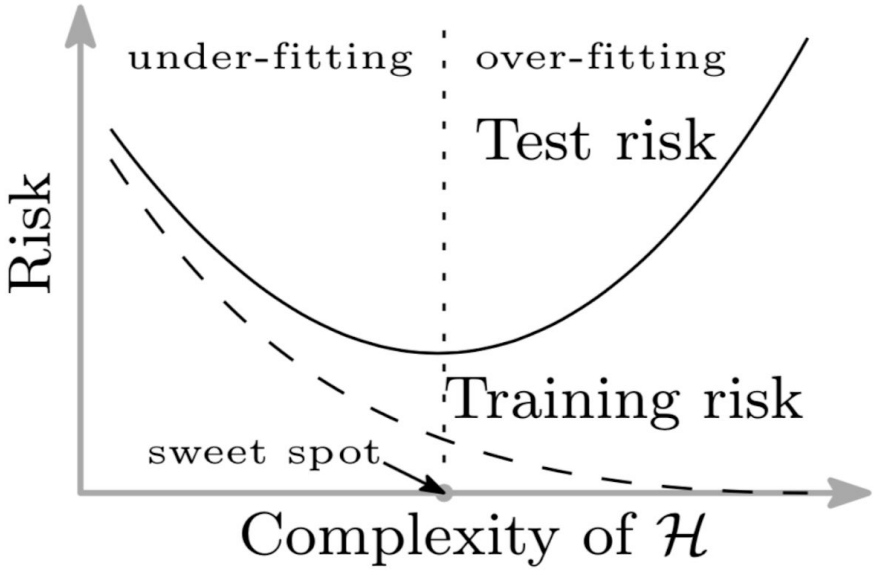
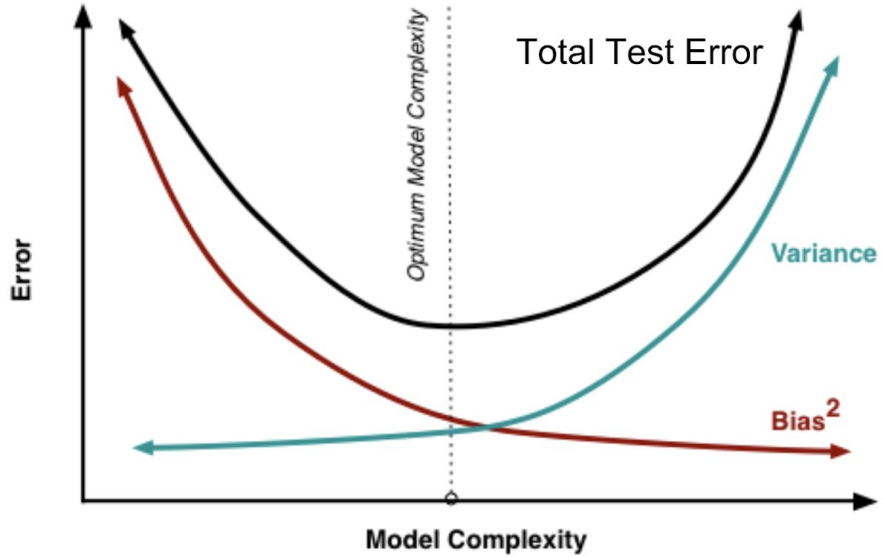
Are deep NN (dramatically) overfitted?

