SEARCHING FOR UNIVERSALITY IN DEEP LEARNING

Marco Gherardi
10/5/2024

software engineering

(Margaret Hamilton)

software engineering

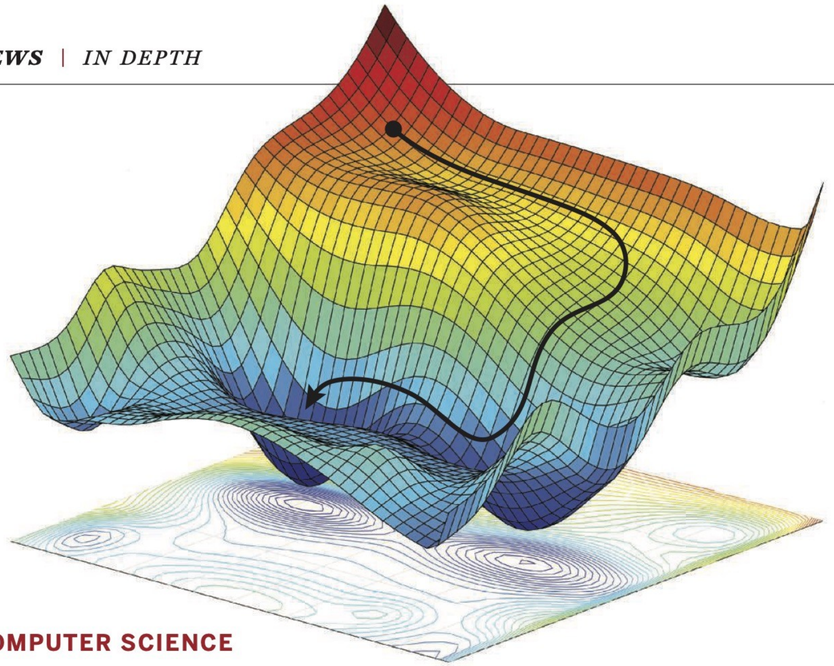
(Margaret Hamilton)

AI engineering?

(Ali Rahimi)

NO!

NEWS | ***IN DEPTH***



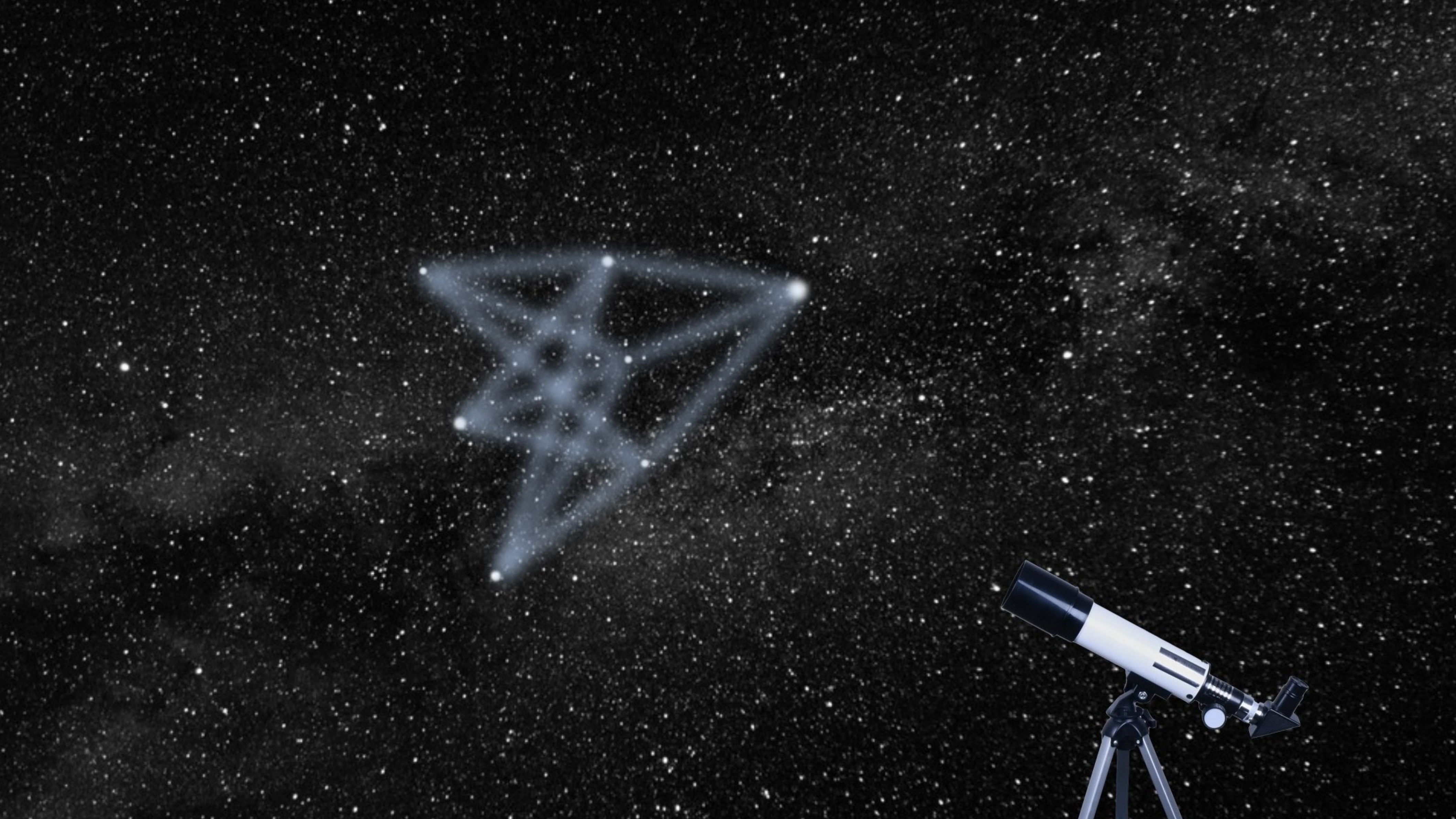
COMPUTER SCIENCE

Has artificial intelligence become alchemy?

SCIENCE 360, 6388 (2018)

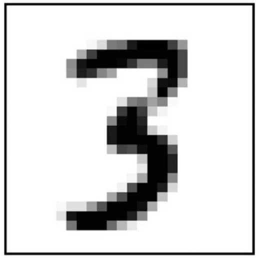
“Many of us feel like we’re operating on an alien technology

“[Problems happen] because we apply brutal optimization techniques to loss surfaces that we don’t understand



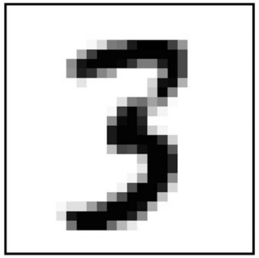
WHAT IS DEEP LEARNING?


some clever algorithm
“three”



$f_{\boldsymbol{\theta}}(\boldsymbol{x})$ with a lot of parameters

"three"



⇨ $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ with a lot of parameters ⇨ “three”

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \hat{\sigma} \circ A^{(L)} \circ \cdots \circ \sigma \circ A^{(2)} \circ \sigma \circ A^{(1)}(\boldsymbol{x})$$

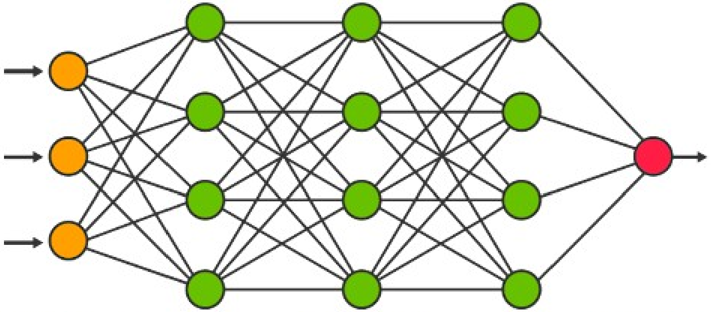
neural network

layers

affine transformation

$$A^{(l)}(\boldsymbol{x}) = W^{(l)}\boldsymbol{x} + \boldsymbol{b}^{(l)}$$

how to fix the parameters?



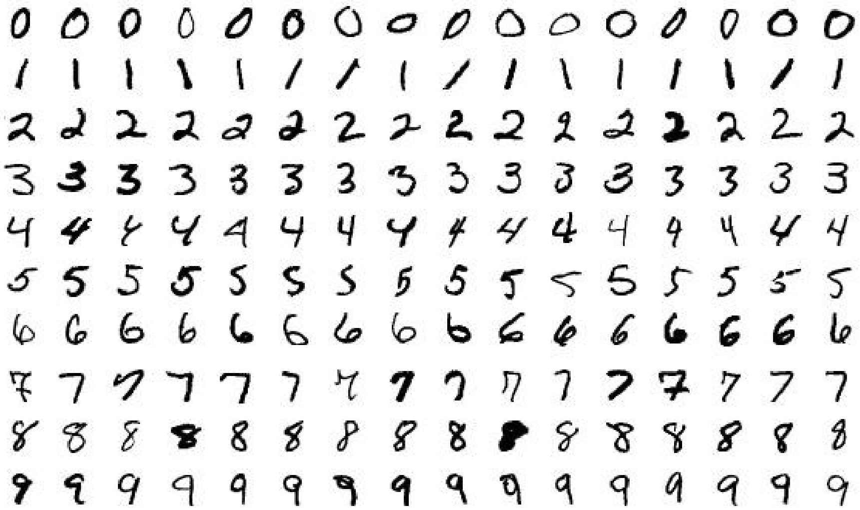
PARAMETERS ARE FIXED BY TRAINING (FITTING)

training set

$$\mathcal{D} = \{(\boldsymbol{x}^\mu, y^\mu)\}_\mu$$

loss function

$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x}, y) \in \mathcal{D}} \text{dist}\,(y, f_{\boldsymbol{\theta}}(\boldsymbol{x}))$$



MNIST

$$\boldsymbol{\theta}_{\text{opt}} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$$

brutal optimization

THE "LOSS LANDSCAPE"

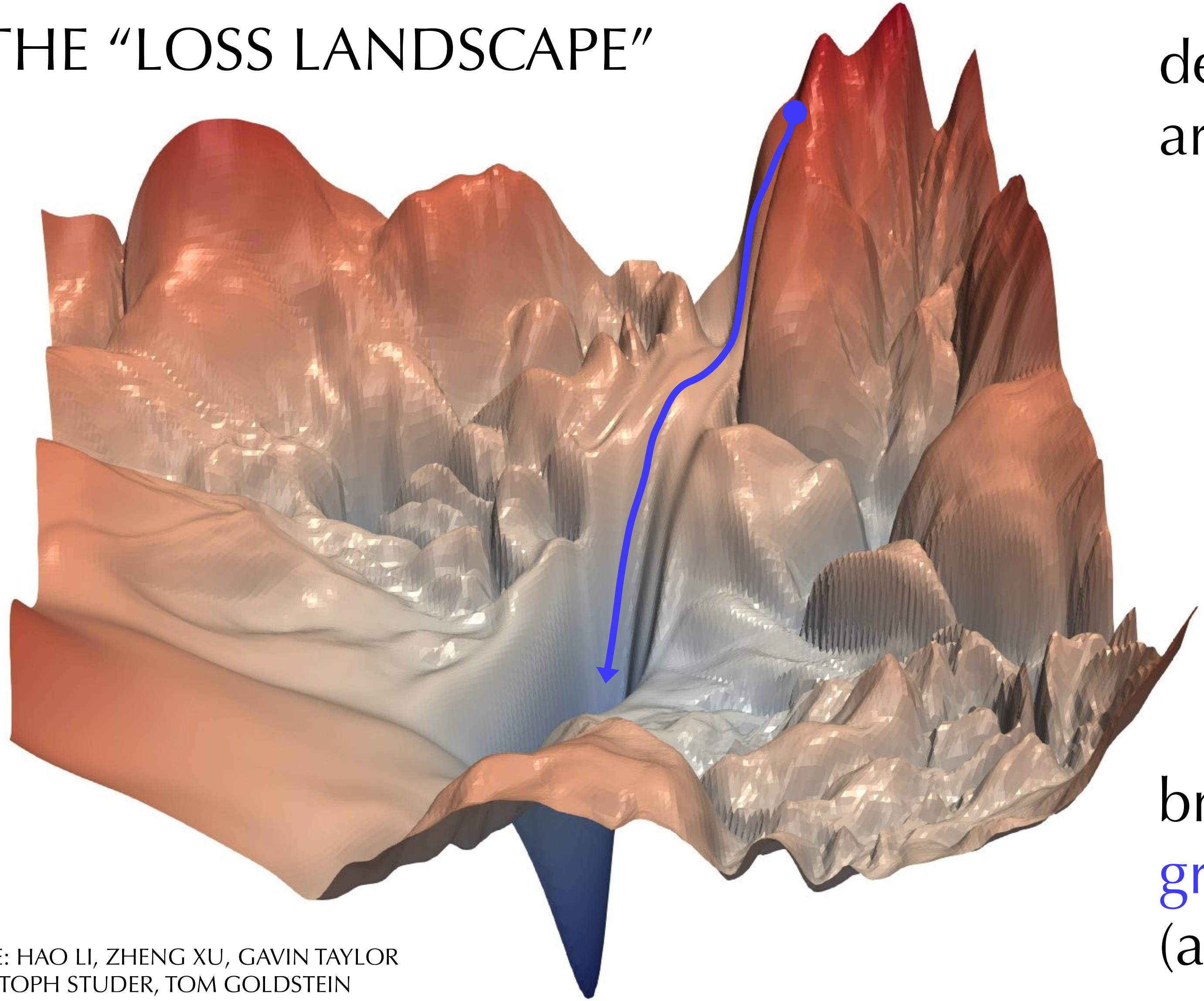
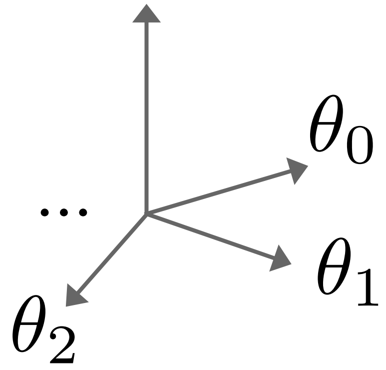


