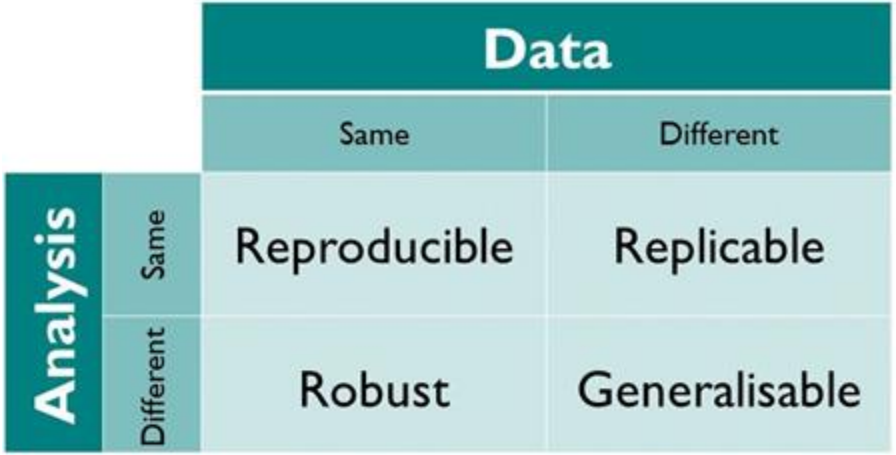
When reproducibility is not enough

Luca Ferranti, Aalto University

THIS IS THE WORST
EXPERIMENT I HAVE EVER REPRODUCED
BUT YOU DID REPRODUCE IT
imgflip.com

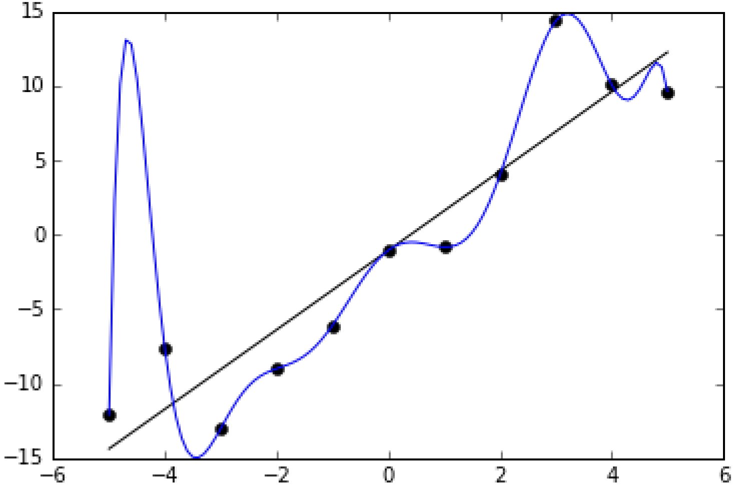
Reproducible < Replicable

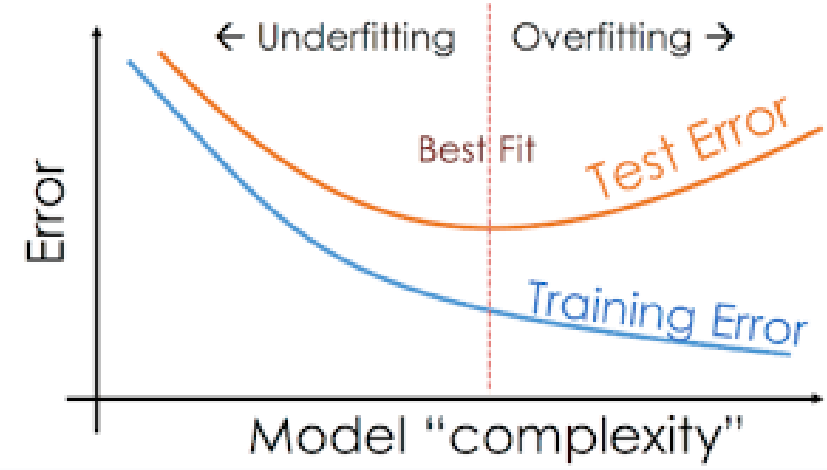


		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Issue: your experiments work only with a specific set of data under a very specific setup

What is overfitting?



