
# Curriculum Learning: An Efficient Learning Paradigm

Session 2 : Curriculum Learning Theory

Sanka Mohottala
Sri Lanka Institute of Information Technology (SLIIT)

# Curriculum Learning via Data Complexity

- Bengio et al. [1] introduced Curriculum Learning based on increasing data complexity.
- Pre-defined CL methods comes under this and many of the Self-Paced Learning (SPL) and Transfer Teacher methods also come under this.

- Definition

Given that z is a (x,y) data point from the dataset and P(z) is the training distribution from which the learner is learning a function of interest.

$$0 \leq W_\lambda(z) \leq 1 \qquad 0 \leq \lambda \leq 1 \qquad W_1(z) = 1$$

$$Q_\lambda(z) \propto W_\lambda(z)P(z) \quad \forall z$$

such that $\int Q_\lambda(z)dz = 1$. Then we have

$$Q_1(z) = P(z) \quad \forall z.$$

- The entropy of distributions gradually increases along λ
- The weight for any example increases along λ

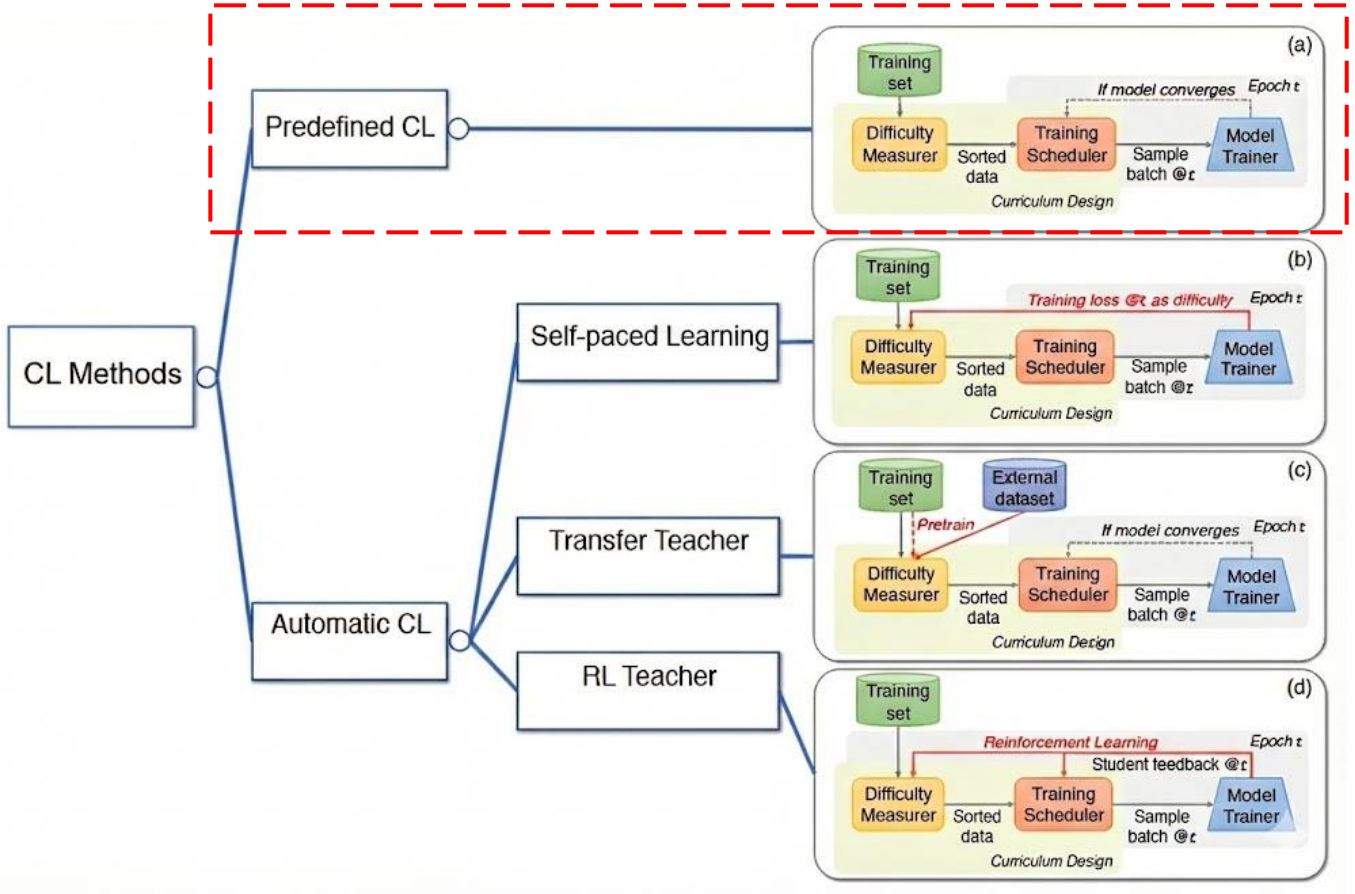



Fig 1: Categorization of CL methods based on scoring method

[1] Y. Bengio, et al. Curriculum learning. In ICML, 41–48, 2009

