The Emergence Theory of Representation Learning

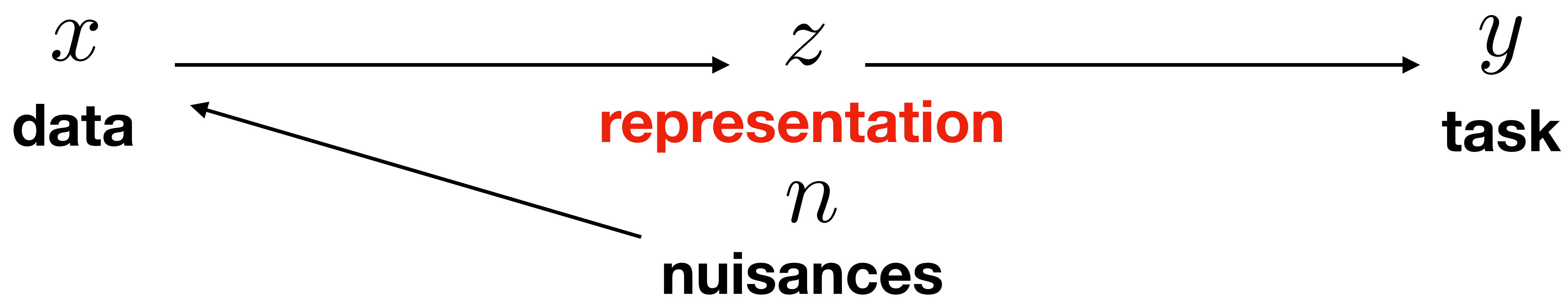
Stefano Soatto, Alessandro Achille
UCLA & AWS

October 3, 2019

menu

- **Part I: What is a representation?** Desiderata
  - What variational principle(s) define optimal representations?
- **Part II: What does Deep Learning have to do with it?** Generalization and the Information Lagrangian
  - Duality and the Emergence Bound
  - Where is the information in deep neural networks?
- **Part III: What if the task is not known completely?**
  - Critical Learning Periods
  - The space of learning tasks
  - Task Topology, Task Reachability

Desiderata of Representations



$y = x$

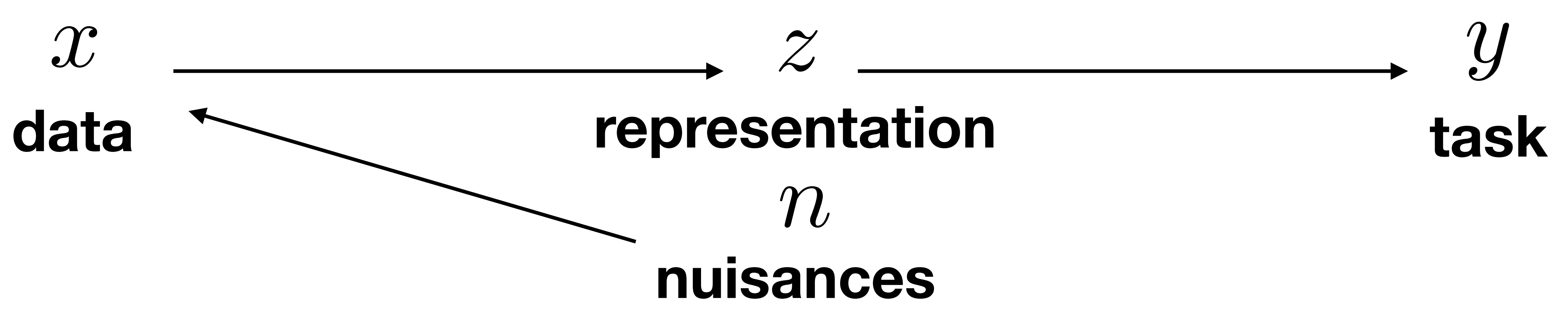
(compression, auto encoding, prediction); encompasses supervised, un-supervised, self-supervised, semi-supervised…

Desiderata of Representations

$$\begin{tikzcd} \underset{\textbf{data}}{x} \arrow[r] & \underset{\textbf{representation}}{z} \arrow[r] & \underset{\textbf{task}}{y} \\ & \underset{\textbf{nuisances}}{n} \arrow[ul] & \end{tikzcd}$$

- Sufficient (for the task) $\quad I(y;z) = I(y;x)$
- Invariant (to nuisances) $\quad n \perp y \ \Rightarrow I(n;z) = 0$
- Minimal $\quad I(x;z) = \text{minimal}$

A Variational Principle?



