Mean teachers are better role models:

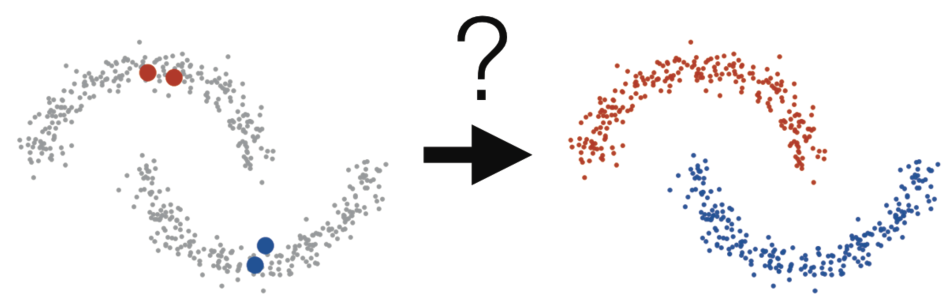
Weight-averaged consistency targets improve semi-supervised deep learning results

Antti Tarvainen & Harri Valpola

THE CURIOUS AI COMPANY

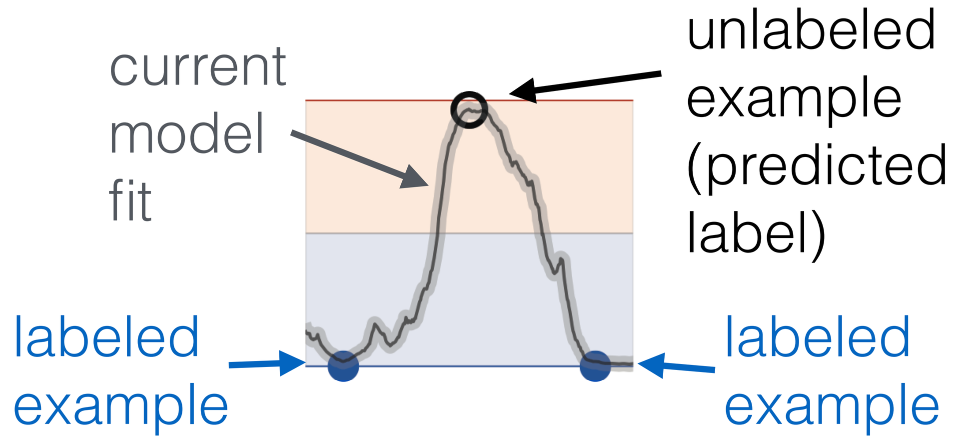
History: Applying noise to propagate labels

One way of doing semi-supervised learning is to propagate label information through the data to previously unlabeled examples.
If two examples are similar, they probably have the same label. This idea can be used in training by perturbing the input or an intermediate representation slightly and requiring that the label remains the same.

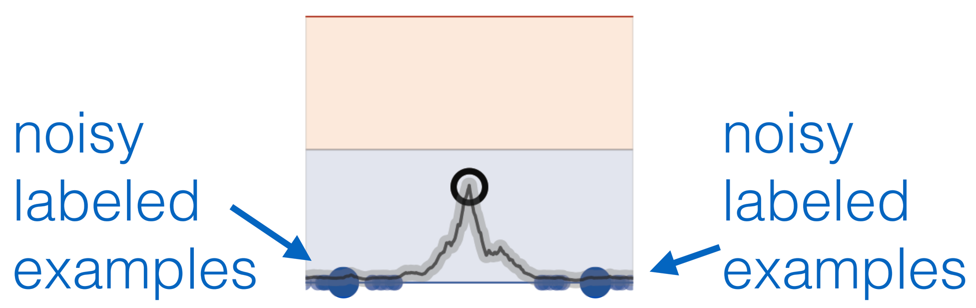


(Imagine that this happens simultaneously in many dimensions, many of which are learned intermediate representations.)