IMAGE: HAO LI, ZHENG XU, GAVIN TAYLOR
CHRISTOPH STUDER, TOM GOLDSTEIN

determined by architecture and data

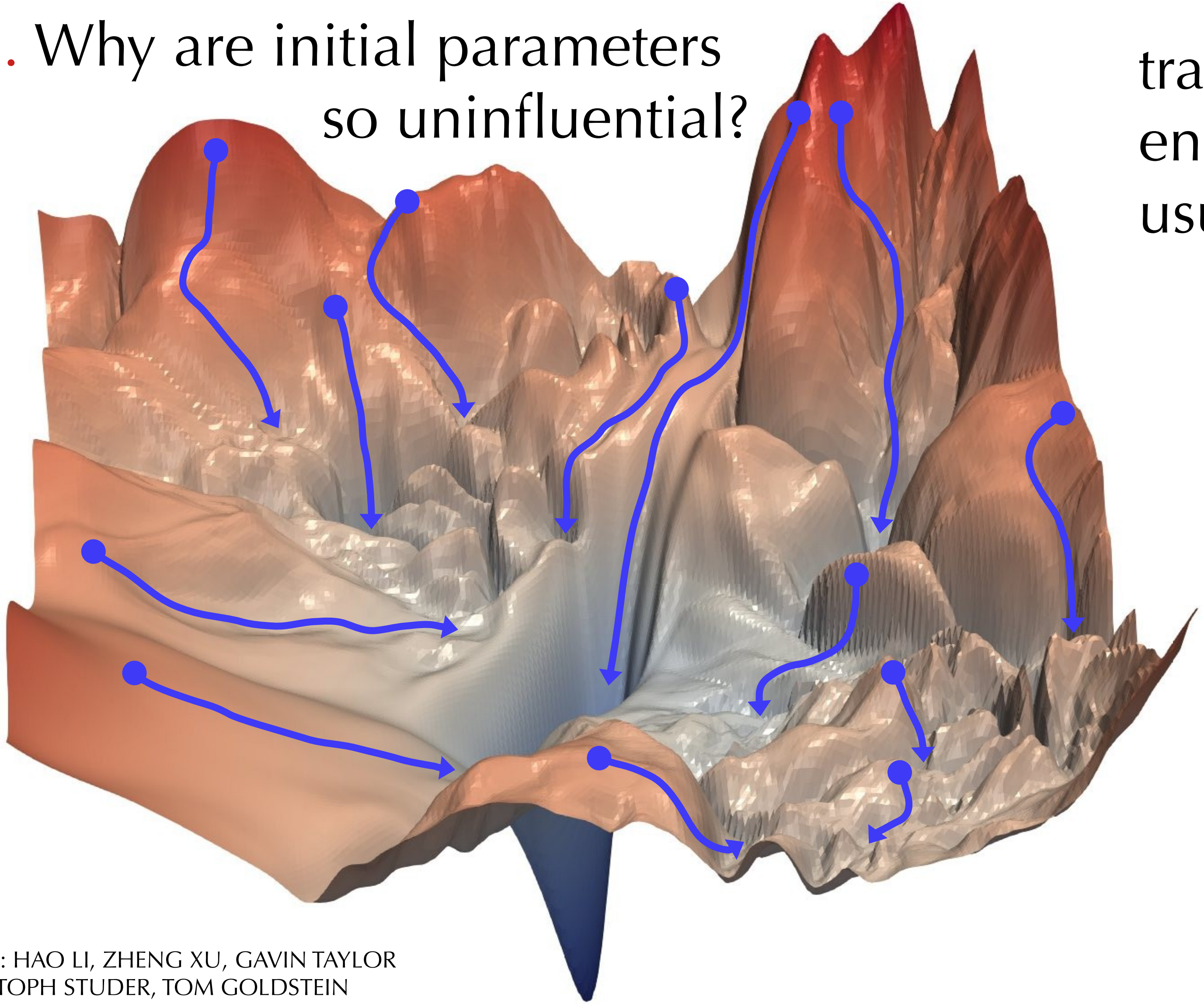
$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$$



brutal optimization:
gradient descent
(and variants)

TWO QUESTIONS

1. Why are initial parameters so uninfluential?



training dynamics ends in different, usually good, points

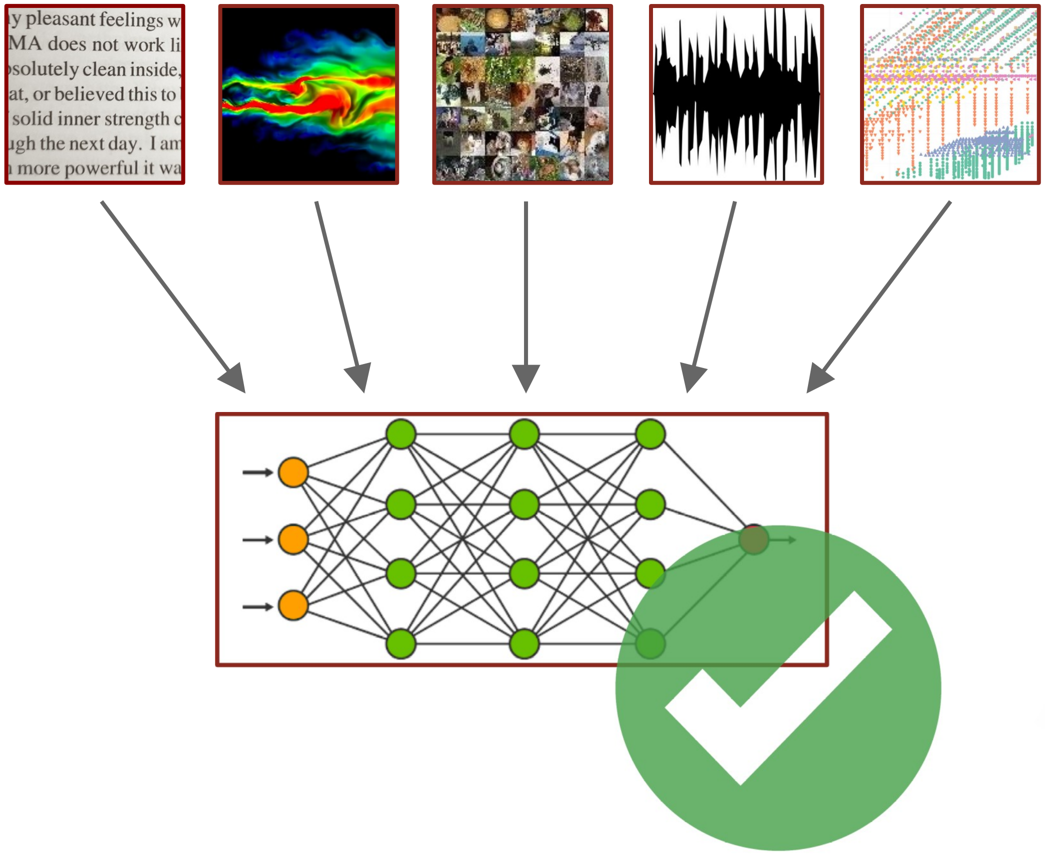
redundancy in the representation

⇓

symmetry?
robustness?

IMAGE: HAO LI, ZHENG XU, GAVIN TAYLOR
CHRISTOPH STUDER, TOM GOLDSTEIN

2. Why do the same architectures work for different data?



SOME TYPE OF UNIVERSALITY IN THE LANDSCAPE WHICH WE ARE MISSING

PHYSICS OF DEEP LEARNING

WHY PHYSICS ?

NATURE PHYSICS | VOL 16 | JUNE 2020 | 602-604 | www.nature.com/naturephysics

comment

Check for updates

Understanding deep learning is also a job for physicists

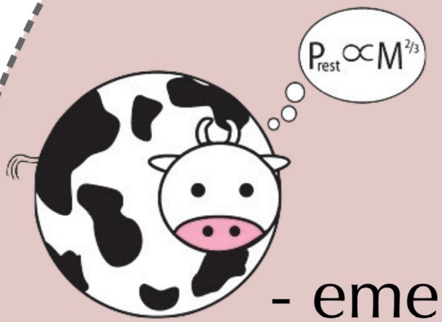
Automated learning from data by means of deep neural networks is finding use in an ever-increasing number of applications, yet key theoretical questions about how it works remain unanswered. A physics-based approach may help to bridge this gap.

Lenka Zdeborová

1. useful toolset (statistical mechanics, mean field, models)

2. aesthetics (unification, simple laws, unexpected phenomena)

3. statistical physics is the science of universality and relevance



MODELING

- emergent phenomena
- effective theories
- simple models

THEORETICAL

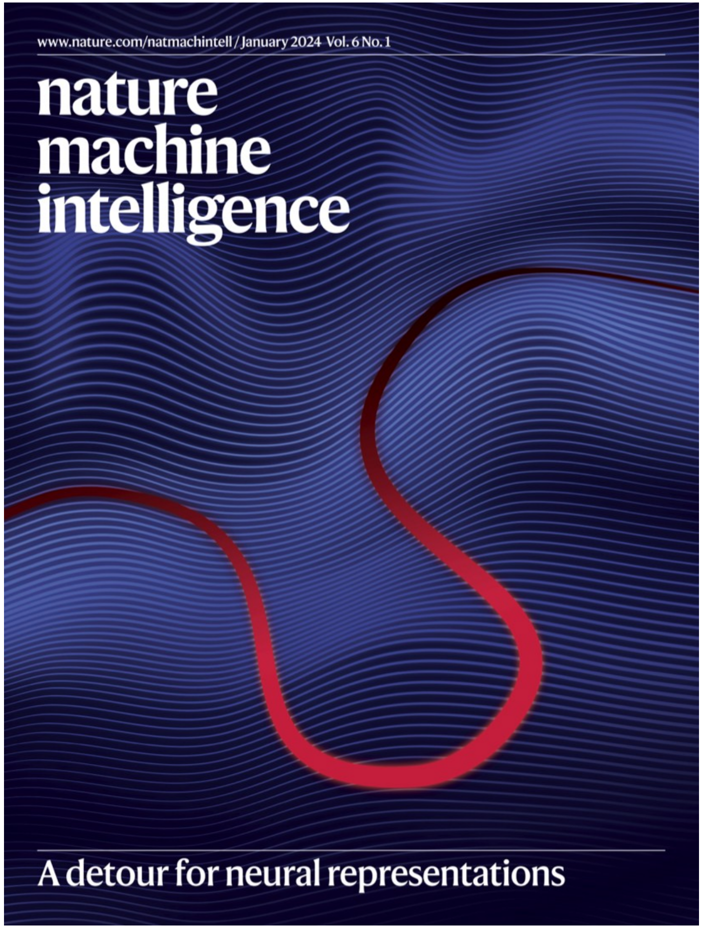
$$Z_{\mathcal{D}}(\beta) = \int D\boldsymbol{\theta}\; e^{-\beta \mathcal{L}(\boldsymbol{\theta};\mathcal{D})}$$

- analytic computations
- solvable limits

EXPERIMENTAL

look for interesting phenomena (*in silico*)

UNIVERSAL TRAINING DYNAMICS



CICERI, CASSANI, OSELLA, ROTONDO, VALLE, GHERARDI
NAT. MACH. INTELL. 6, 40 (2024)

joint work with

PIETRO ROTONDO (UNIPR)

MATTEO OSELLA (UNITO)

FILIPPO VALLE (UNITO)

SIMONE CICERI

LORENZO CASSANI

FROM PARAMETERS TO INTERNAL REPRESENTATIONS

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \hat{\sigma} \circ A^{(L)} \circ \cdots \circ \sigma \circ A^{(2)} \circ \sigma \circ A^{(1)}(\boldsymbol{x})$$

$\boldsymbol{\theta}_t$

$h_t(\boldsymbol{x}) \in \mathbb{R}^H$

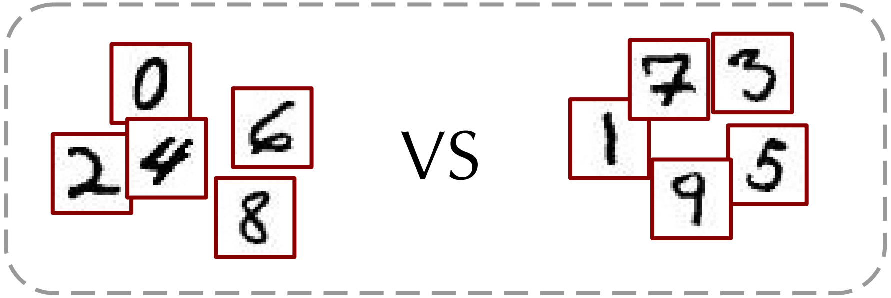
FROM PARAMETERS TO INTERNAL REPRESENTATIONS

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \hat{\sigma} \circ A^{(L)} \circ \cdots \circ \sigma \circ A^{(2)} \circ \sigma \circ A^{(1)}(\boldsymbol{x})$$