- Sufficient (for the task) $I(y; z) = I(y; x)$
- Invariant (to nuisances)
- Minimal (information) $I(z; x) = \text{smallest}$

$$\min_{q(z|x)} \mathcal{L} \doteq H_{p,q}(y|z) + \beta I(z; x)$$

Information Bottleneck (IB)

[Tishby-Bialek-Pereira ’99]

A Variational Principle?

- Claim: z is sufficient, n a nuisance; then

$$\underbrace{I(z;n)}_{\text{invariance}} \leq \underbrace{I(z;x)}_{\text{minimality}} - \underbrace{I(x;y)}_{\text{constant}}$$

- and there exists a nuisance for which equality holds

A. Achille and S. Soatto, *On the Emergence of Invariance and Disentangling in Deep Representations;* JMLR 2018; ArXiv:1706.01350

Examples

- **Nuisances have a group structure: Maximal Invariance**
  - **Localization (SLAM)**
  - **Diffeo/homeomorphisms of the domain and range of an image:**
    - General viewpoint and illumination invariants (Attributed Reeb Trees) [Sundaramoorthi, Varadarajan, etc.] 2005-2009
  - **Local affine domain & range transformations:**
    - DSP-SIFT [Dong] 2011-2015
  - **Non-invertible nuisances:**
    - Occlusions, Scale… Give up on Maximal Invariance

This Information Bottleneck is wishful thinking

- The task is a function of (test) data *we have not yet seen*!

$$z \sim p(z|x)$$

- The Information Bottleneck is a statement of desire

$$\min_{q(z|x)} \mathcal{L} \doteq H_{p,q}(y|z) + \beta I(z;x)$$

A. Saxe et al., On the Information Bottleneck Theory of Deep Learning, ICLR 2018

Desiderata of Deep Learning

GSD

$$\mathcal{D} \longrightarrow q_w(y|x) \longrightarrow p(y|x)$$

dataset **model**

$$\{x_i, y_i\} \sim p \quad i = 1, \dots, N$$

?

Empirical Cross-Entropy

SGD

$$q_w(y|x) = \arg\min H_{p,q}(\mathcal{D}|w) + \beta I(\mathcal{D}; w)$$

$$- \sum_{i=1}^{N} \log q_w(y_i|x_i)$$

generalization

$$L_{test}(q_w(y|x)) \leq \frac{1}{N(1 - 1/2\beta)} \left[H_{p,q}(\mathcal{D}|w) + \overbrace{\beta KL(q(w|\mathcal{D})||p(w))}^{+\beta I(\mathcal{D}; w)}\right]$$

PAC-Bayes bound (Catoni, 2008; McAllester, 2013)

The Information Lagrangian

$$\mathcal{D} \longrightarrow q_w(y|x) \longrightarrow p(y|x)$$

dataset **model**

$$\{x_i, y_i\} \sim p \quad i = 1, \ldots, N$$

Past data (training set)

?

$$\boxed{\mathcal{L}_N(w) \doteq H_{p,q}(\mathcal{D}|w) + \beta I(\mathcal{D}; p(w))}$$

Future data (test sample)

$$\boxed{\mathcal{L} \doteq H_{p,q}(y|z) + \beta I(z; x)}$$

$$L_{test}(q_w(y|x)) \leq \frac{1}{N(1 - 1/2\beta)} \left[H_{p,q}(\mathcal{D}|w) + \beta KL(q(w|\mathcal{D})||p(w))\right]$$

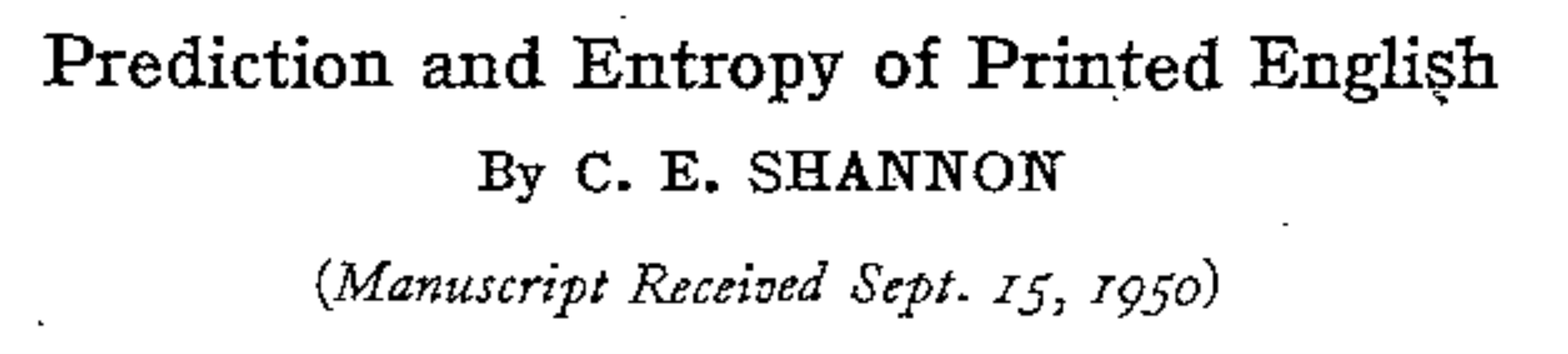
A few questions (preview)