Deep Double Descent & Lottery Ticket Hypothesis

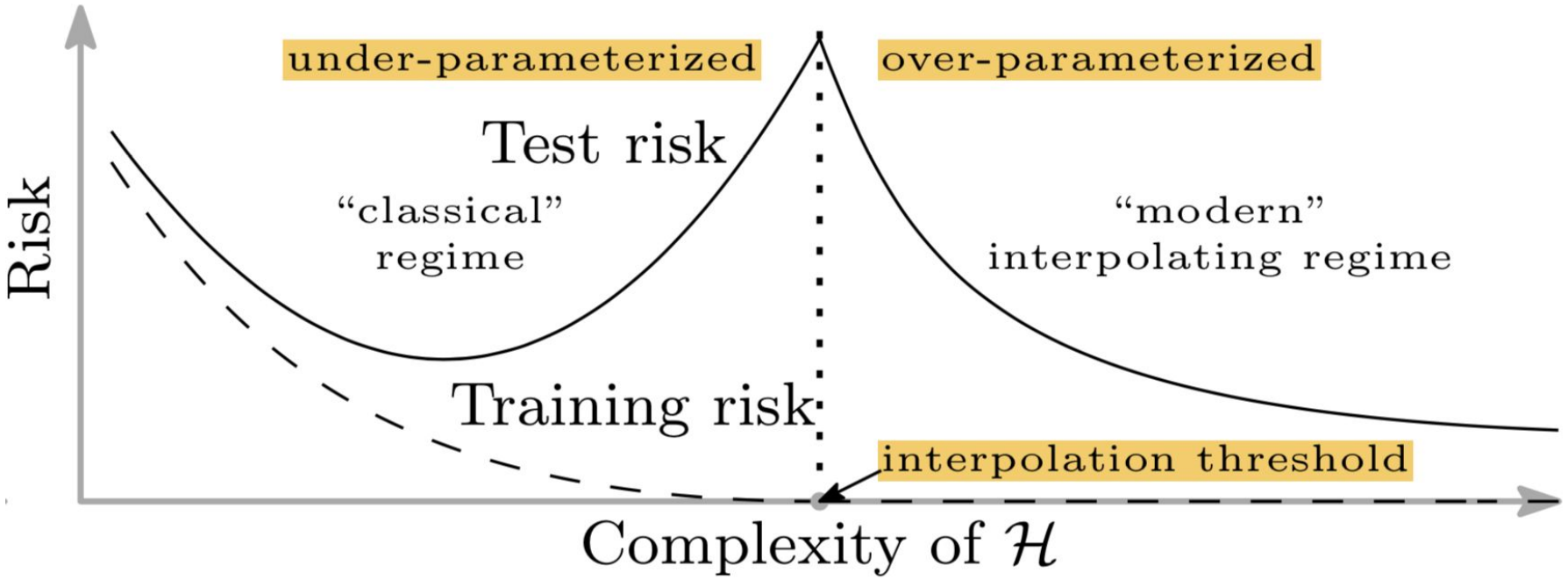
Traditional risk curve & bias-variance trade-off



