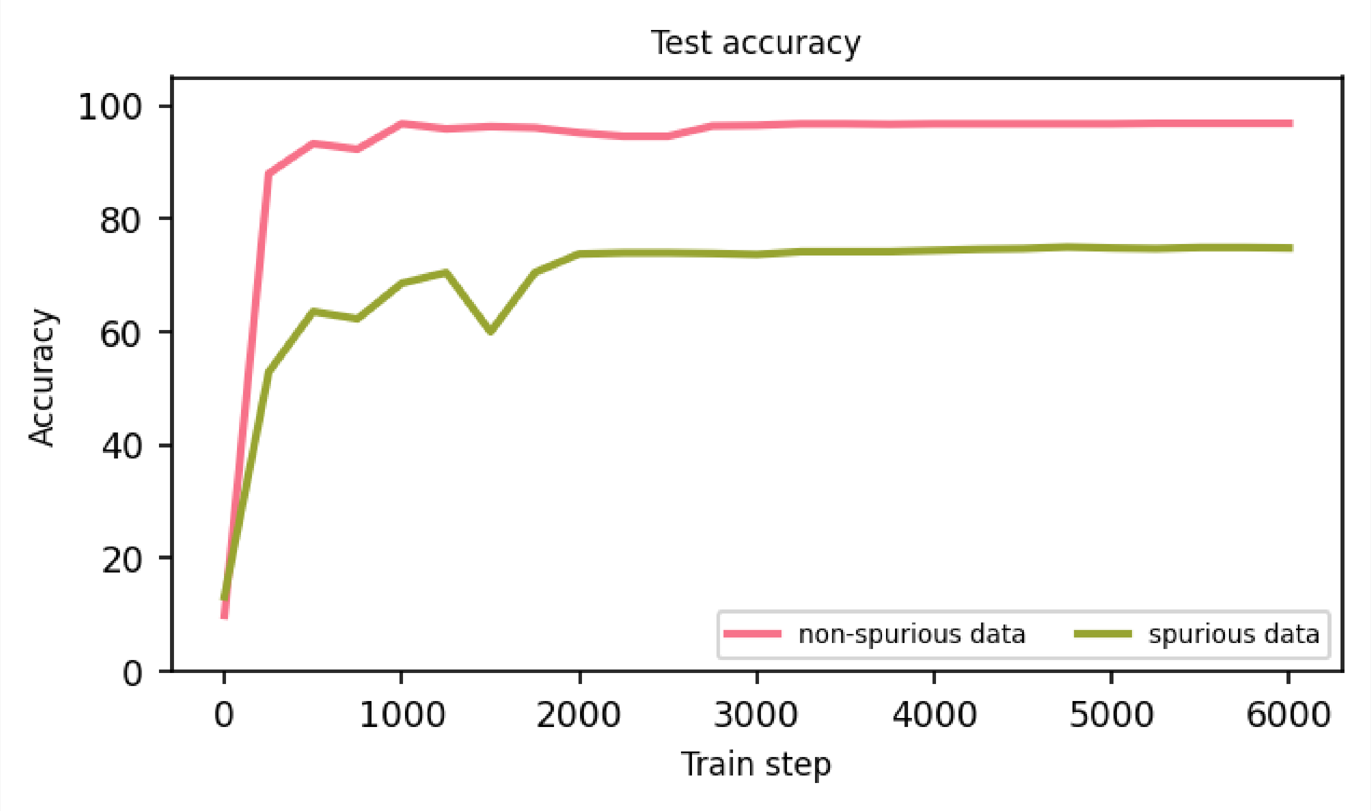
MNIST-1d



MNIST-1d
with spurious
correlations

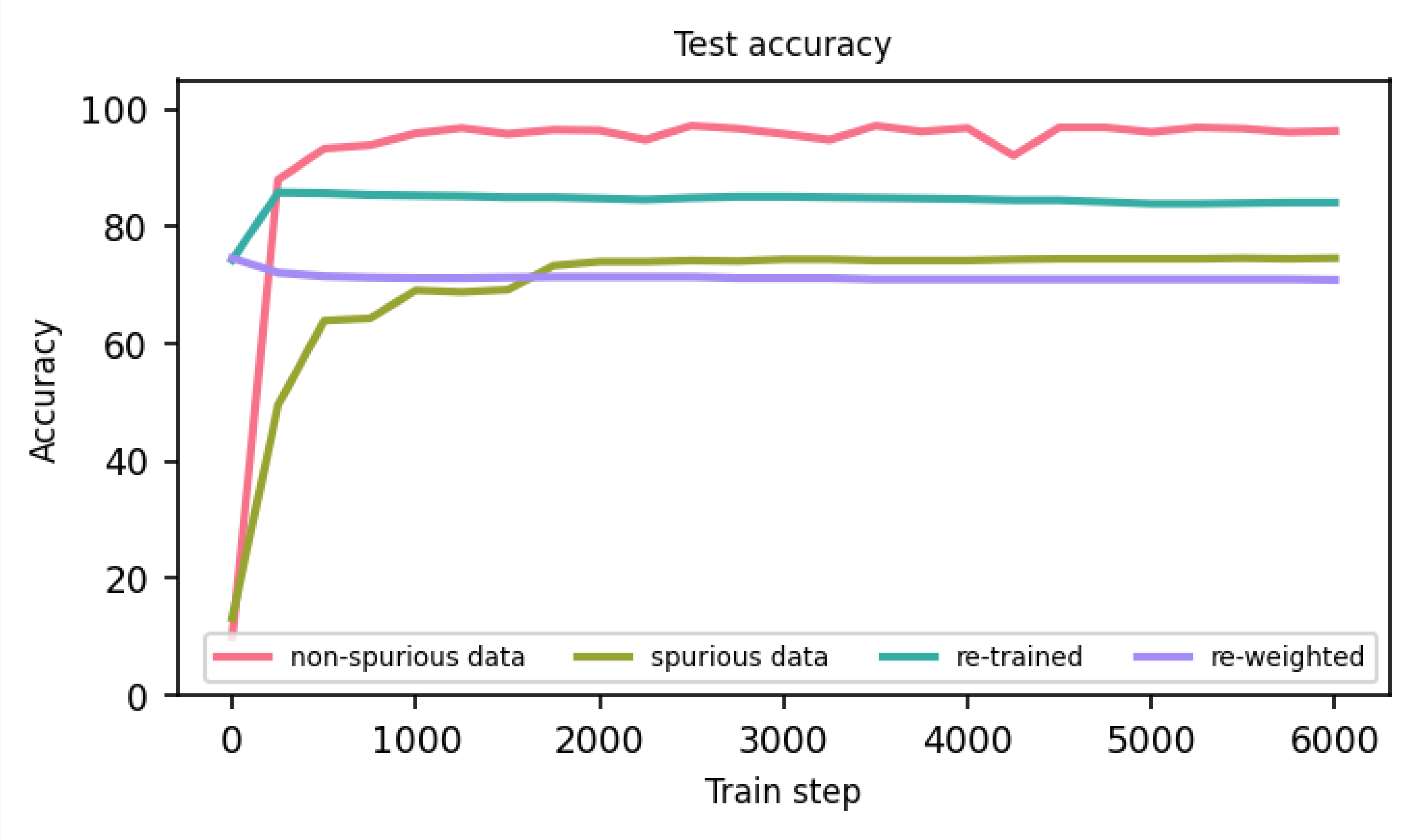
Experiments

- But this comes with the cost of inaccuracy when test data is missing the same spuriously correlated features