# What Does Curriculum Learning do?

Researchers have shown that CL improves **convergence** and **generalization** including Bengio et al. [1].

https://colab.research.google.com/drive/1j4gnqKjj2LGQ5lt-JH275xMriSvctInb?usp=sharing

# Results on Convergence and Generalization

Convex optimization (with less noisy data)

- Data was created for two classes from two random gaussian distributions with 2 input dimensions.
- Bayes classifier was used to measure difficulty.
- SVM was trained using “Easy only” vs “Hard only” approach and resulted in better generalization (16% vs 17%).

Non-Convex optimization

- Perceptron was trained from a similarly generated data from uniform distributions with high input dimensions.
- Easy data (relevant) are on one half-space and some Hard data (irrelevant) on other half-space. (or using margin)
- Perceptron trained for 500 time and averaged with results in Fig 2.

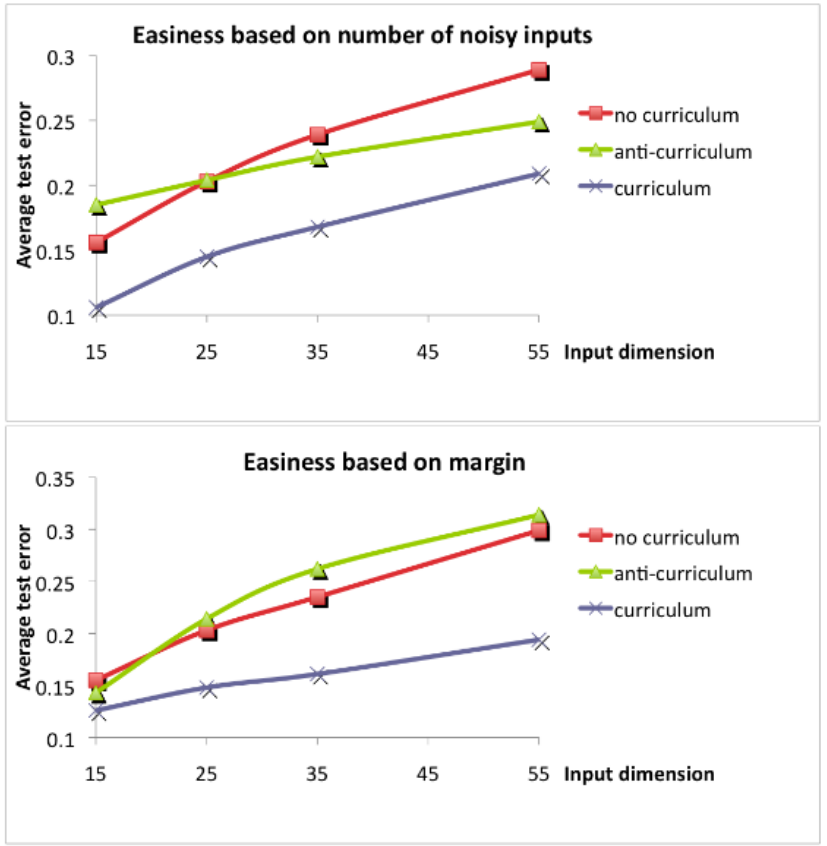



Fig 2: Perceptron based experiment results

# Why Does Curriculum Learning work?

- Existing research [1-3] study effectiveness of CL from two perspectives, as an optimization problem (Guiding) and as a data distribution problem (denoising).