$\boldsymbol{\theta}_t$ $\qquad$ $h_t(\boldsymbol{x}) \in \mathbb{R}^H$

binary classification task

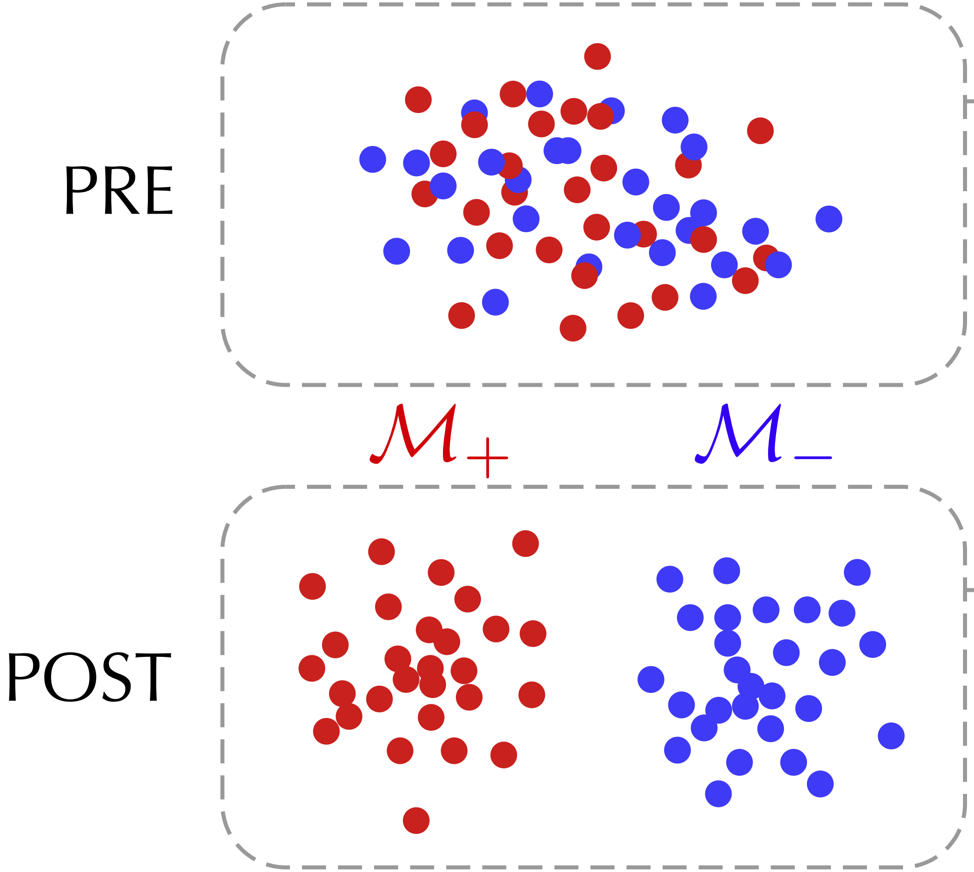
$y^\mu = \pm 1$

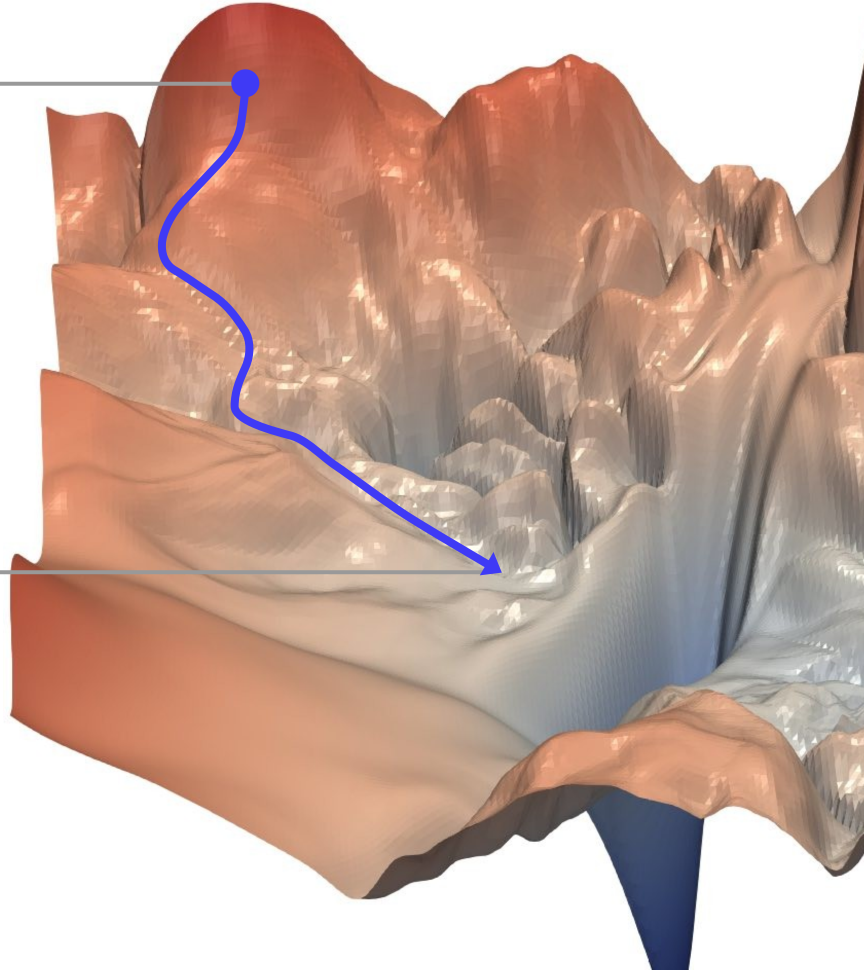


$$\mathcal{D} = \{(\boldsymbol{x}^\mu, y^\mu)\}_\mu \implies \mathcal{M}_\pm(t) = \{h_t(\boldsymbol{x}^\mu) \mid y^\mu = \pm 1\}_\mu$$

internal representations of the two classes: $\mathcal{M}_+$ $\mathcal{M}_-$

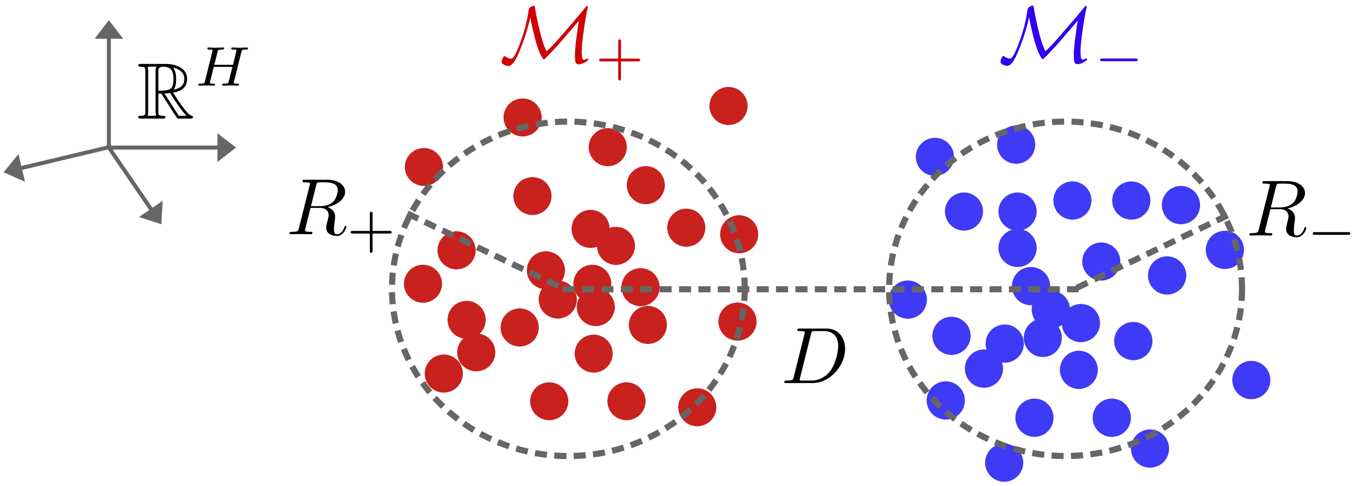
TRAINING DISENTAGLES INTERNAL REPRESENTATIONS





ROTONDO, LAGOMARSINO, GHERARDI, PRR (2020)
CHUNG, LEE, SOMPOLINSKY, PRX (2018)

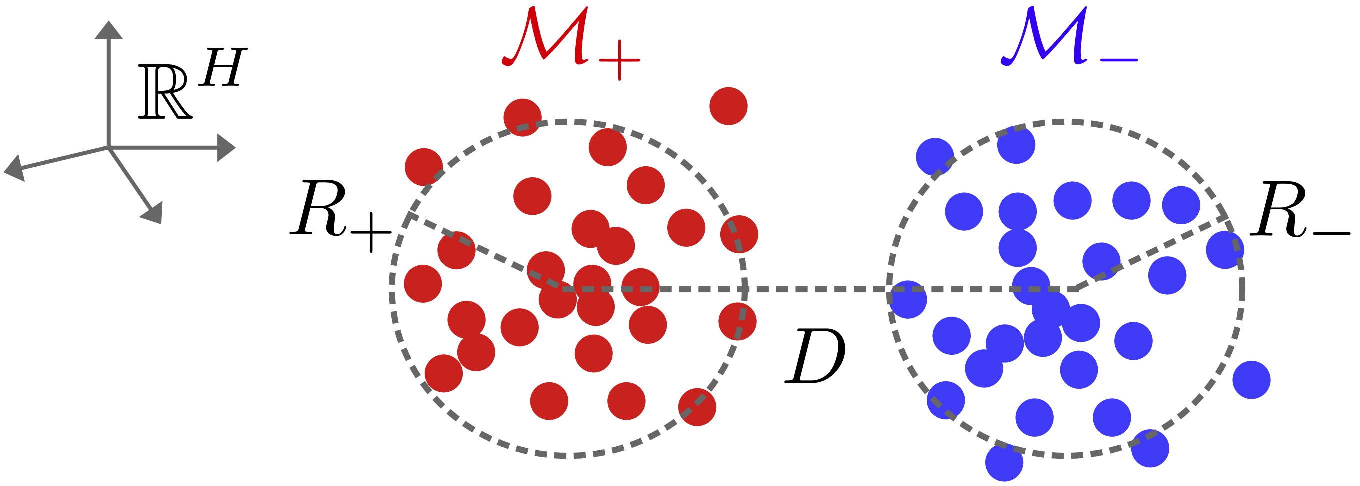
SIMPLE MEASURES OF ENTANGLEMENT

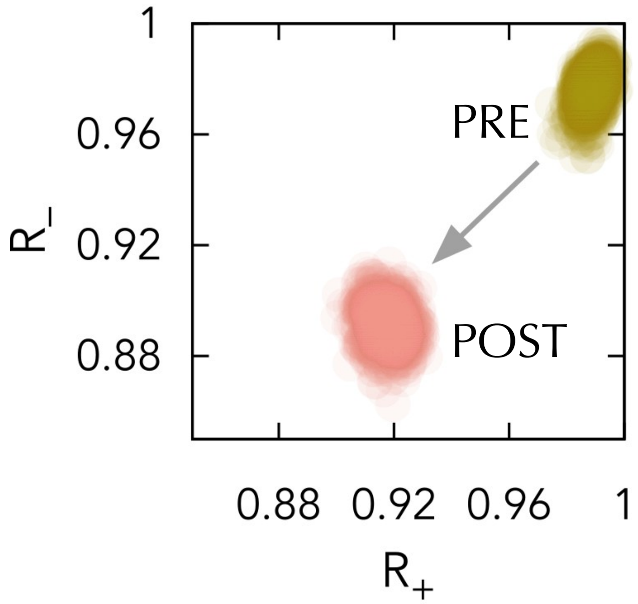


$$R^2_\pm(t) = \frac{1}{2n^2_\pm} \sum_{\boldsymbol{a},\boldsymbol{b} \in \mathcal{M}_\pm(t)} \|\boldsymbol{a} - \boldsymbol{b}\|^2$$

$$D(t) = \left\| \frac{1}{n_+} \sum_{\boldsymbol{a} \in \mathcal{M}_+(t)} \boldsymbol{a} - \frac{1}{n_-} \sum_{\boldsymbol{a} \in \mathcal{M}_-(t)} \boldsymbol{a} \right\|$$