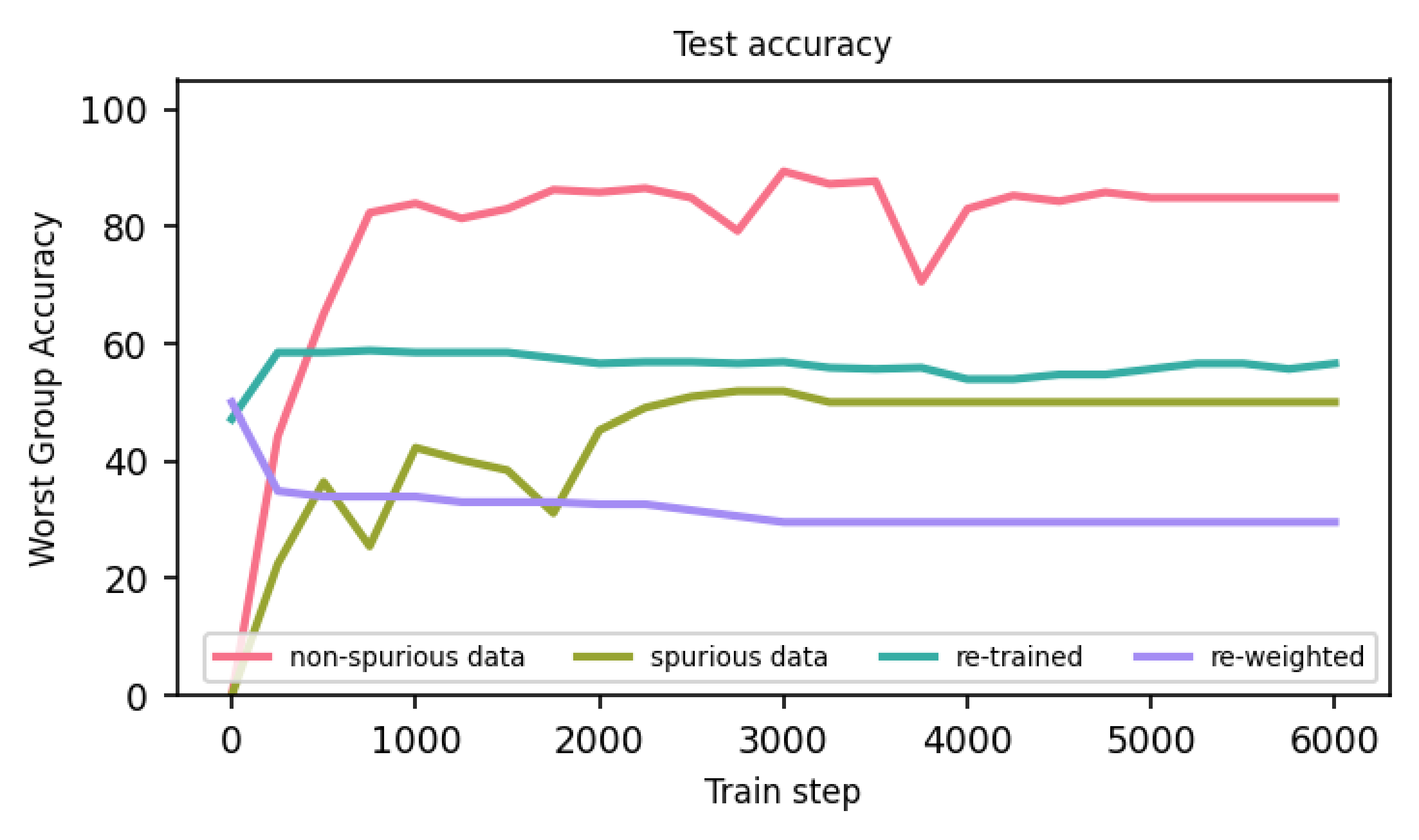
Experiments

- Some strategies from [1] and [2] attempt to address the problem via re-training or re-weighting the model



Experiments

- This effect is even worse when measuring worst group accuracy!



Questions

- How much spurious noise can we add before the problem becomes intractable?
- How much data do we need to re-train with in order to fix a model?
- Why does re-weighting perform poorly in this example?
- Can we fix it by giving more attention to the group with the worst accuracy?