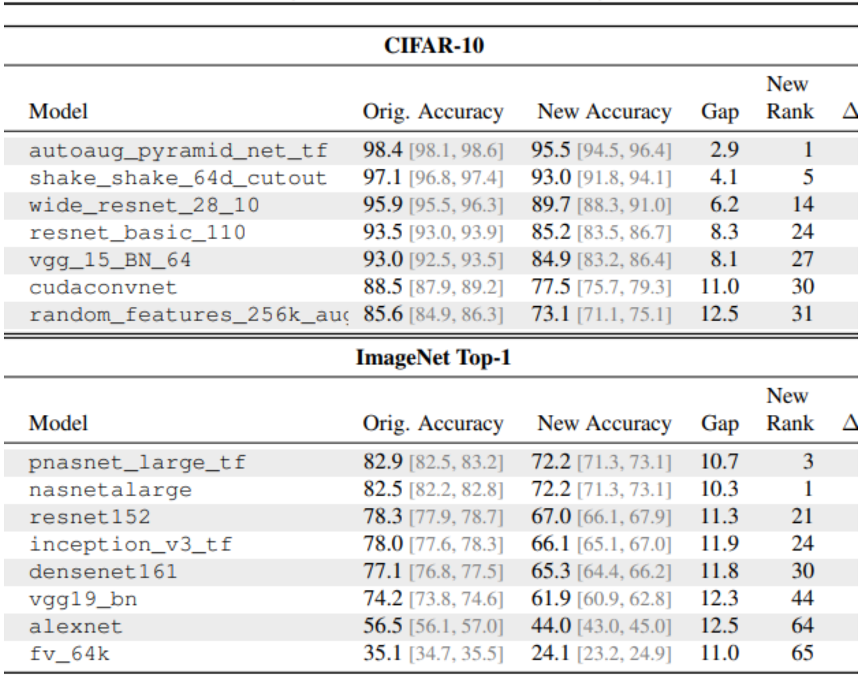


What is overfitting here?

- Generally overfitting is referred *to training data*
- However, one can have overfitting **to whole datasets / setups**
    - Results only with a given dataset and worse elsewhere
    - Results only work for a specific randomseed (fixed train-split)
    - Data leakage between training and test data
    - Tuned to one metric / leaderboard

Datasets overfitting

[ref] Do ImageNet Classifiers Generalize to ImageNet?
https://proceedings.mlr.press/v97/recht19a/recht19a.pdf?utm_source=chatgpt.com



CIFAR-10

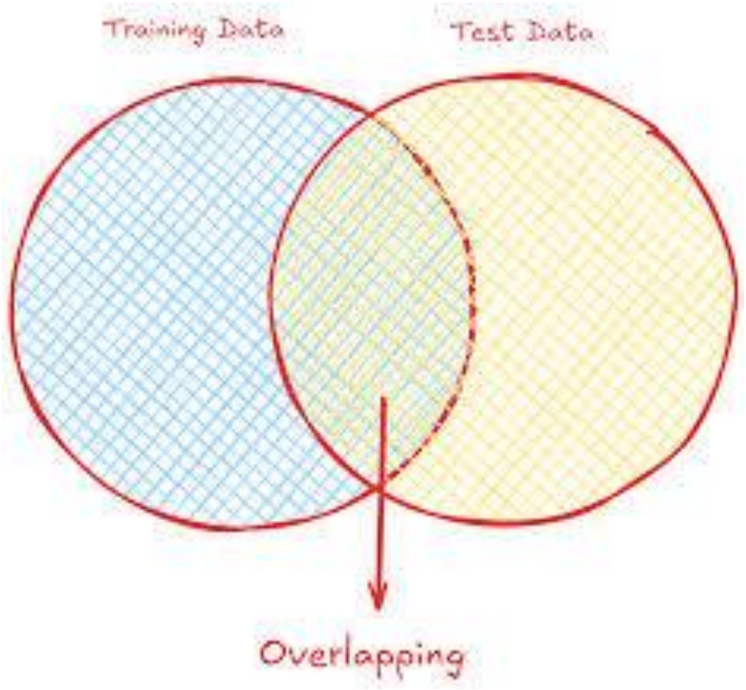
Model	Orig. Accuracy	New Accuracy	Gap	New Rank	Δ
autoaug_pyramid_net_tf	**98.4** [98.1, 98.6]	**95.5** [94.5, 96.4]	2.9	1	
shake_shake_64d_cutout	**97.1** [96.8, 97.4]	**93.0** [91.8, 94.1]	4.1	5	
wide_resnet_28_10	**95.9** [95.5, 96.3]	**89.7** [88.3, 91.0]	6.2	14	
resnet_basic_110	**93.5** [93.0, 93.9]	**85.2** [83.5, 86.7]	8.3	24	
vgg_15_BN_64	**93.0** [92.5, 93.5]	**84.9** [83.2, 86.4]	8.1	27	
cudaconvnet	**88.5** [87.9, 89.2]	**77.5** [75.7, 79.3]	11.0	30	
random_features_256k_au	**85.6** [84.9, 86.3]	**73.1** [71.1, 75.1]	12.5	31	