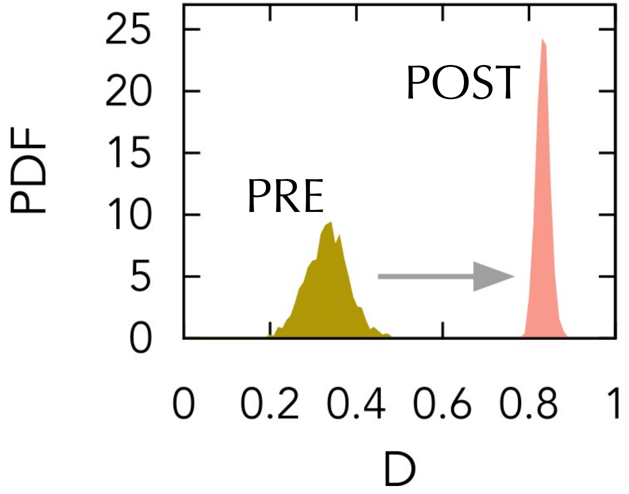
SIMPLE MEASURES OF ENTANGLEMENT

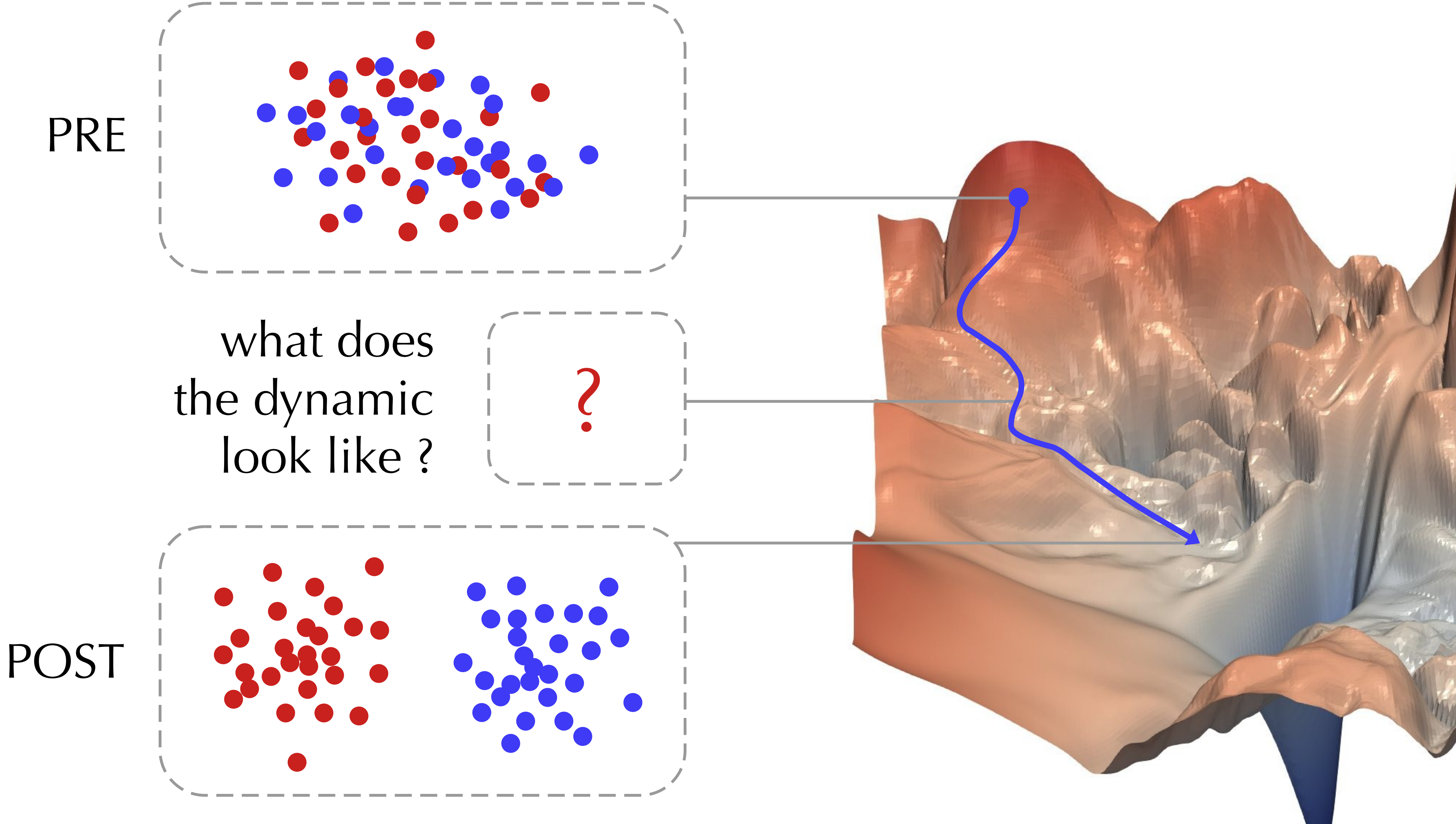




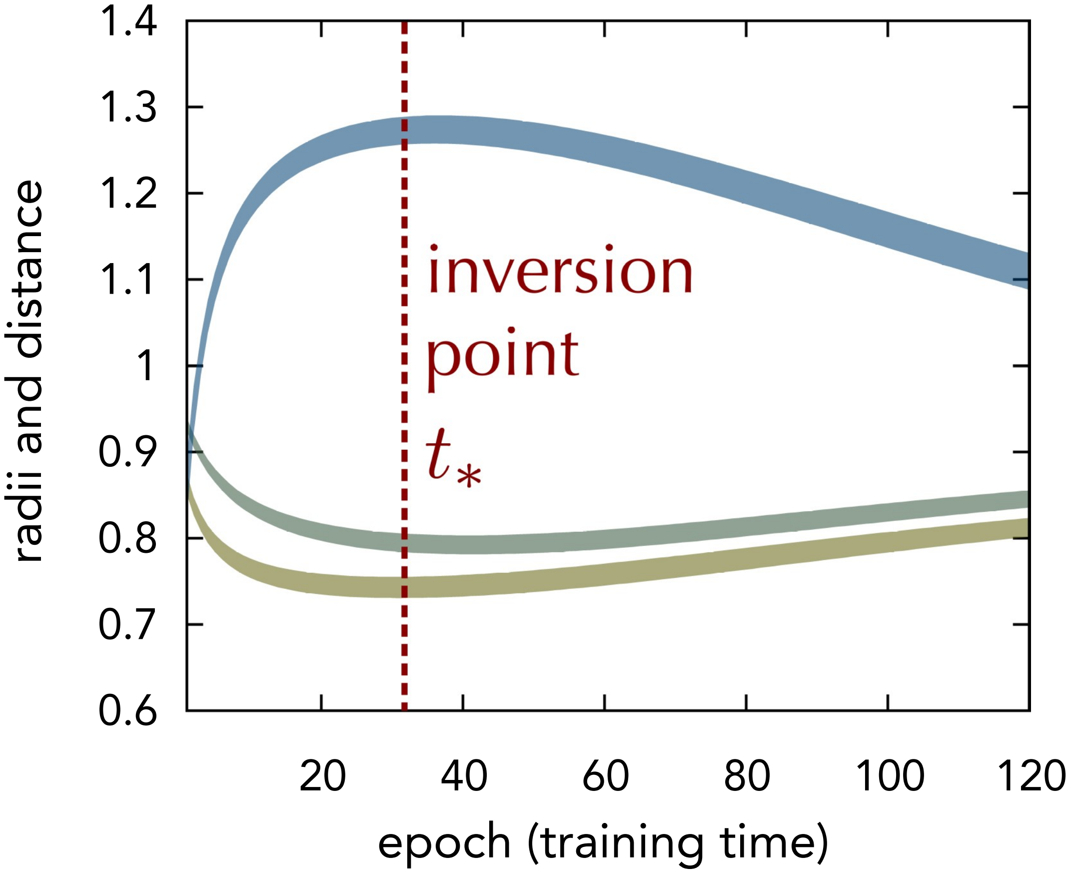
$$R^2_\pm(t) = \frac{1}{2n^2_\pm} \sum_{\boldsymbol{a},\boldsymbol{b}\in\mathcal{M}_\pm(t)} \|\boldsymbol{a}-\boldsymbol{b}\|^2$$

$$D(t) = \left\| \frac{1}{n_+} \sum_{\boldsymbol{a}\in\mathcal{M}_+(t)} \boldsymbol{a} - \frac{1}{n_-} \sum_{\boldsymbol{a}\in\mathcal{M}_-(t)} \boldsymbol{a} \right\|$$




PRE
what does
the dynamic
look like ?
?
POST

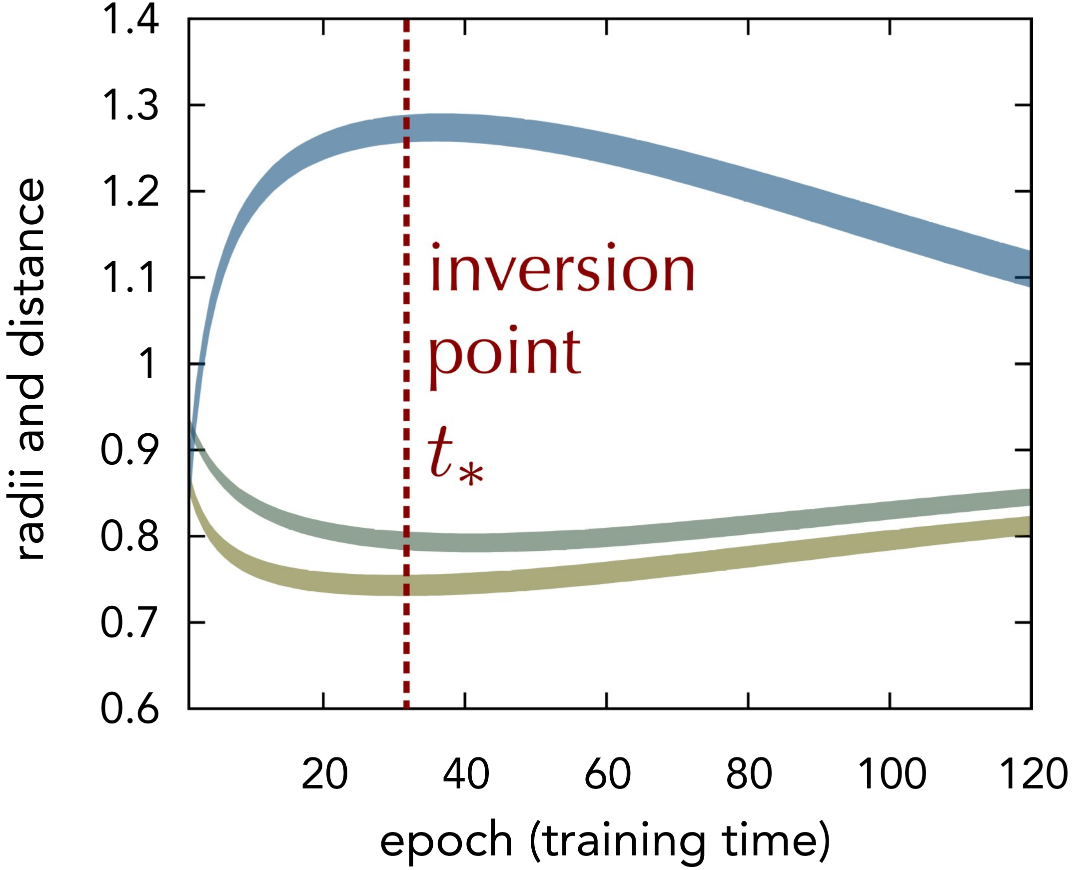
DYNAMIC IS NON-MONOTONIC



is the inversion point universal?

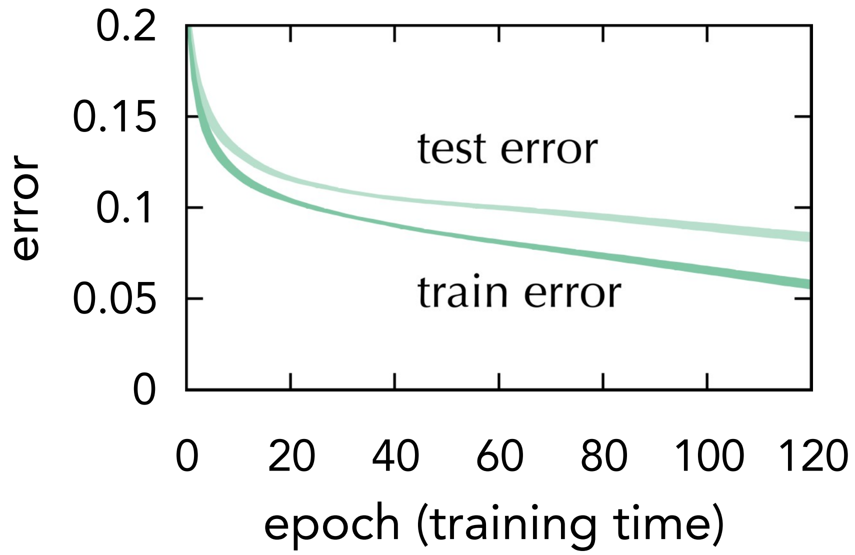
is it a property of the specific dynamics or of the loss landscape?

DYNAMIC IS NON-MONOTONIC

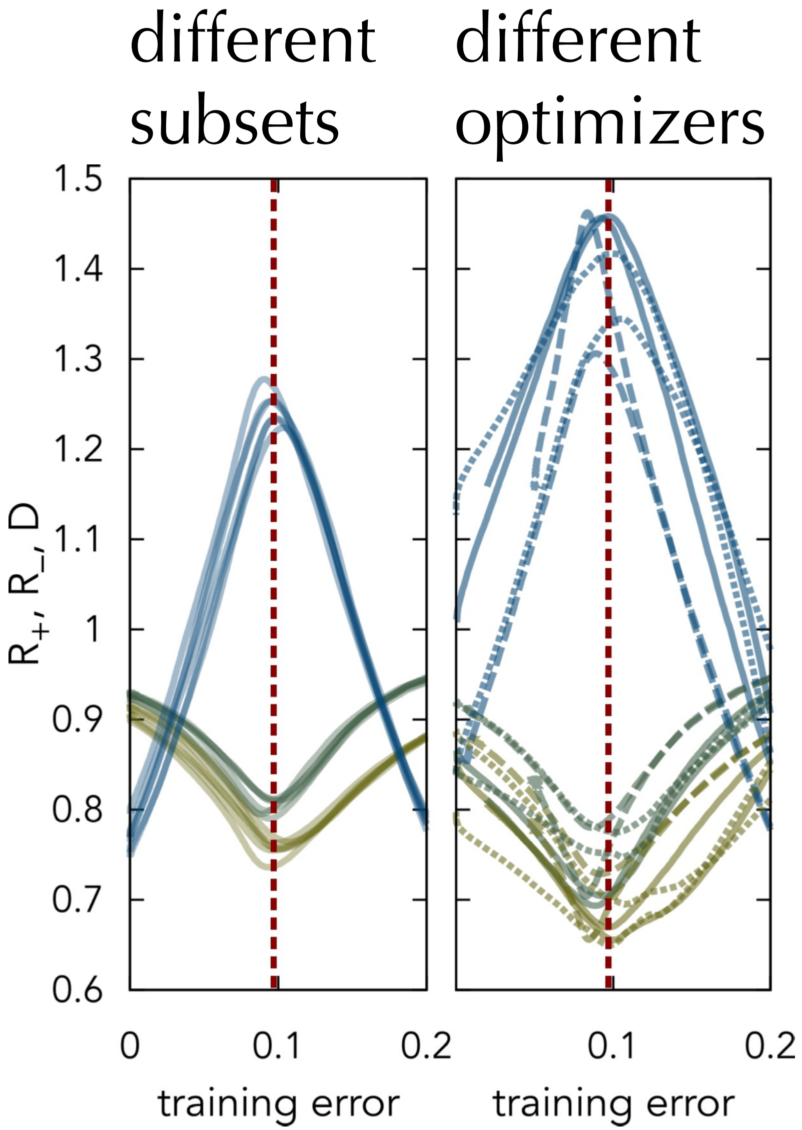


an "invariant" measure of training progress: the training error

$$\epsilon_{\rm tr} = 1 - \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}} \delta_{f_{\boldsymbol{\theta}_t}(\boldsymbol{x}),y}$$