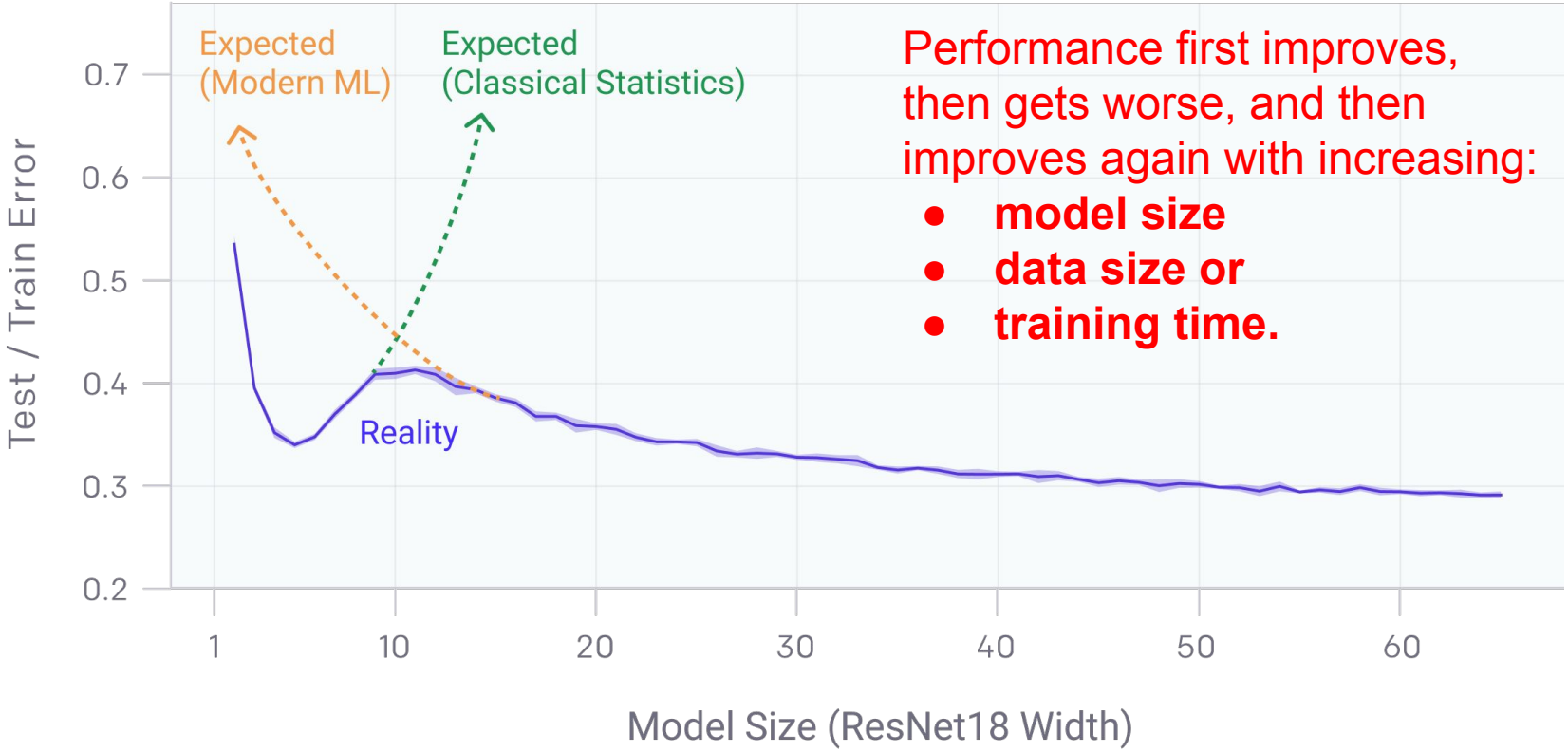


Traditional machine learning uses a U-shape risk curve to measure the bias-variance trade-off and quantify how generalizable a model is.

New double-U-shaped bias-variance risk curve for deep NN



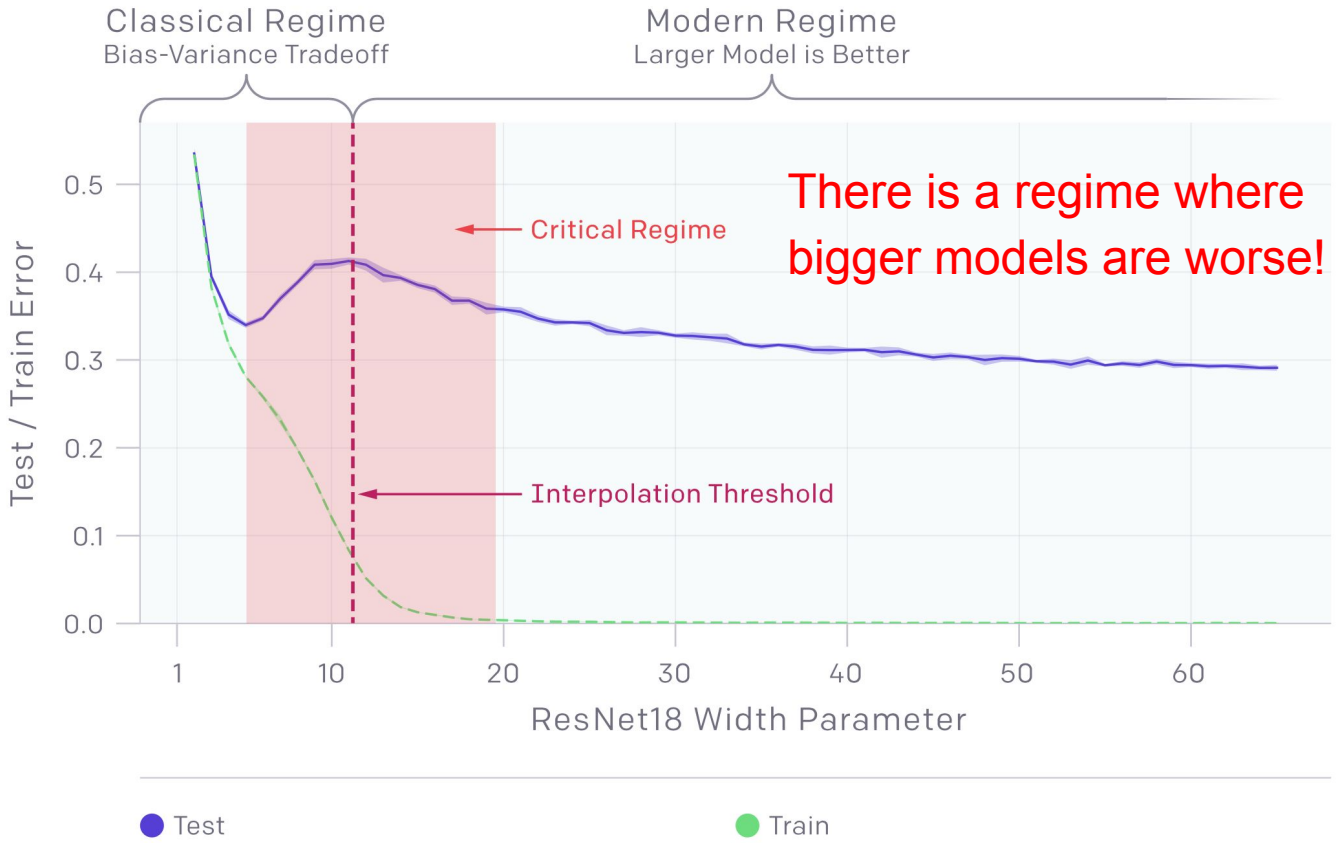
Deep Double Descent?