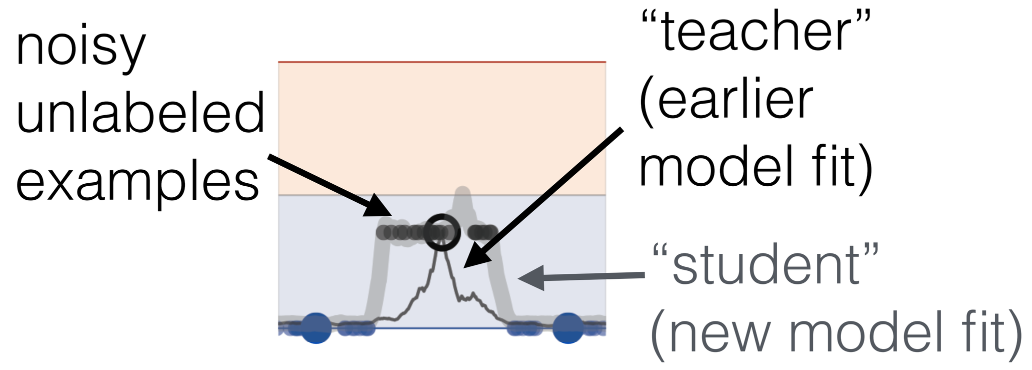
Without regularisation, the model is free to choose any function that fits the labeled data points.
Problem: Does not generalise well.



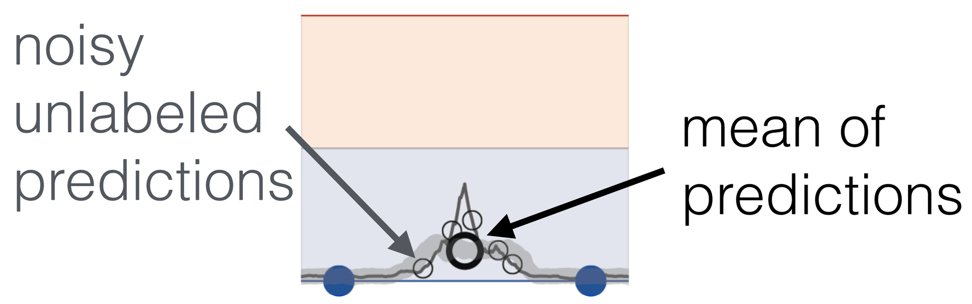
2012: **Dropout** [1] and other types of perturbation smoothen the fitted function around labeled examples.
Problem: Does not exploit unlabeled data.



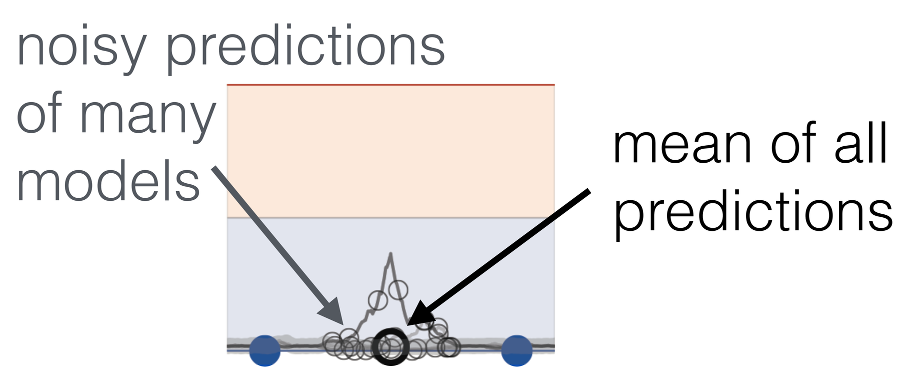
2015: **Virtual adversarial training** [2] and **Ladder Γ model** [3] smoothen the function around both labeled and unlabeled data.
Problem: Biased towards current predictions.



2016: **Π model** [4] applies noise also on the teacher side to obtain a better approximation of the correct label.
Problem: Still fairly biased.