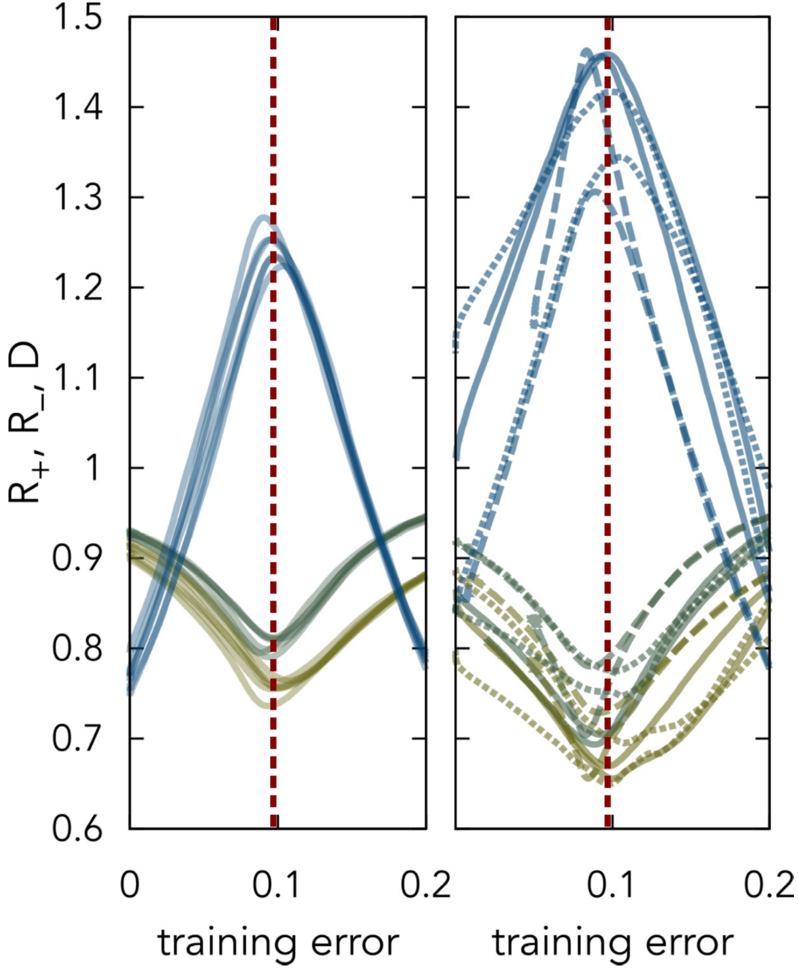


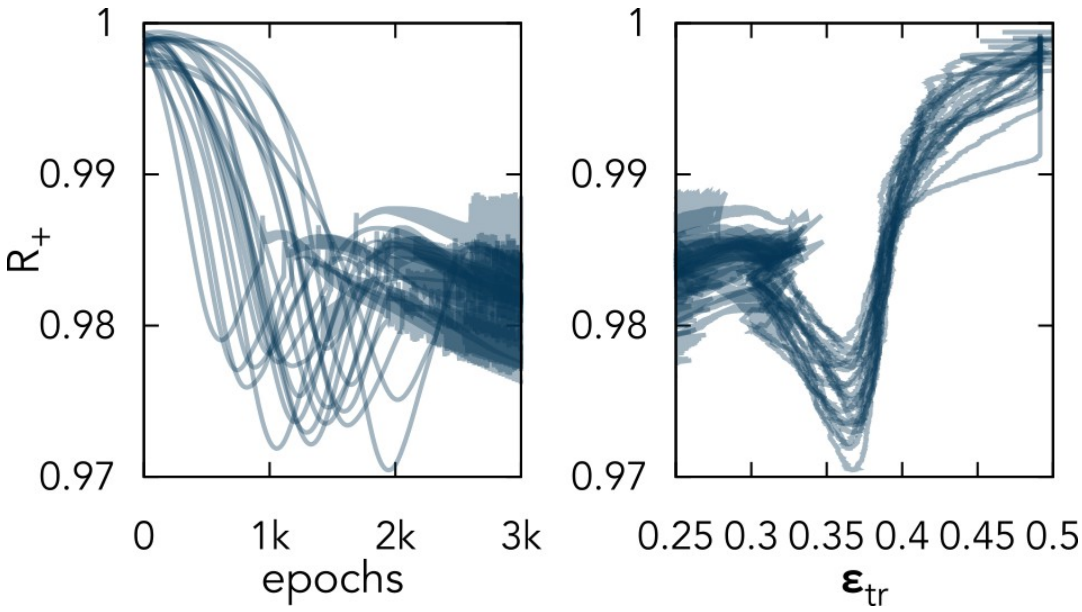
UNIVERSALITY



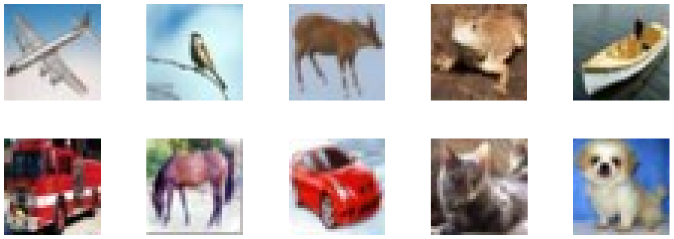
UNIVERSALITY

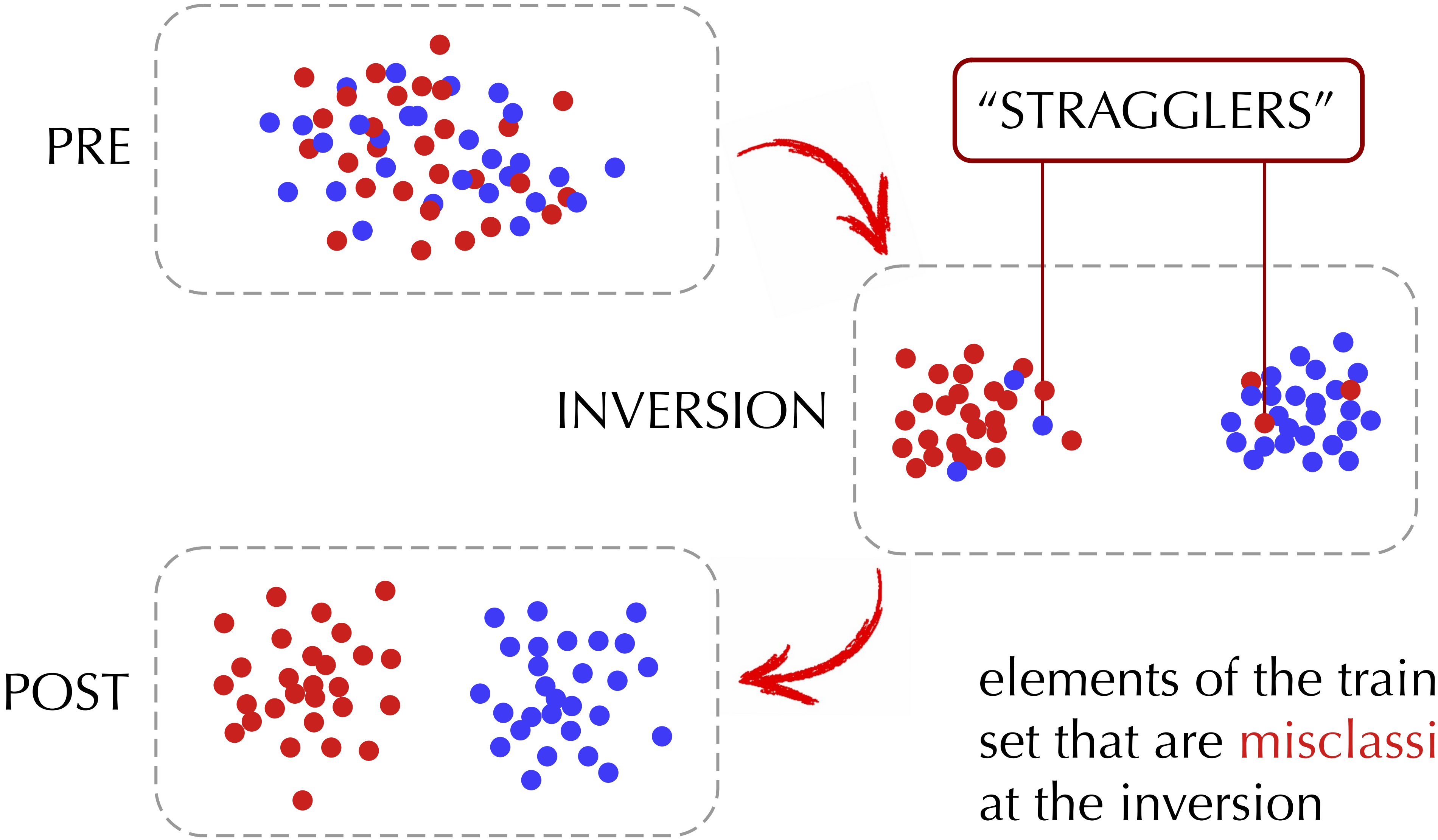
different subsets | different optimizers