- What is the relation between the two bottlenecks?
- What “information”? the weights are fixed, and there is only one dataset!
- What is the “prior”? and the “posterior”?
- The second term of the Information Lagrangian is not there in practice!

Measuring Information by Adding Noise

$$L(w) = H_{p,q}(\mathcal{D}|w) + \beta\, \mathrm{KL}(q(w|\mathcal{D}) \,\|\, p(w))$$

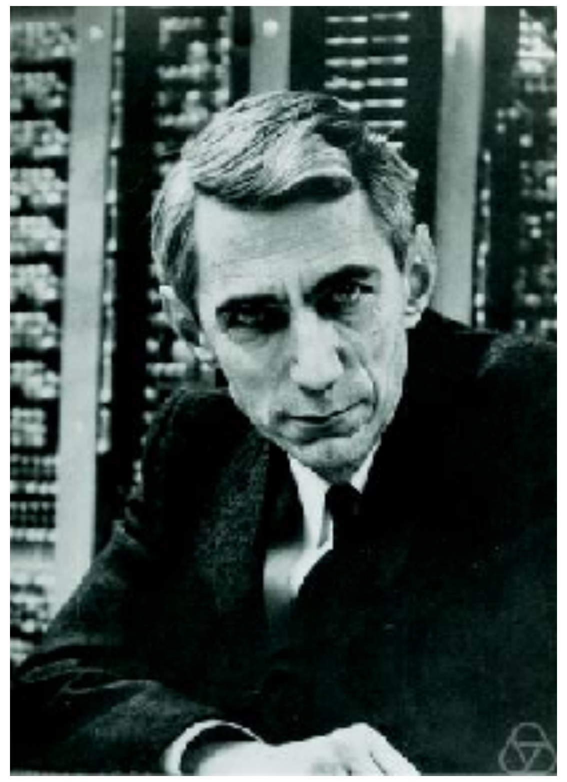
Idea: We can estimate the amount of information contained in the weights by corrupting them with noise and measuring the decrease in performance.



Prediction and Entropy of Printed English

By C. E. SHANNON

(Manuscript Received Sept. 15, 1950)



Example: Shannon (1951) estimates the information content of the English language by corrupting random letters and measuring the reconstruction error of English speakers.

“Thif is a vevy moisy party” → “This is a very noisy party”

C. Shannon, *Prediction and Entropy of Printed English*, Bell System Technical Journal, *1951*

The Information in a Deep Neural Network

$$L(w) = \underbrace{H_{p,q}(\mathcal{D}|w) + \beta\, \mathrm{KL}(q(w|\mathcal{D})}_{\text{output of training}} \,\|\, \underbrace{p(w)}_{\text{fixed prior}})$$