Optimization perspective (Guide)

- CL can be interpreted as a continuation method.
- Takes a smoother loss landscape first and find a global picture from it first.
- Then fine-tune a better minima using less smooth loss landscape information.
- This can be interpreted as a type of transfer learning i.e., some semantic understanding is obtained from smooth loss that help to learn from complex loss as well.

Data distribution (denoise)

- Datasets contain noise (incorrect labels, ambiguity of features), CL can be seen as separating noisy from clean by working mostly with clean data.

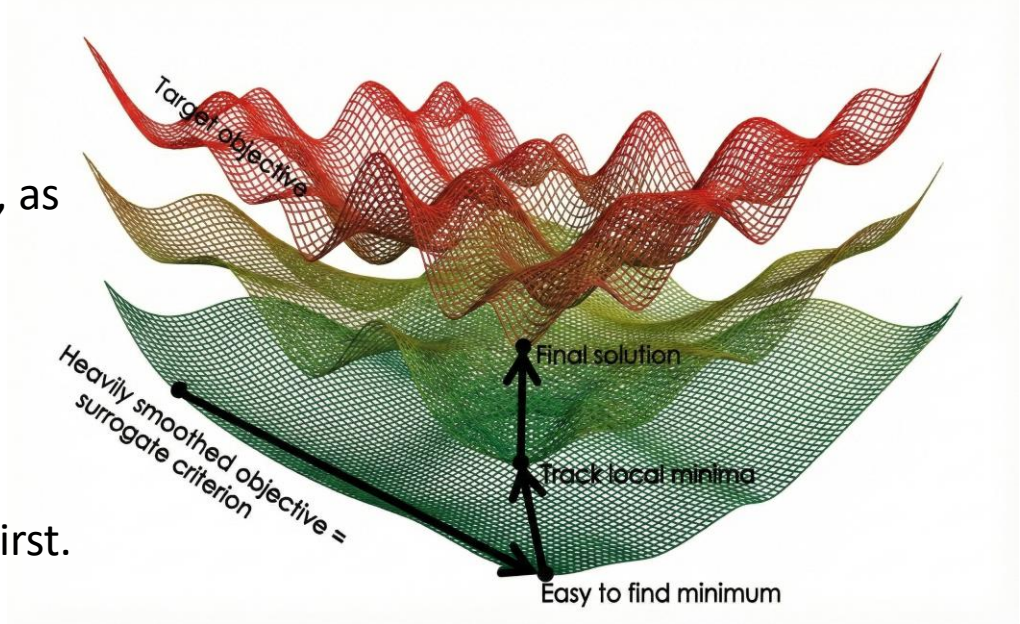



Fig 3: loss landscapes for easy to hard data subsets

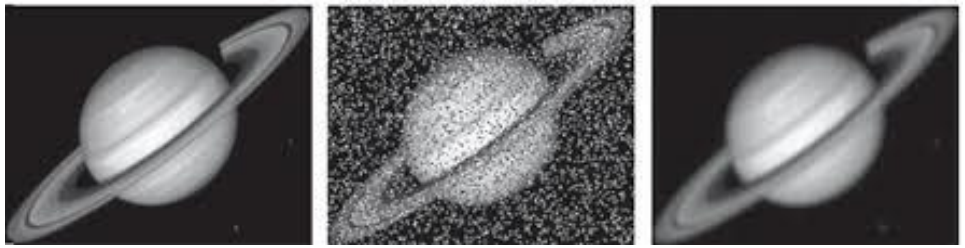
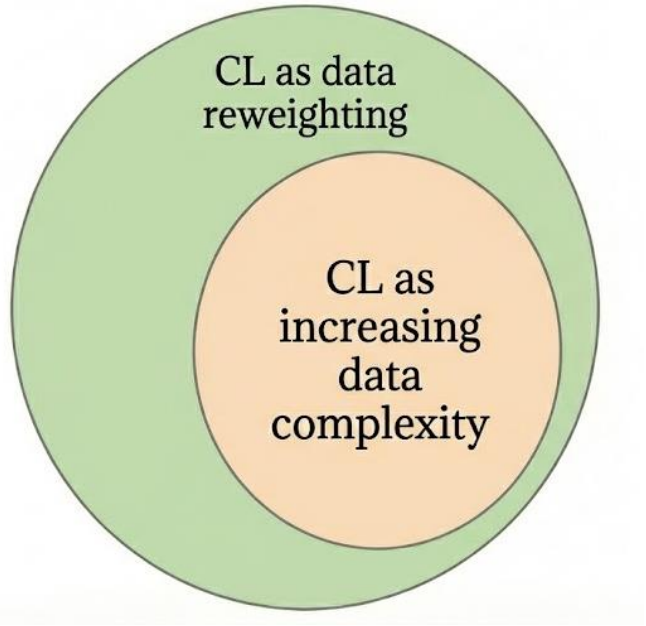


Fig 4: Noise effect on data

[2] D. Weinshall, et al. Curriculum learning by transfer learning: Theory and experiments with deep networks. In ICML, 2018.
[3] T. Gong, et al. Why curriculum learning & self-paced learning work in big/noisy data: A theoretical perspective. Big Data & Information Analytics, 2016.

# When Anti-CL works better….

- While CL works as data is given in increasing complexity, there are some research that show that works in the opposite direction – when given in reversed order (i.e., Anti-CL).
  - Hard data mining [4]
- This results in relaxing the conditions of CL definition and presenting CL as a data sampling method – without a pre-defined directionality.