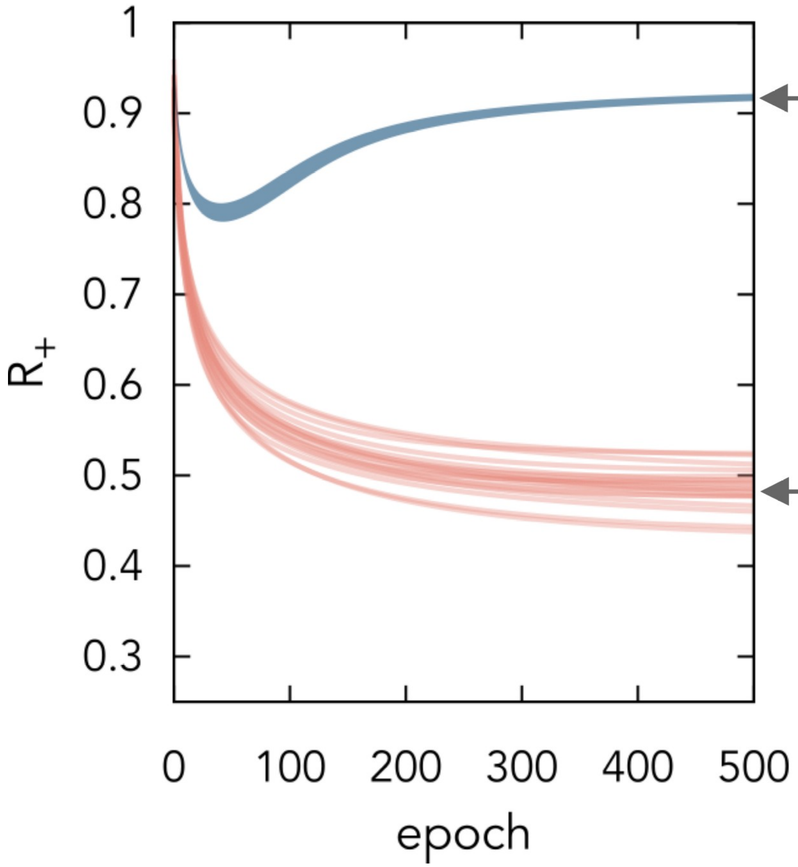
CIFAR-10
dataset





elements of the training set that are misclassified at the inversion

STRAGGLERS CAUSE THE INVERSION



full training set

$\mathcal{D}$

pruned training set

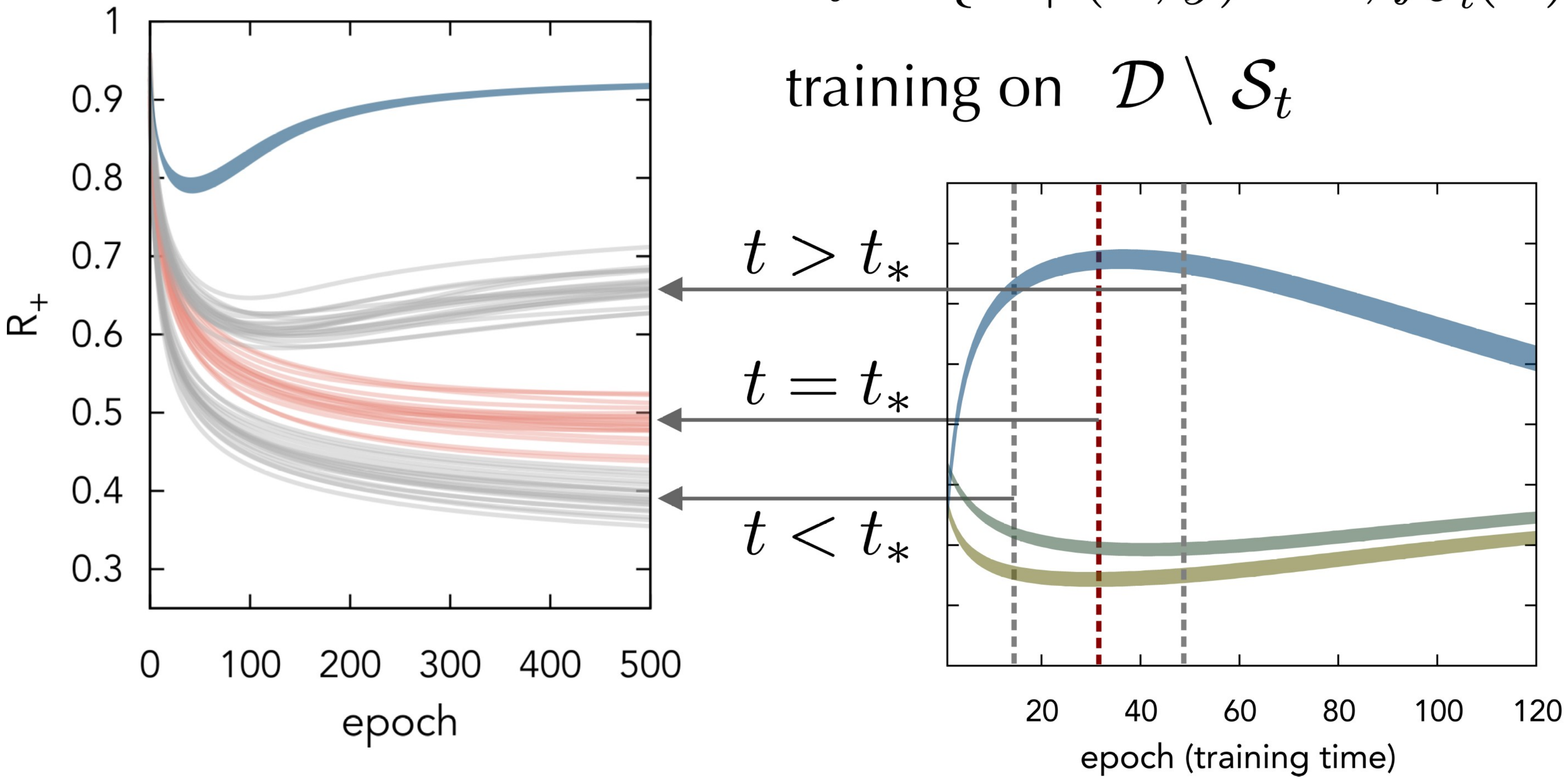
$\mathcal{D} \setminus \mathcal{S}_{t_*}$

the stragglers

STRAGGLERS CAUSE THE INVERSION

$$\mathcal{S}_t = \{\boldsymbol{x} \mid (\boldsymbol{x}, y) \in \mathcal{D}, f_{\boldsymbol{\theta}_t}(\boldsymbol{x}) \neq y\}$$

training on $\mathcal{D} \setminus \mathcal{S}_t$



STRAGGLERS ARE EXCEPTIONALLY STABLE

1. find misclassified training-set elements from two random initializations

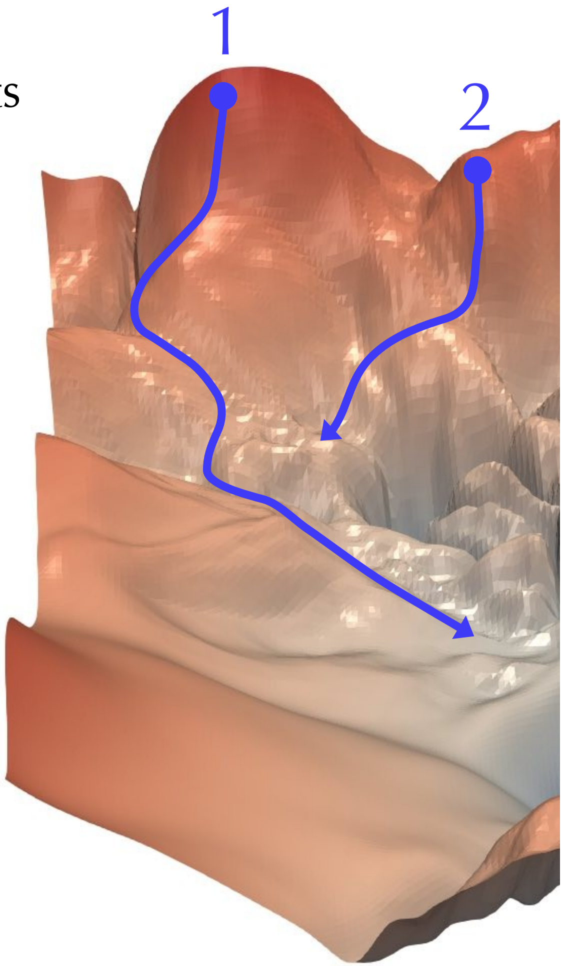
$$\mathcal{S}_t^1 \quad \mathcal{S}_t^2$$

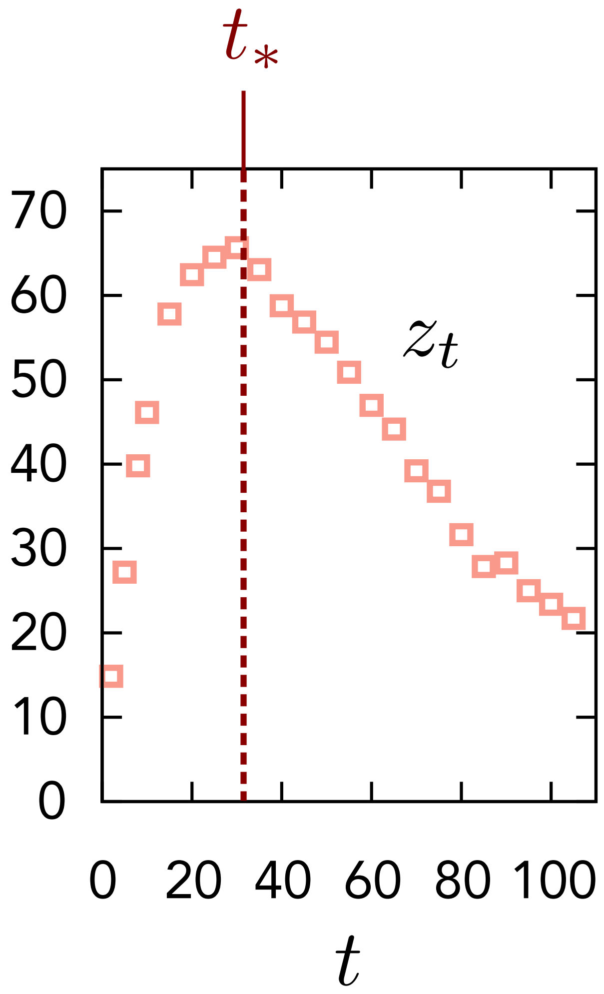
2. count common training-set elements

$$M_t = \left| \mathcal{S}_t^1 \cap \mathcal{S}_t^2 \right|$$

3. repeat and compare with null model (hypergeometric)

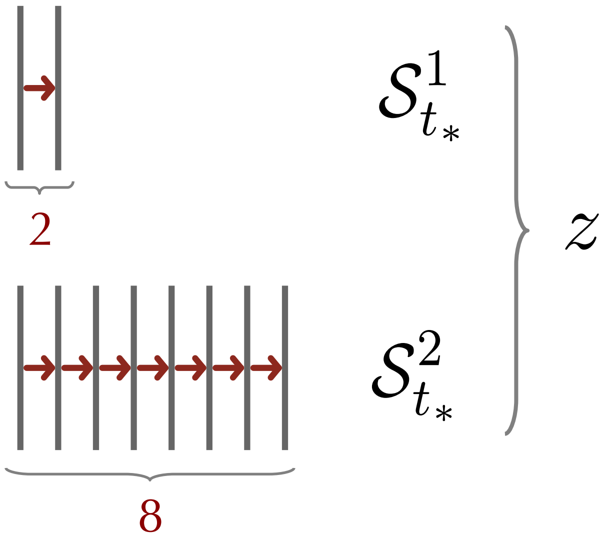
$$z_t = \frac{\langle M_t \rangle - \hat{M}_t}{\sigma_{M_t}}$$





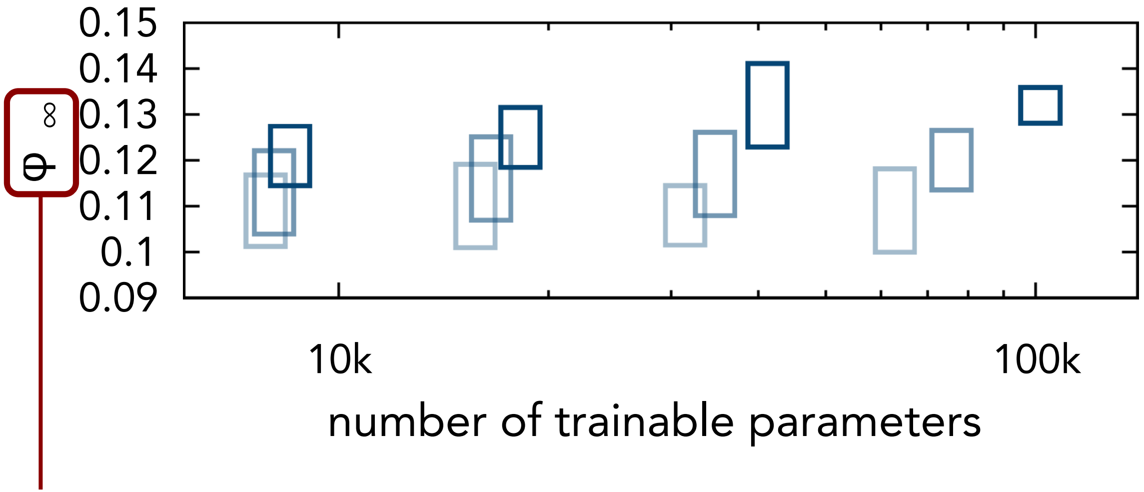
ARPIT, JASTRZEBSKI, ET AL., ICML (2017)