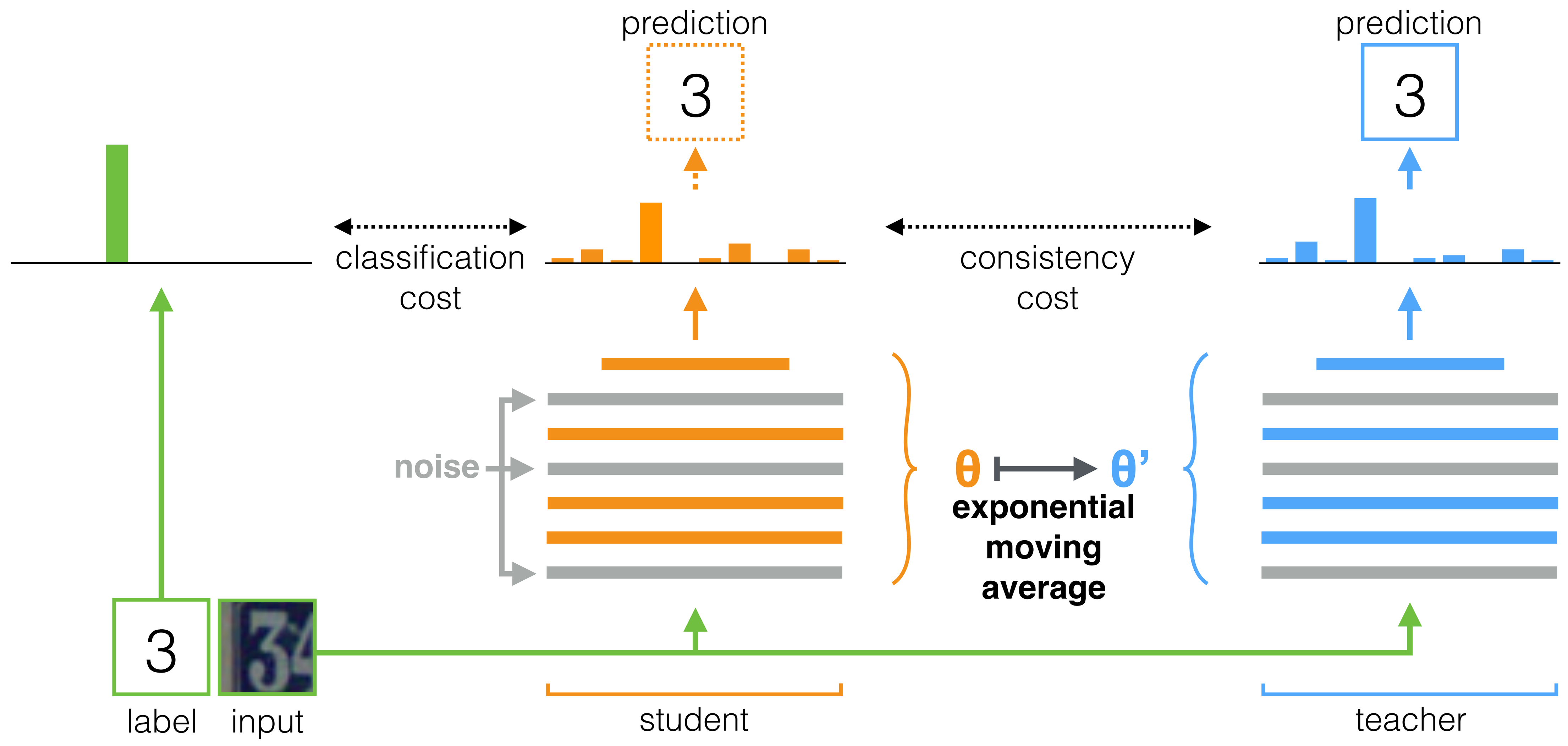
2016: **Temporal ensembling** [4] averages predictions over epochs to form less biased targets.
Problem: Cannot be applied to large datasets.



Our contribution: Mean teacher

We form a **Mean teacher** by averaging student weights over time. Averaged weights are known to improve prediction accuracy [5]. A better teacher should lead to better results. And it does!

On our experiments, mean teacher is more accurate than the previous state of the art. Unlike Temporal ensembling, it can be used with datasets of any size, including on-line learning.



Results on the SVHN dataset

	73257 images 250 labels	73257 images 500 labels	73257 images 73257 labels	573257 images 500 labels
Supervised only	42.65 ± 2.68	22.08 ± 0.73	2.81 ± 0.07	22.08 ± 0.73
GAN [6]		18.44 ± 4.8		
Π model [4]	12.94 ± 1.68[a]	6.65 ± 0.53	**2.54 ± 0.04**	3.26 ± 0.14[a]
Temporal ensembling [4]		5.12 ± 0.13	2.74 ± 0.06	
Mean teacher	**4.35 ± 0.50**	**4.18 ± 0.27**	**2.50 ± 0.05**	**2.46 ± 0.06**

[a]Our implementation