ImageNet Top-1

Model	Orig. Accuracy	New Accuracy	Gap	New Rank	Δ
pnasnet_large_tf	**82.9** [82.5, 83.2]	**72.2** [71.3, 73.1]	10.7	3	
nasnetalarge	**82.5** [82.2, 82.8]	**72.2** [71.3, 73.1]	10.3	1	
resnet152	**78.3** [77.9, 78.7]	**67.0** [66.1, 67.9]	11.3	21	
inception_v3_tf	**78.0** [77.6, 78.3]	**66.1** [65.1, 67.0]	11.9	24	
densenet161	**77.1** [76.8, 77.5]	**65.3** [64.4, 66.2]	11.8	30	
vgg19_bn	**74.2** [73.8, 74.6]	**61.9** [60.9, 62.8]	12.3	44	
alexnet	**56.5** [56.1, 57.0]	**44.0** [43.0, 45.0]	12.5	64	
fv_64k	**35.1** [34.7, 35.5]	**24.1** [23.2, 24.9]	11.0	65	

Data leakage

- Accidentally using test data for the training

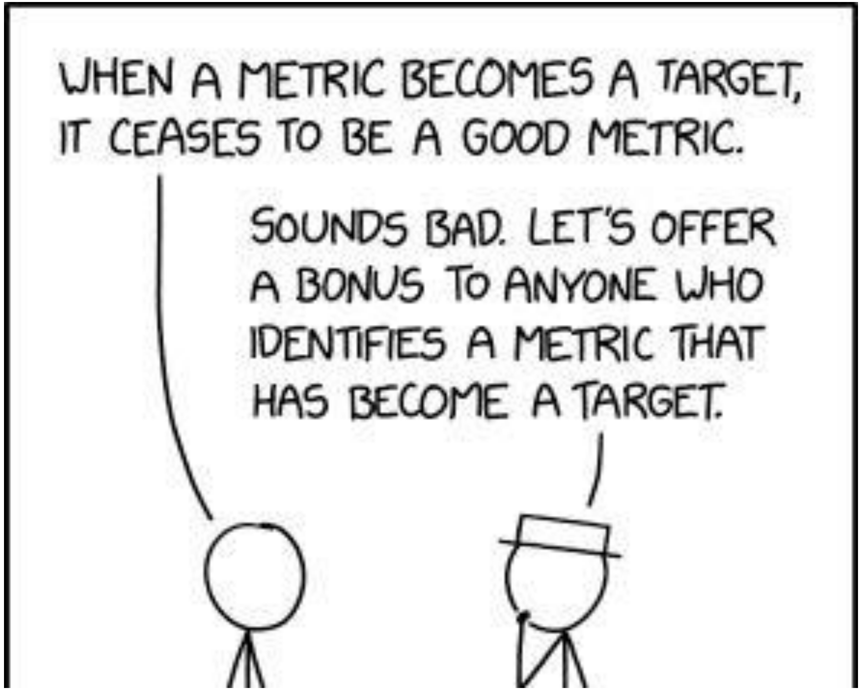


Hyperparameters tuning

- Large hyperparameter sweeps on a single validation split
- Picking the single best run = picking the top of a noisy distribution
- Pipeline is fully specified and reproducible…
- …but the claim “method X is better” is not robust

Leaderboard optimization

- Repeat experiments and choose model / parameters that make your test data perform better in a given leaderboard


WHEN A METRIC BECOMES A TARGET, IT CEASES TO BE A GOOD METRIC.
SOUNDS BAD. LET'S OFFER A BONUS TO ANYONE WHO IDENTIFIES A METRIC THAT HAS BECOME A TARGET.

How to mitigate?

- Test with multiple random seeds
- Test on multiple data sets
- Do ablation study
- Don’t choose hyperparameters based on test data results

reproducibility is the beginning not the end!

Beware, **reproducibility is the beginning not the end!**

- You can reproduce results that
  - work only in one specific case
  - work because of hacks

Once you have learnt to make reproducible research, make it replicable!