STRAGGLERS ARE CONSERVED ACROSS ARCHITECTURES



with different fully-connected architectures

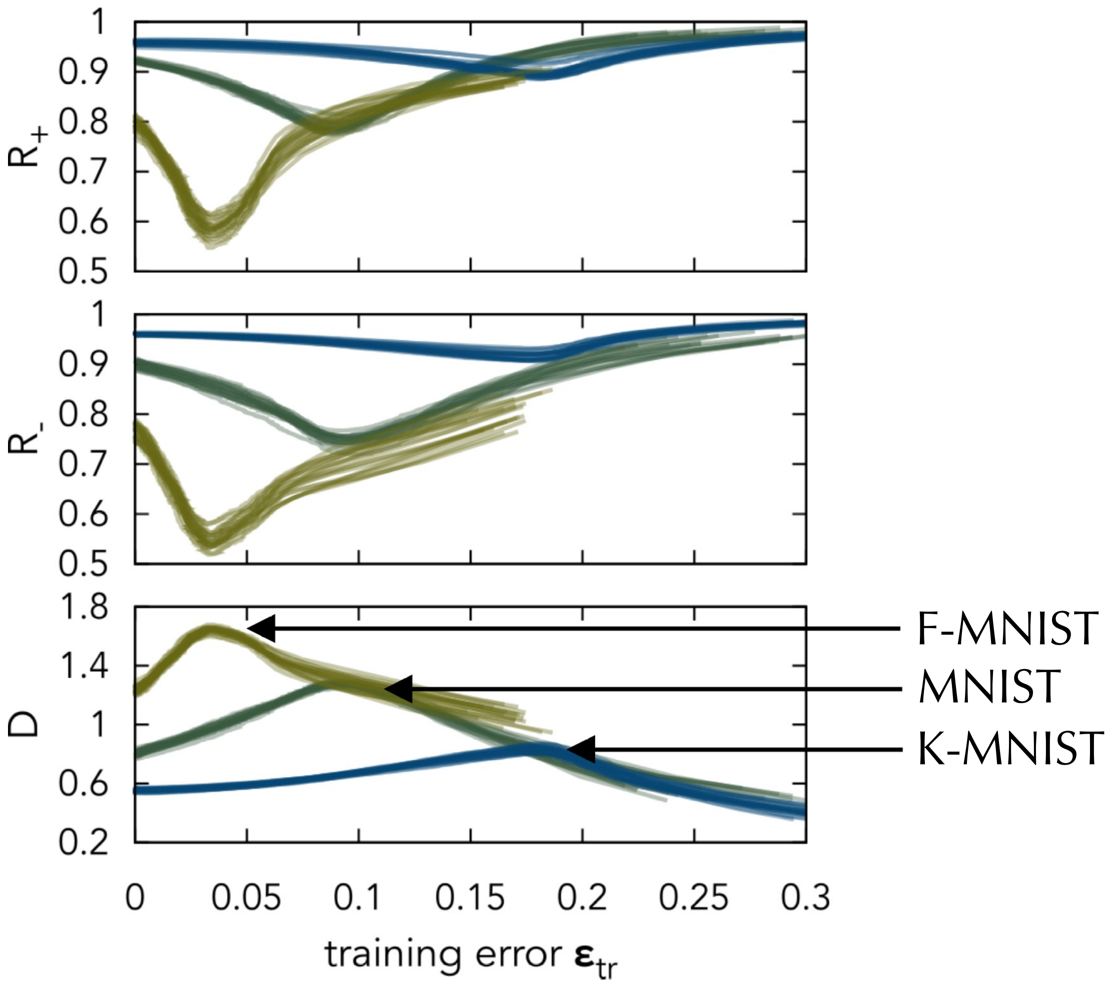
$$z = 40 \sim 50$$

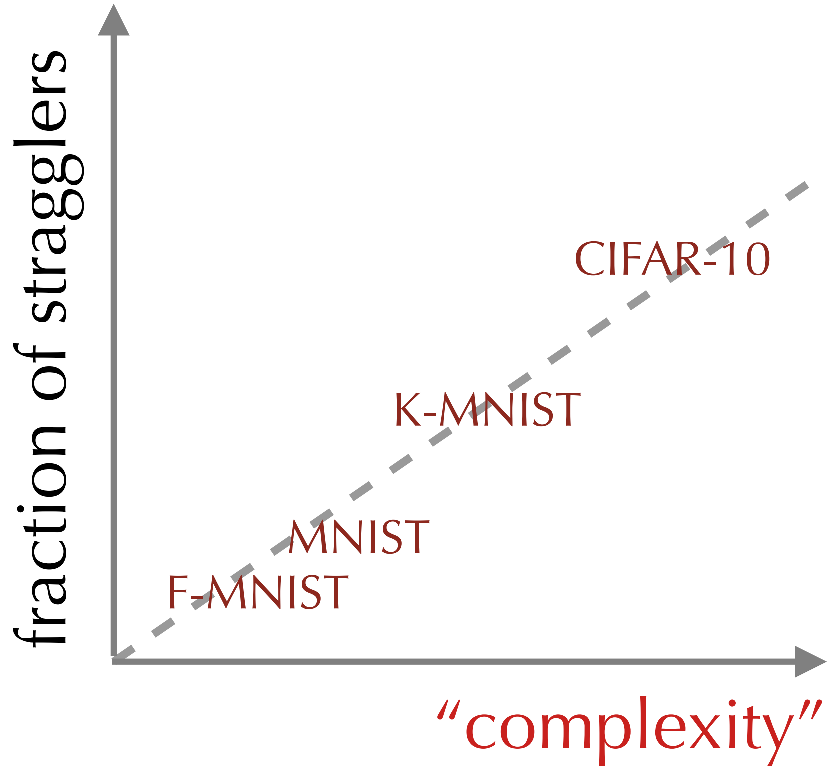


fraction of stragglers in large dataset depends very weakly on architecture

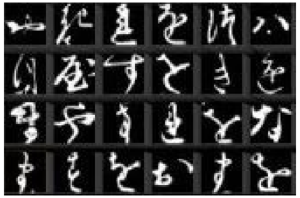
UNIVERSALITY

STRAGGLERS IN OTHER DATA SETS





fashion MNIST



Kuzushiji MNIST

HOW DO STRAGGLERS AFFECT GENERALIZATION ?

	training set	test error
F-MNIST	$\mathcal{D}$	3.5 %
	$\mathcal{D} \setminus \mathcal{S}_{t_*}$	4 %
MNIST	$\mathcal{D}$	5 %
	$\mathcal{D} \setminus \mathcal{S}_{t_*}$	10 %
K-MNIST	$\mathcal{D}$	2 %
	$\mathcal{D} \setminus \mathcal{S}_{t_*}$	30 %

$\mathcal{D}$ beats $\mathcal{D} \setminus \mathcal{S}_{t_*}$

HOW DO STRAGGLERS AFFECT GENERALIZATION ?

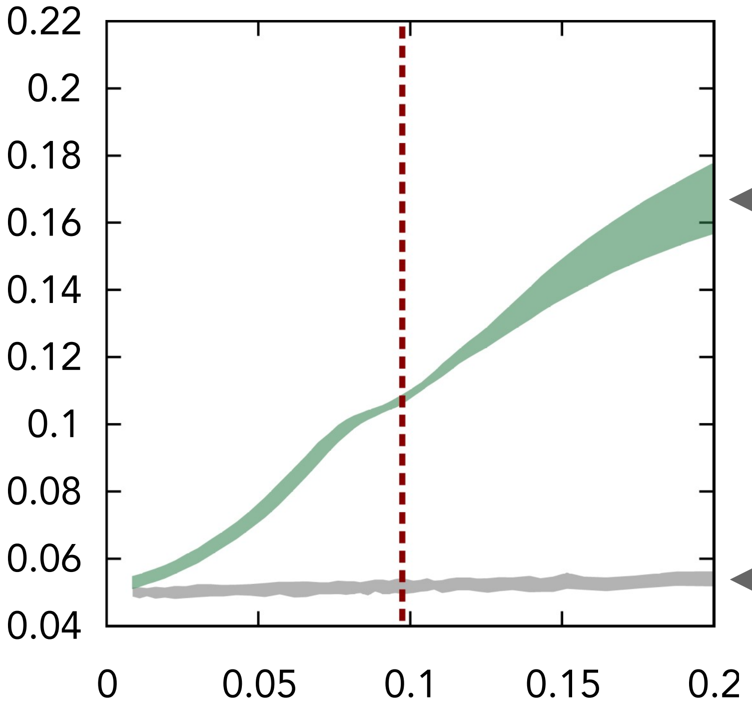
	training set	test error
F-MNIST	$\mathcal{D}$	3.5 %
	$\mathcal{D} \setminus \mathcal{S}_{t_*}$	4 %
MNIST	$\mathcal{D}$	5 %
	$\mathcal{D} \setminus \mathcal{S}_{t_*}$	10 %
K-MNIST	$\mathcal{D}$	2 %
	$\mathcal{D} \setminus \mathcal{S}_{t_*}$	30 %

$\mathcal{D}$ beats $\mathcal{D} \setminus \mathcal{S}_{t_*}$

stragglers determine generalization ?

it's not that simple !

what about $\mathcal{D} \setminus \mathcal{S}_t$?



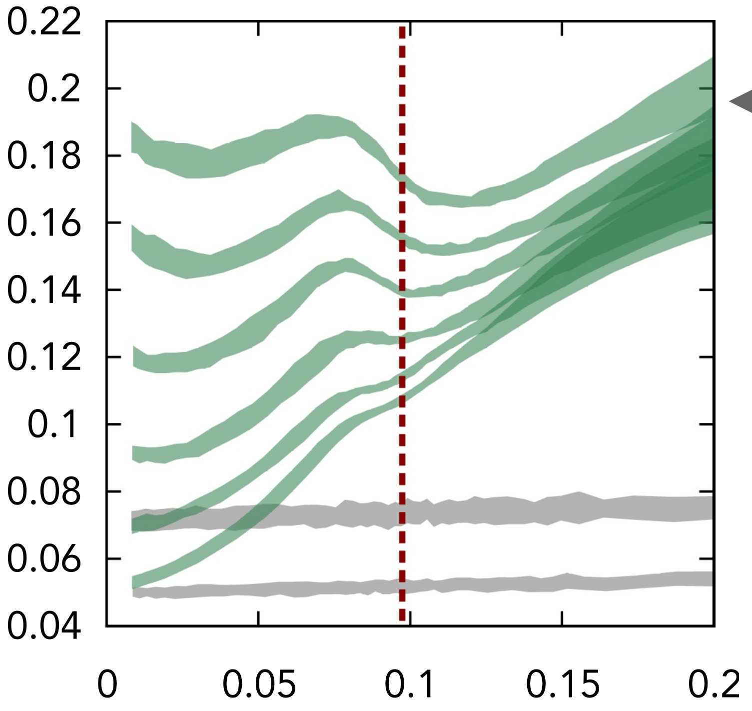
test error of model trained on $\mathcal{D} \setminus \mathcal{S}_t$

test error of model trained on $\mathcal{D} \setminus \mathcal{R}_t$

random subset with the same cardinality as $\mathcal{S}_t$

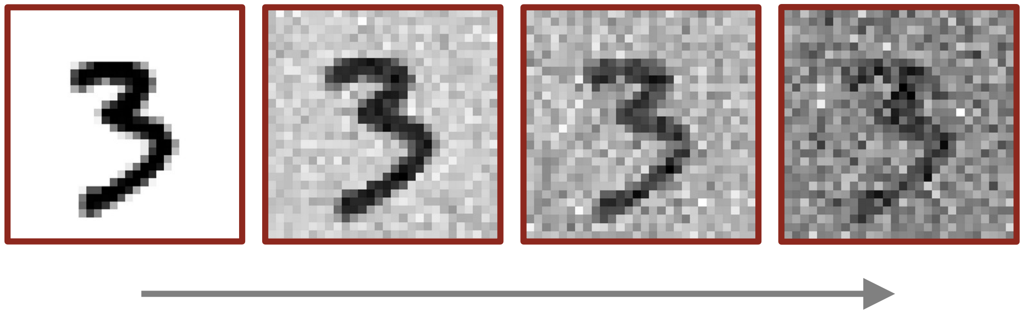
$$\epsilon_{\mathrm{tr}}(t) = \frac{|\mathcal{S}_t|}{|\mathcal{D}|}$$

STRAGGLERS HAMPER OUT-OF-DISTRIB GENERALIZATION

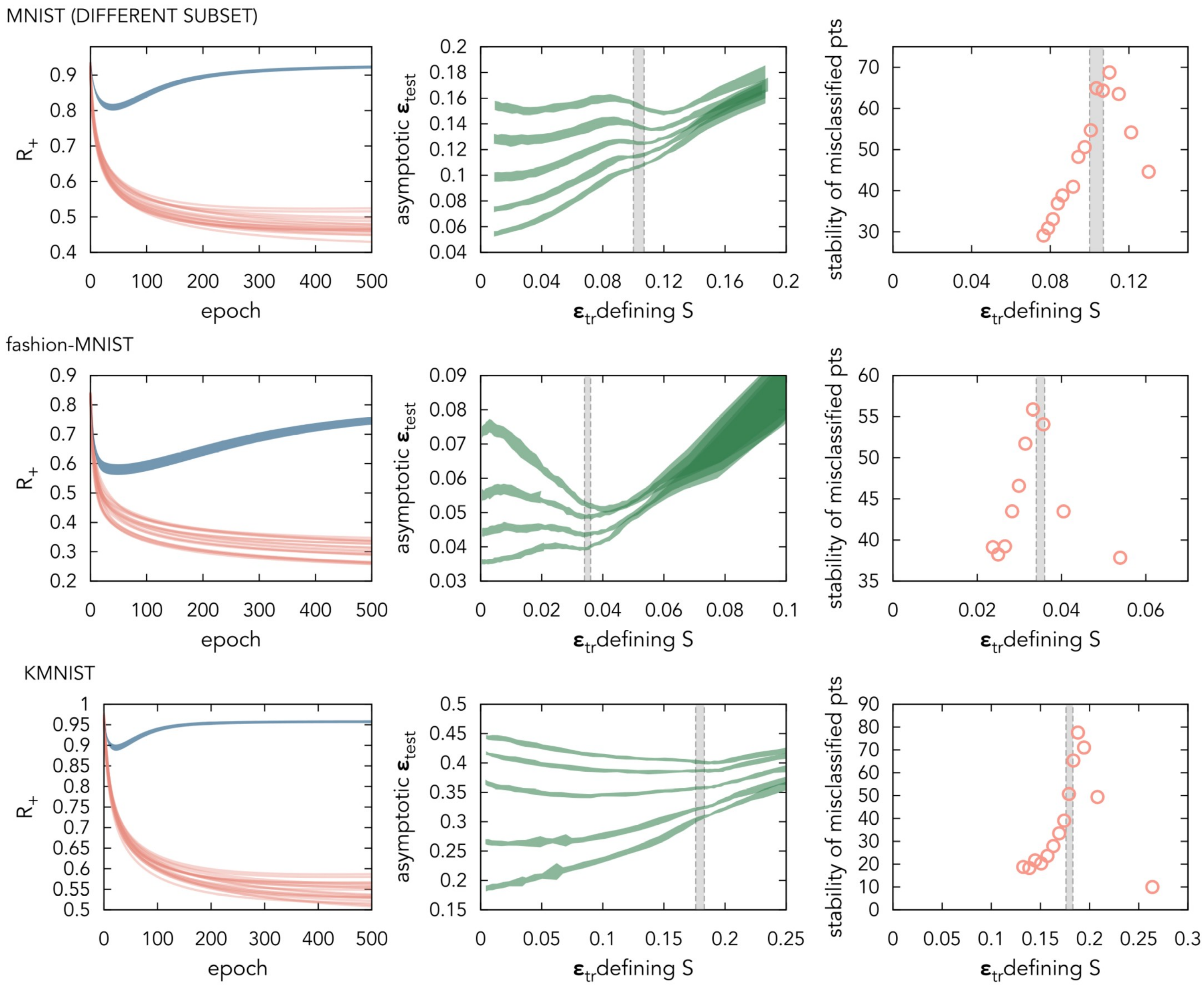


$$\epsilon_{\mathrm{tr}}(t) = \frac{|\mathcal{S}_t|}{|\mathcal{D}|}$$

test error evaluated on increasingly noisy test sets



white noise

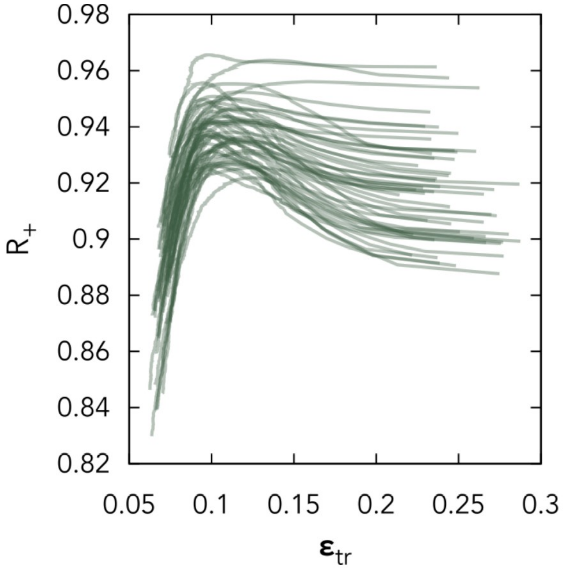


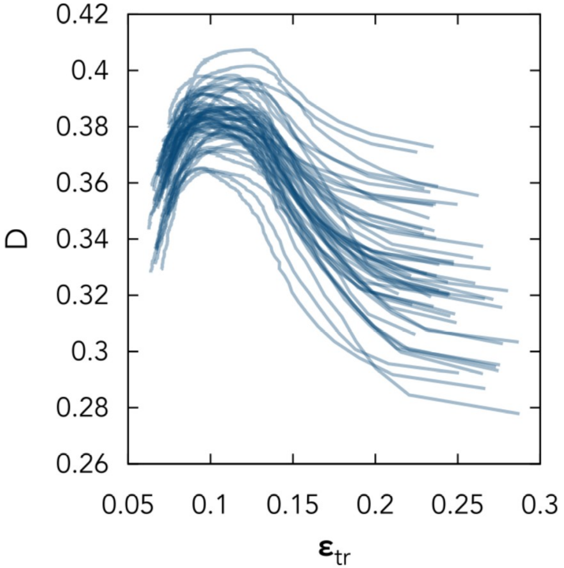
stragglers seem to be a universal property of empirical data sets

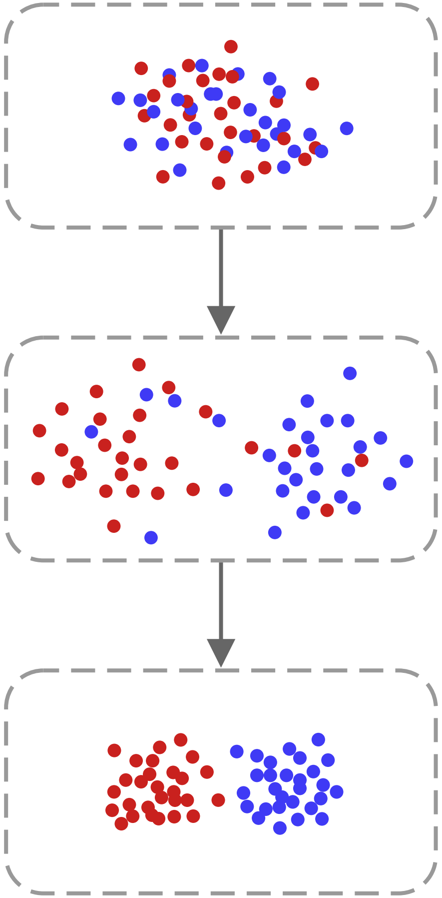
THESES AVAILABLE !