**CL by Dynamic Instance Hardness (DIH) [4]**

- Instantaneous hardness generally measured at the start of each epoch with the model's feedback on all samples (i.e., loss).
- But this is highly sensitive due to constant change and [4] propose a moving average thus taking history of hardness (loss) into consideration.
- Model is trained to focus on hard data since they are the ones model is having a hard time learning.

| Curriculum | CIFAR10 | CIFAR100 | Food-101 | ImageNet |
|---|---|---|---|---|
| Rand mini-batch | 96.18 | 79.64 | 83.56 | 75.04 |
| SPL | 93.55 | 80.25 | 81.36 | 73.23 |
| MCL | 96.60 | 80.99 | 84.18 | 75.09 |
| DIHCL-Rand, Loss | 96.76 | 80.77 | 83.82 | 75.41 |
| DIHCL-Rand, dLoss | 96.73 | 80.65 | 83.82 | 75.34 |
| DIHCL-Exp, Loss | 97.03 | 82.23 | 84.65 | 75.10 |
| DIHCL-Exp, dLoss | 96.40 | 81.42 | 84.75 | 75.62 |
| DIHCL-Beta, Flip | 96.51 | 81.06 | 84.94 | 76.33 |

Fig 5: Accuracy of DIH

- DIH experiments show robustness to randomness of training thus verify the credibility of anti- CL performance.

[4] Zhou, T., Wang, S., & Bilmes, J. (2020). Curriculum learning by dynamic instance hardness. *Advances in Neural Information Processing Systems*, *33*
Codes: https://github.com/tianyizhou/DIHCL.

# Generalization vs Efficiency

- While many research show the advantages of CL (some of anti-CL), there are some work [5] that raise skepticism on CL direction.