Fisher Information: Gaussian prior

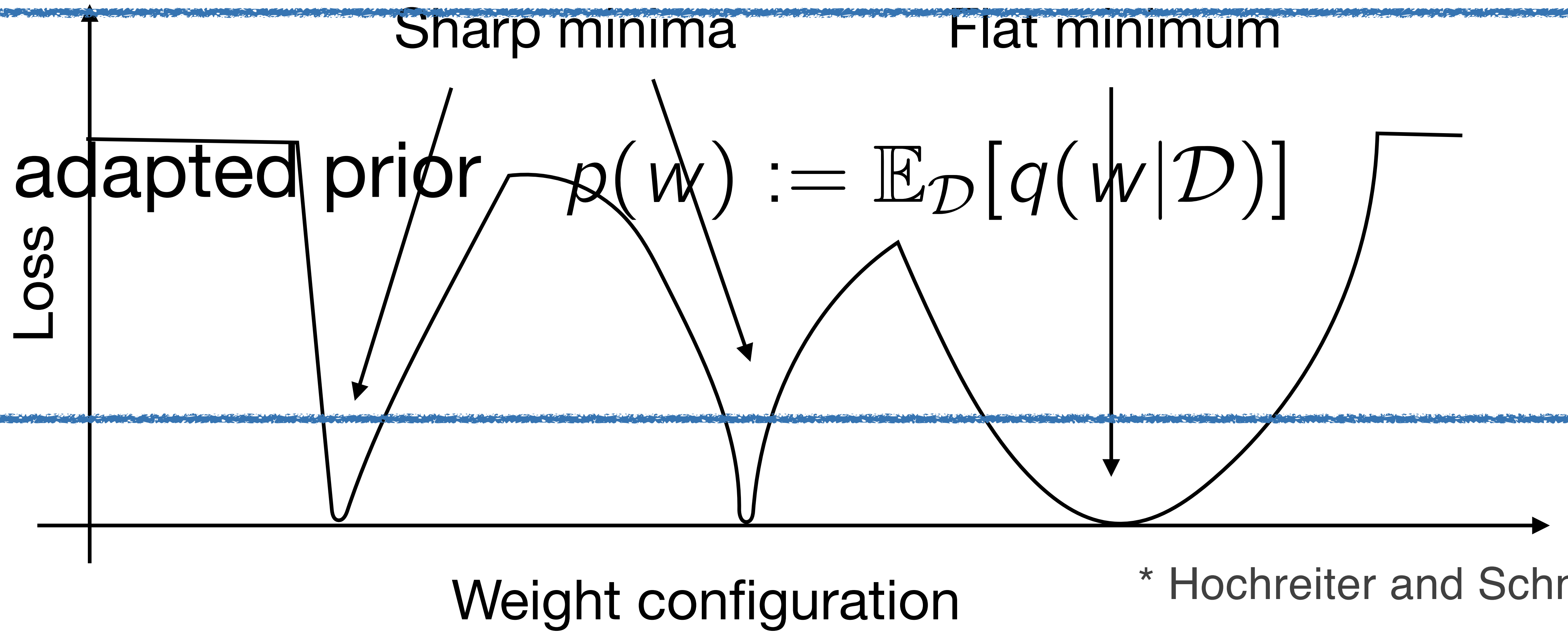
$$\mathrm{KL} = \frac{\|w\|^2}{\lambda^2} + \log|2\lambda^2 NF + I|$$

F = curvature of loss landscape

$\Rightarrow$ Implicitly minimized by SGD

Shannon Information: adapted prior $p(w) := \mathbb{E}_{\mathcal{D}}[q(w|\mathcal{D})]$

$$\mathbb{E}_{\mathcal{D}}[\mathrm{KL}] = I(w;\mathcal{D})$$



* Hochreiter and Schmidhuber, *Flat Minima*, *Neural Computation* ,*1997*

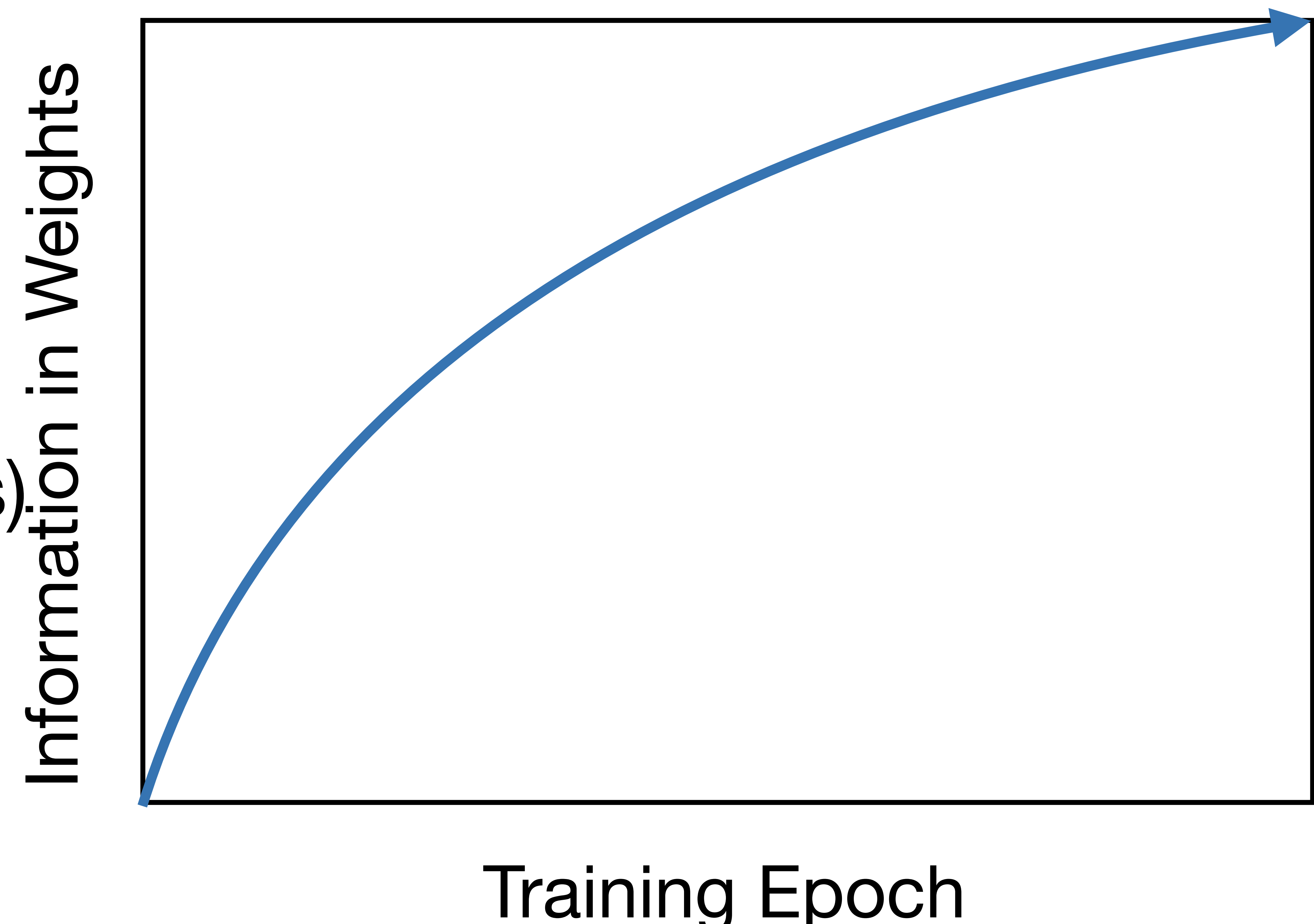
A. Achille et al., *The Information Complexity of Learning Tasks, their Structure and their Distance,* ArXiv 2019