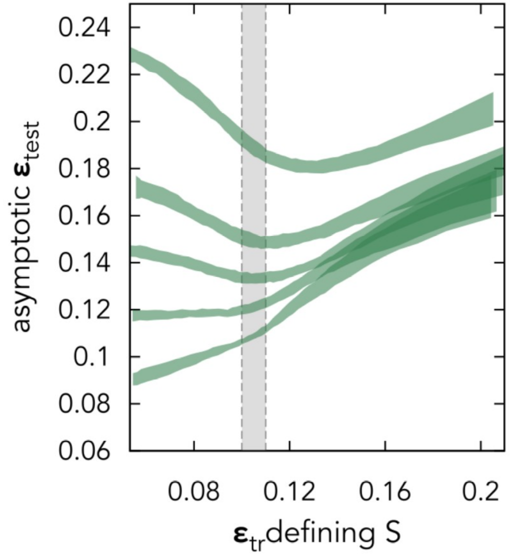
simple CNN

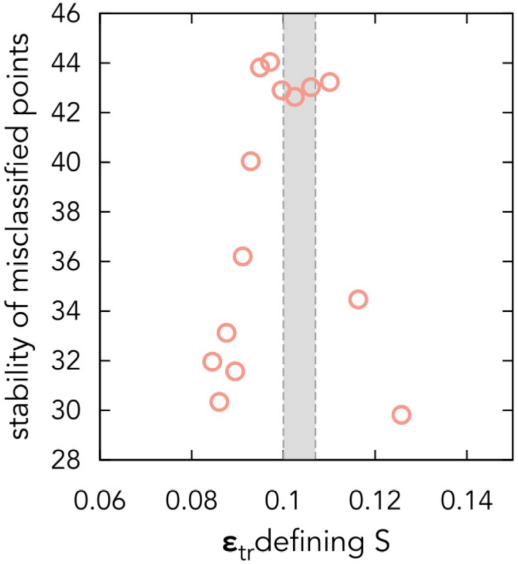
→ Conv(10,4,4)
→ Tanh
→ Flatten
→ Linear





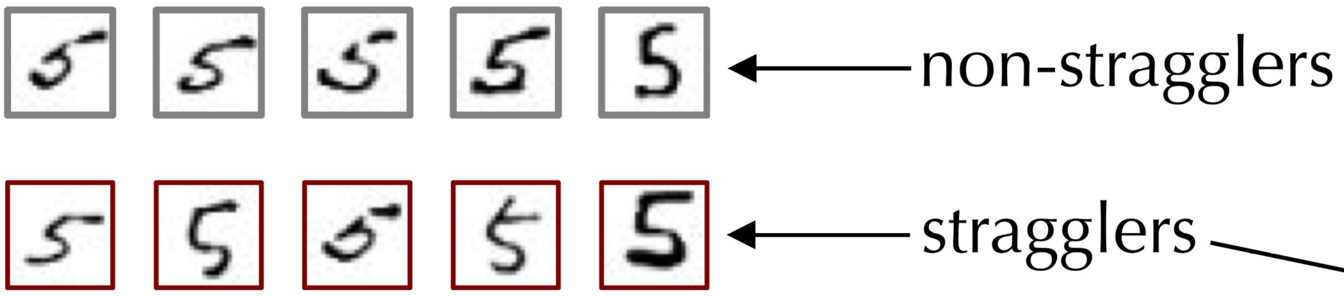


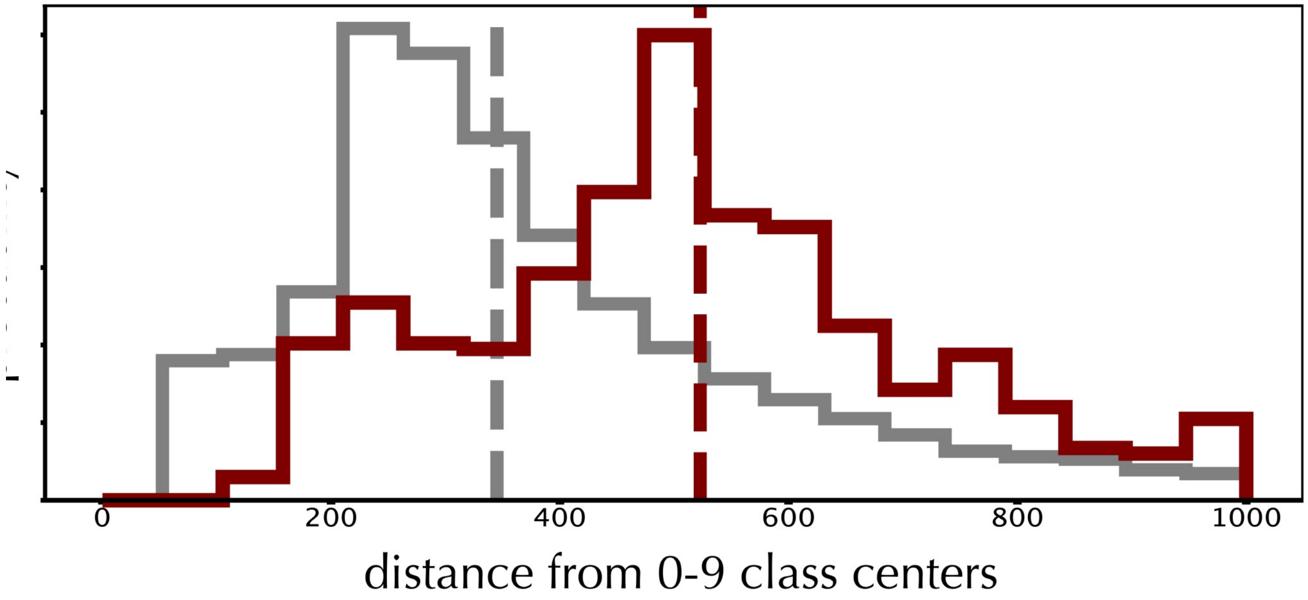


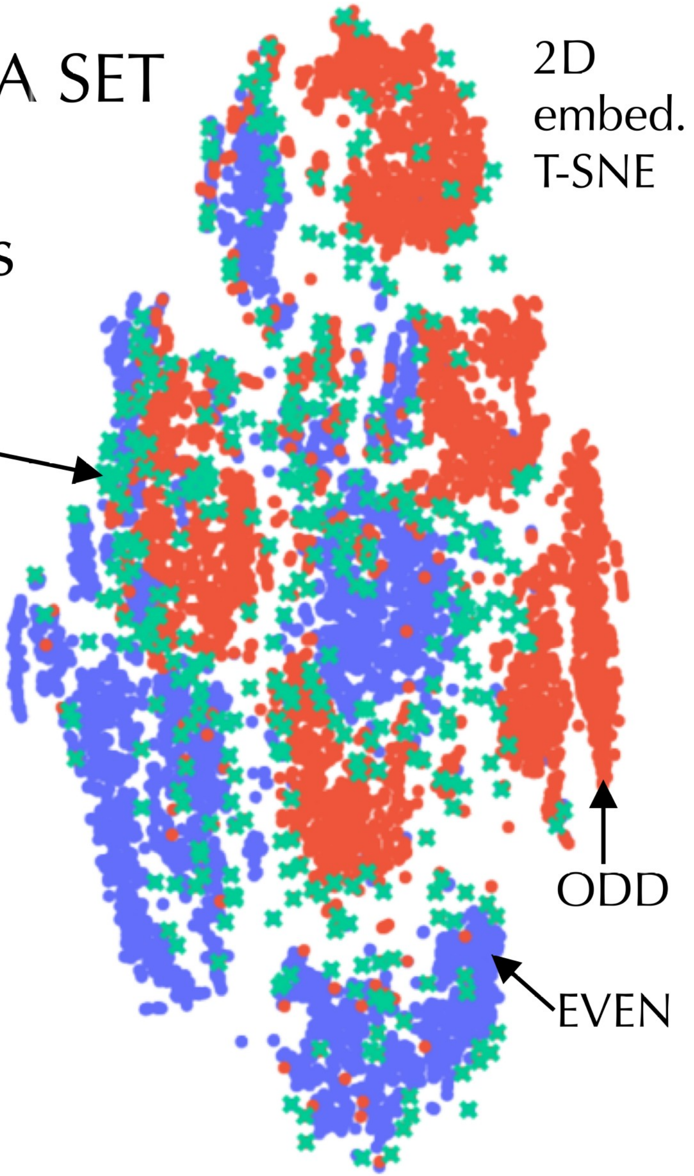


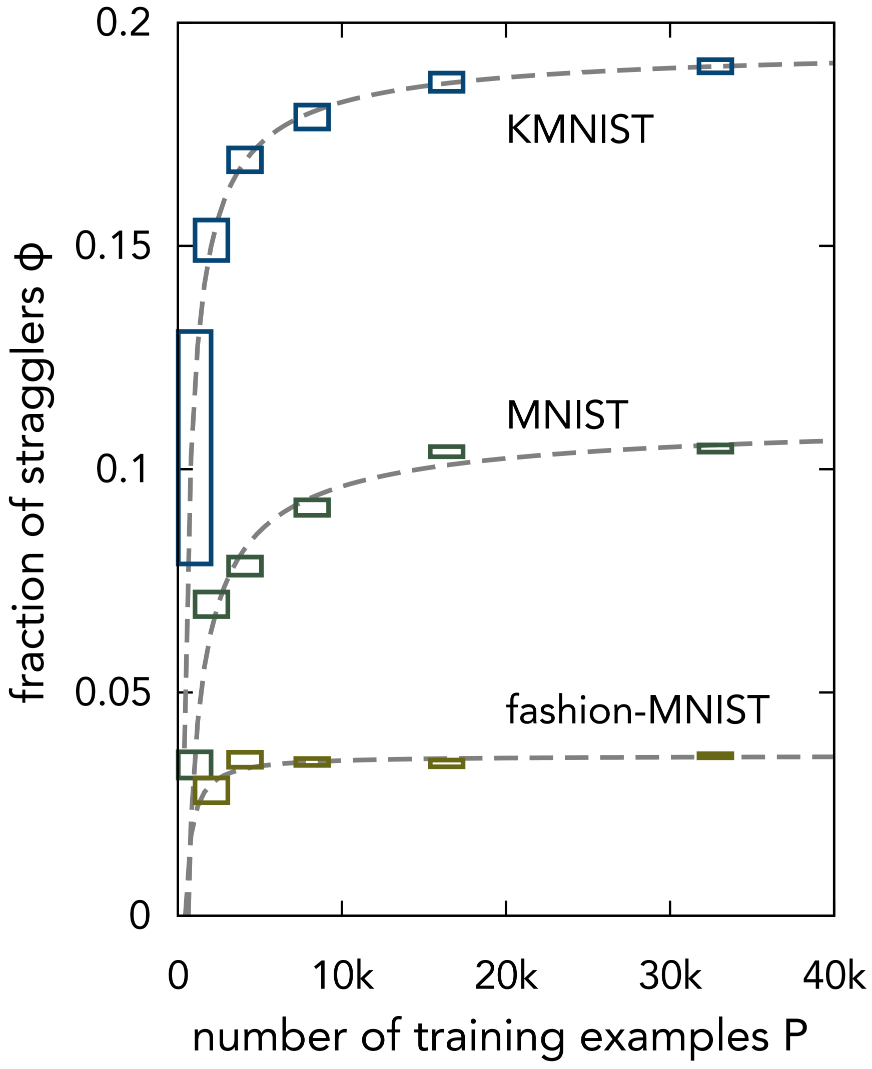
stragglers seem to universally affect different architectures

STRAGGLERS ARE PERIPHERAL IN THE DATA SET









finite-size scaling

$$\phi(P) \approx \phi_\infty \left[1 - \left(\frac{P}{P_0} \right)^{-\gamma} \right]$$