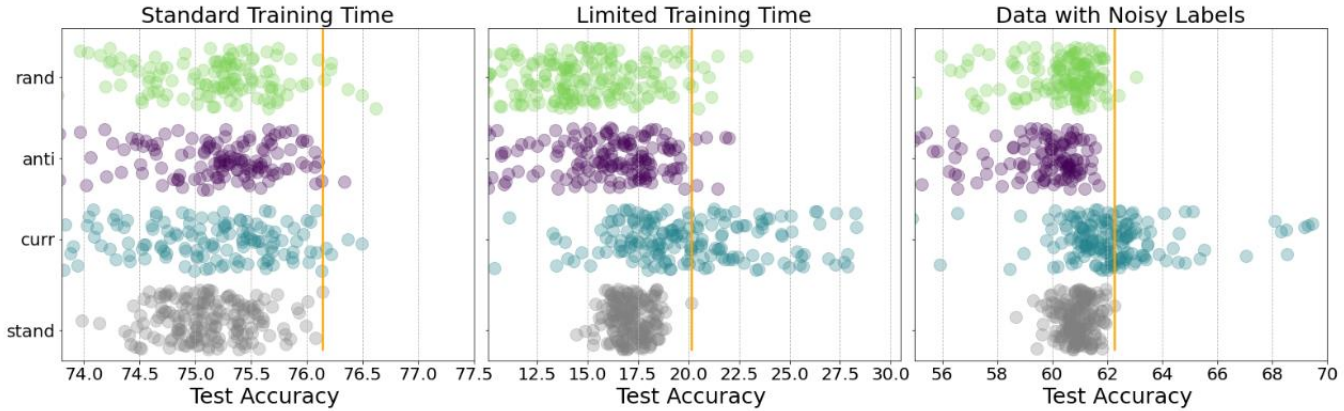



Fig 6: Accuracy vs training time vs noise robustness

- They [5] show that for some benchmark datasets, CL have only marginal benefits, and that randomly ordered samples perform similar to CL and anti-CL.
- Suggests that CL benefits are due to the dynamic training set size.
- Experiments demonstrate that CL <u>(not anti-CL)</u> can indeed improve the performance with limited training time budget or in existence of noisy data.
  - This suggest the presence of training time efficiency and data-efficiency.

*Experiments are done with only 3 transfer teacher scoring functions.*

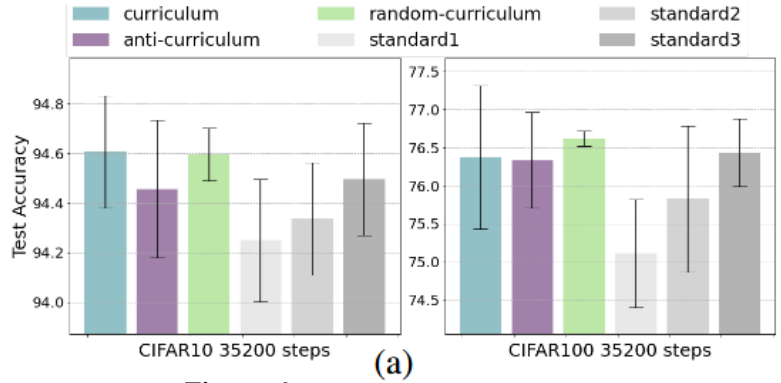



Fig 7: Average accuracy

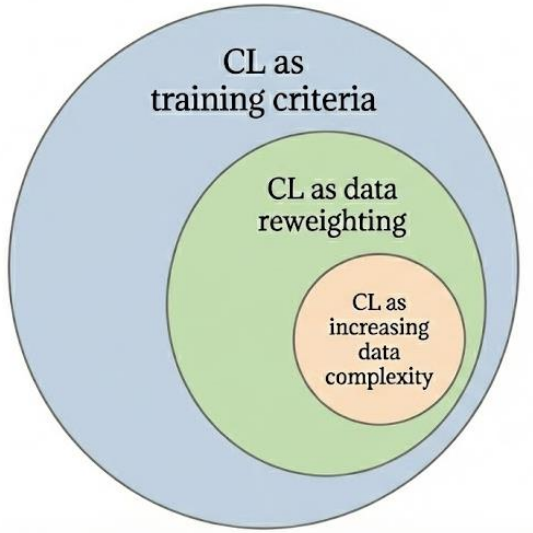
Hypothesis 1: Learning dynamics are different under CL and need to revisit the optimizer hyperparameter tuning.
Hypothesis 2: CL favors some architectures (and data modalities) over others.
Hypothesis 3: Datasets should contain sufficient complexity/difficulty variation.