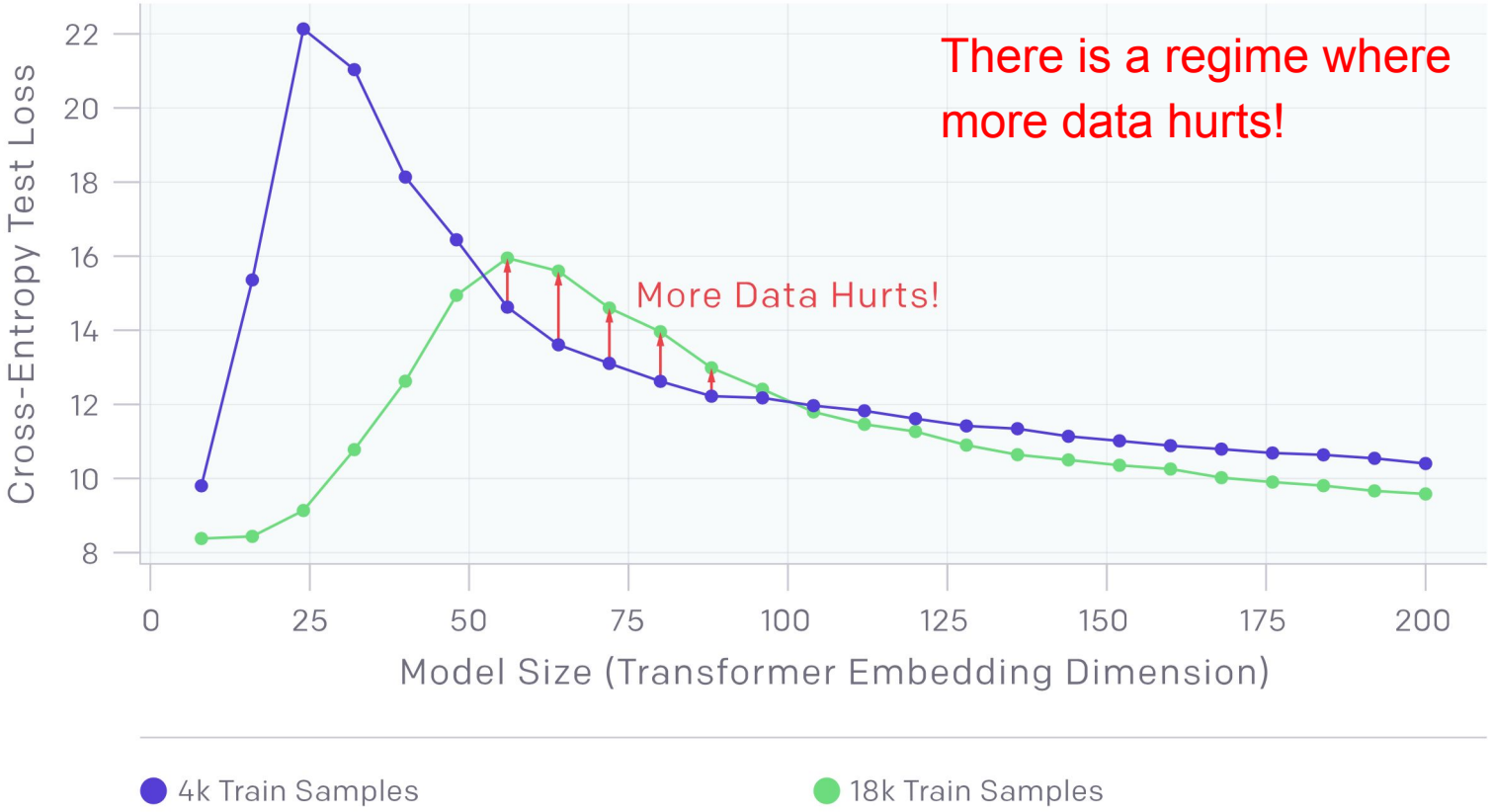
Performance first improves, then gets worse, and then improves again with increasing:

- **model size**
- **data size or**
- **training time.**

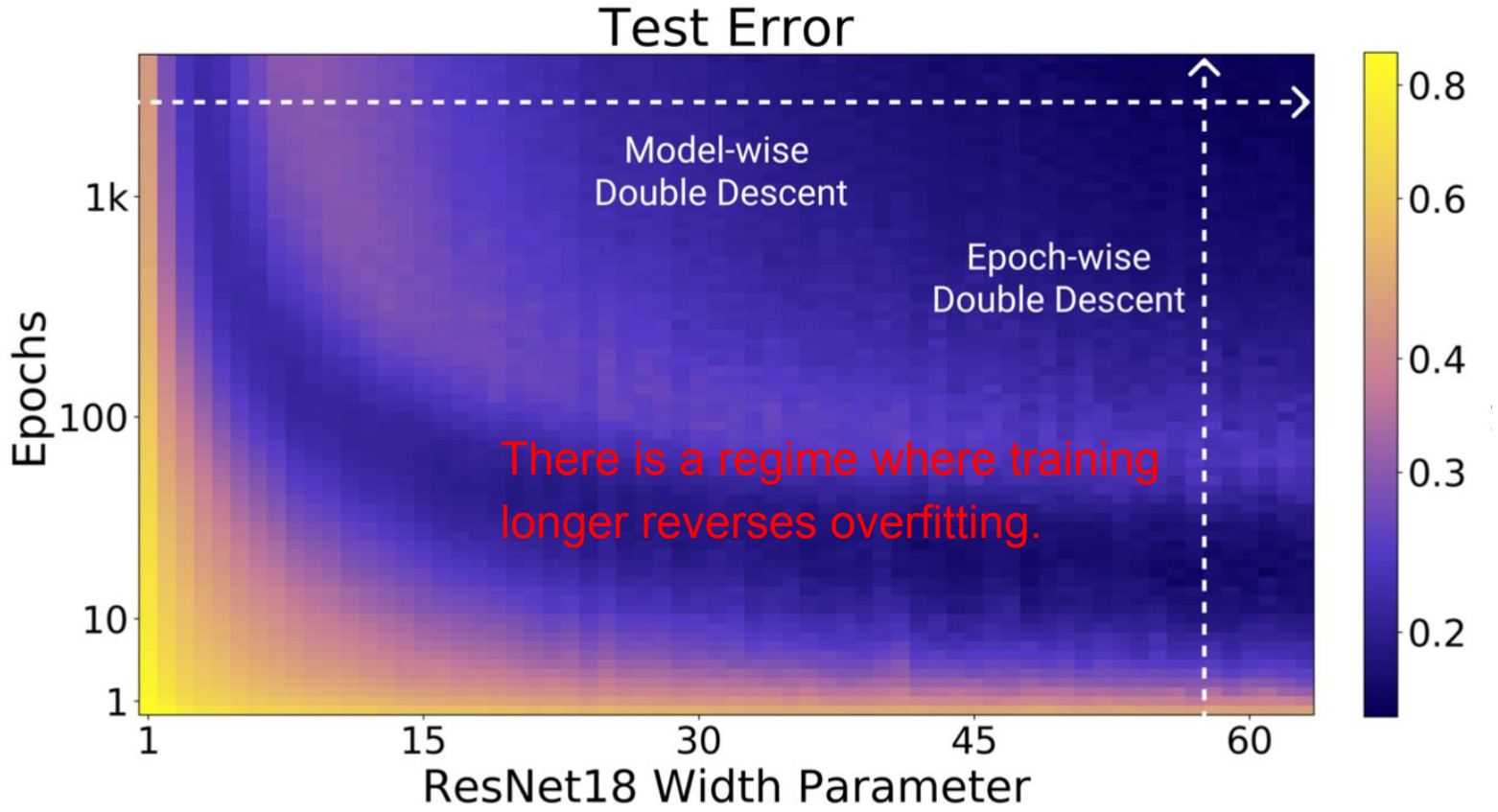
Model-wise double descent



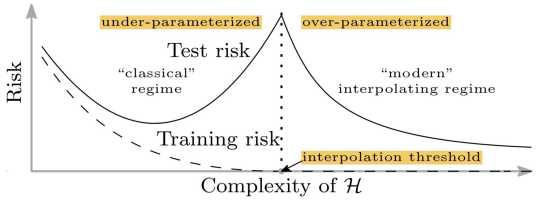
Sample-wise non-monotonicity



Epoch-wise double descent



Intuition?