[5] Wu, X., Dyer, E., & Neyshabur, B. (2020). When do curricula work? *arXiv:2012.03107*.

# Generalization of Curriculum Learning

- CL can be further generalized by taking it beyond data-level and into model-level, task-level etc.
- Under this definition, CL can bet thought of as a training criteria.
- Many research [6][7] have explored this direction.

- In Curriculum by smoothing [6], model capacity was increased with a curricular by incorporating an implicit inductive bias.
- Starts as a low pass filter and gradually include high frequency information of feature maps along epochs.
  - Making the CNN filters robust to the feature maps.
  - Can be interpreted as an augmentation method.
  - Performs over multiple domains including TFL, generative (VAE) etc.
- In Curriculum dropout [7], fixed dropout rate is replaced by epoch varying dropout rate starting with none to a fixed rate.
  - Performs across multiple sub domains in CV with CNNs.



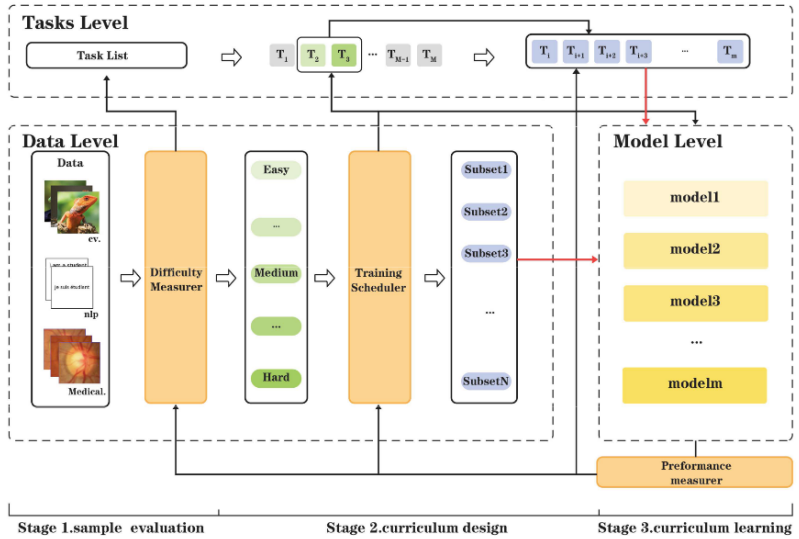




Fig 8: Generalized CL


[6] S. Sinha, et al. Curriculum by smoothing. In NeurIPS, 2020.
[7] P. Morerio, et al. Curriculum dropout. In ICCV, 3544–3552, 2017.

# Limitations of Pre-Defined CL Methods

- Selection Complexity: No standard methodology exists for pairing Difficulty Measurers with Training Schedulers; relies on exhaustive trial-and-error.
- Lack of Adaptivity: Components remain fixed during training, ignoring dynamic feedback from the model.
- Design Bottlenecks: Requires heavy domain expertise and struggles with high-dimensional feature spaces.
- Human-Model Misalignment: Examples considered "easy" by humans are often difficult for models due to differing decision boundaries.
- Hyperparameter Sensitivity: Identifying optimal scheduler hyperparameters is computationally difficult.

How can we design automated scoring and pacing functions that are independent of human heuristics, adapt dynamically to the training trajectory, and remain robust against hyperparameter variations?

# Transfer Teacher (TT)

- Instead of using a human guided score function, a pre-trained model is sued to give the score.
- Pacing function is generally an adaptive one that take the models feedback into consideration.

- In [8], ImageNet-1k pretrained model with an SVM classifier was used to get the score values for CIFAR-10.
- Accuracy improvement and reduced wall clock time.
- In [9], along with transfer learning approach[8], bootstrapping (using a pre-trained model with same architecture and with same dataset) used and obtained similar results.
- In [10], scoring function was improved by taking a cross-validation approach with N pre-trained model.

- Extra pre-training add more wall clock time.

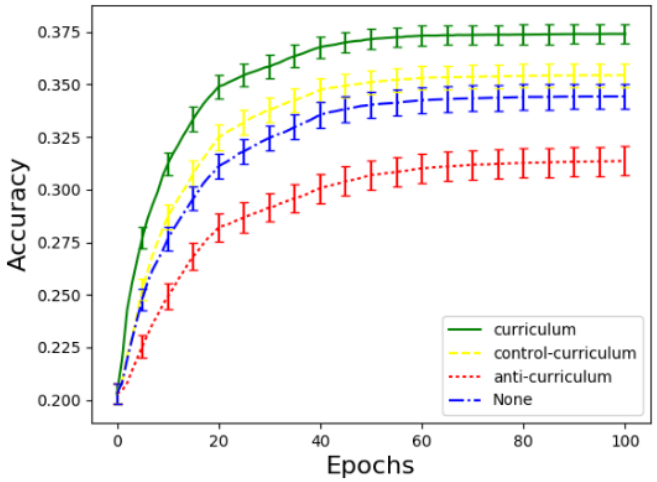



Fig 9: CIFAR-10 with CNN [8]

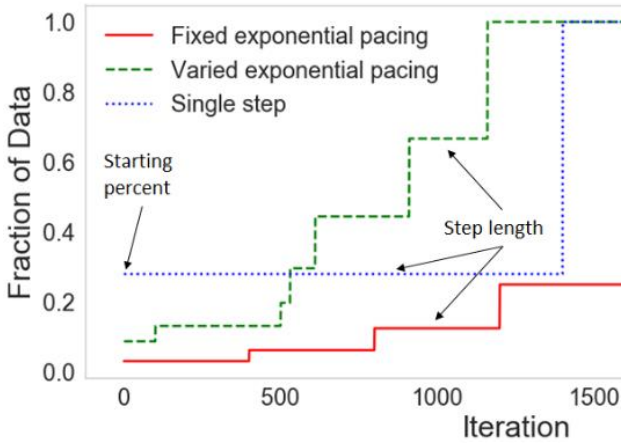
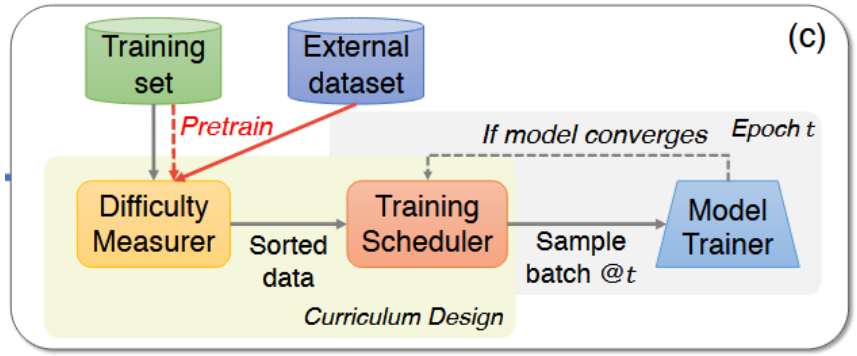



Fig 10: adaptive pacing [9]