At the interpolation threshold, there is effectively only one model that fits the train data, and forcing it to fit even slightly noisy or misspecified labels will destroy its global structure. There are no "good models" which both interpolate the train set and perform well on the test set.

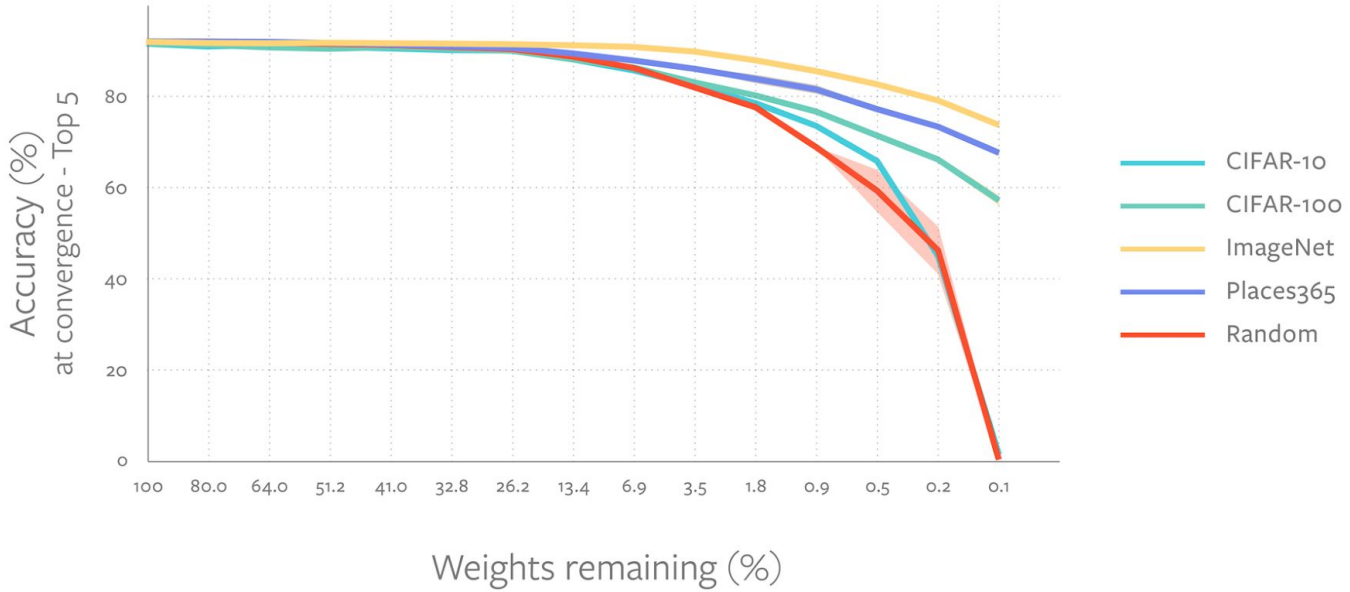
In the over-parameterized regime, there are many models that fit the train set and there exist such good models.

The implicit bias of stochastic gradient descent (SGD) leads it to such good models, for reasons we don't yet understand.

Lottery ticket hypothesis

Click for FAIR blog post with video.

Generalization of winning tickets?



Each line represents a different source dataset for the winning ticket.

Winning tickets generated on ImageNet and Places365 consistently outperformed those generated on smaller datasets on ImageNet.

Sources:

https://lilianweng.github.io/lil-log/2019/03/14/are-deep-neural-networks-dramatically-overfitted.html

https://openai.com/blog/deep-double-descent/

https://www.lesswrong.com/posts/FRv7ryoqtvSuqBxuT/understanding-deep-double-descent

https://ai.facebook.com/blog/understanding-the-generalization-of-lottery-tickets-in-neural-networks/

Compiled by Michael M. Pieler