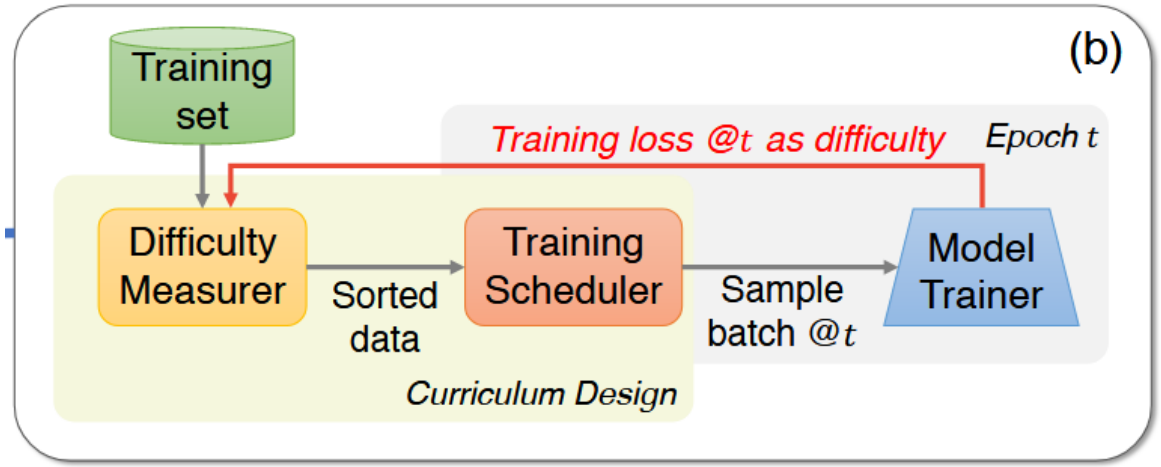




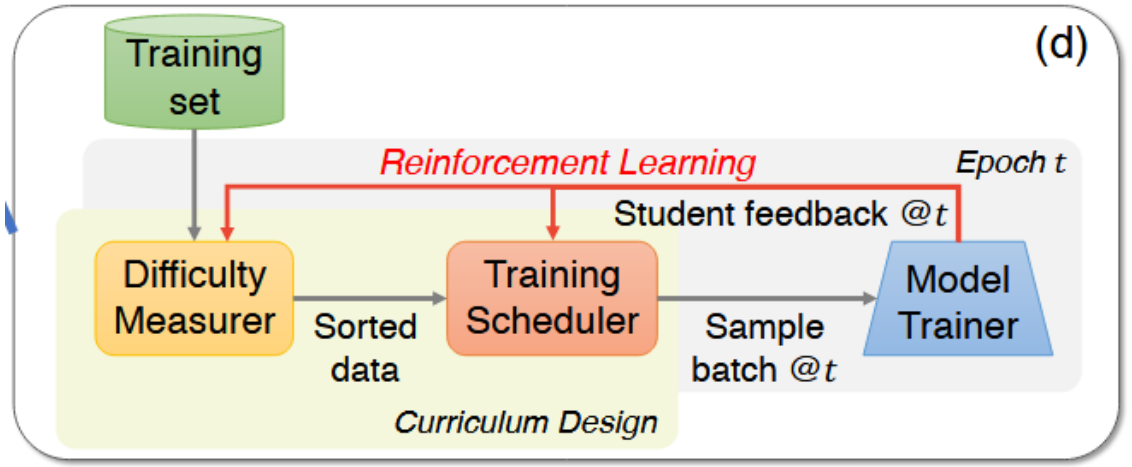
[8] D. Weinshall, et al. Curriculum learning by transfer learning: Theory and experiments with deep networks. In ICML, 2018.
[9] G. Hacohen, et al. On the power of curriculum learning in training deep networks. ICML, 2019.
[10] Xu, Benfeng, et al. "Curriculum learning for natural language understanding." *Proceedings of the 58th annual meeting of the association for computational linguistics*. 2020.

# Other CL Approaches

- self-paced learning (SPL), - TT utilize a teacher model, but a student model can also obtain a score value for data from itself during the training.
- In Bengio et al.[1] pre-defined CL, synthetic data usage was a limitation, and SPL was introduced to mitigate this in [11].




- The SPL and TT only automate the scoring function and still use predefined pacing function, and they only consider one side of the CL.
  - SPL takes the student feedback (i.e., losses) to adjust the curriculum
  - TT leverages the teacher's knowledge to determine the order of presenting data.
- Combining both, RL teacher [12] approach take data selection as ACTION and student feedback as REWARD and STATE.




[11] Kumar, M., Benjamin Packer, and Daphne Koller. "Self-paced learning for latent variable models." *Advances in neural information processing systems* 23 (2010).

[12] Graves, Alex, et al. "Automated curriculum learning for neural networks." *international conference on machine learning*. Pmlr, 2017.