H. Li et al., Visualizing the Loss Landscape of Neural Nets, ICLR 2018

A few questions (preview)

$$L(w) = H_{p,q}(\mathcal{D}|w) + \beta\,\mathrm{KL}(q(w|\mathcal{D}) \,\|\, p(w))$$

- The second term of the Information Lagrangian is not there in practice!
  - (inductive bias of SGD)
- Now that we can compute the Info in the Weights, what does it look like as we learn?
  - (critical learning periods in deep networks).



Relation between Fisher and Shannon

SGD minimizes the Fisher Information of the Weights. However, generalization is governed by the Shannon Information.

Proposition. Assuming the dataset is parametrized in a differentiable way, we have:

$$\boxed{I(w;\mathcal{D})} \approx H(\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[\log\left(\frac{(2\pi e)^k}{|\nabla_{\mathcal{D}} w^* \boxed{F(w^*)} \nabla_{\mathcal{D}} w^{*T}|}\right)\right]$$

Where $w^* = w^*(D)$ is the result of running SGD on dataset D and F(w) is the Fisher Information Matrix in w.

A. and Soatto, *Where is the Information in a Deep Network?*, 2018

Emergence Bound: Simple weights → simple activations

Let z = f_w(x) be a layer of a network, and let z_n be the representation obtained by adding noise to the weights. We define the effective information as $I_{eff}(x; z) = I(x; z_n)$

Theorem (Emergence Bound): Let z = f_w(x) be a layer of a network. To first-order, the effective information in the activations is:

Information in activations

$$\underbrace{\boxed{I_{\mathrm{eff}}(x;z)}}_{\text{Information in activations}} \approx H(x) - \log\left(\frac{(2\pi e)^k}{|\nabla_x f_w(x)^t \; J_f^t \; \underbrace{\boxed{F(w)}}_{\text{Fisher of Weights}} \; J_f \; \nabla_x f_w(x)|}\right)$$

Fisher of Weights

where $F(w)$ is the Fisher Information of the weights, J_f is the Jacobian of f_w w.r.t. w.

A. and Soatto, *Emergence of Invariance and Disentanglement in Deep Representations*, JMLR 2018
A. and Soatto, Where is the Information in a Deep Neural Network?, 2019

Two Bottlenecks

Generalization
Weights
Past

$$\mathcal{D} \longrightarrow w \longrightarrow p(y|x)$$

dataset — weights — real distribution

WHERE IS THE INFORMATION?

Information Lagrangian $\mathcal{L}_N(w) \doteq H_{p,q}(\mathcal{D}|w) + \beta KL(q(w|\mathcal{D})||p(w))$

WHAT PRIORS? WHAT POSTERIOR?

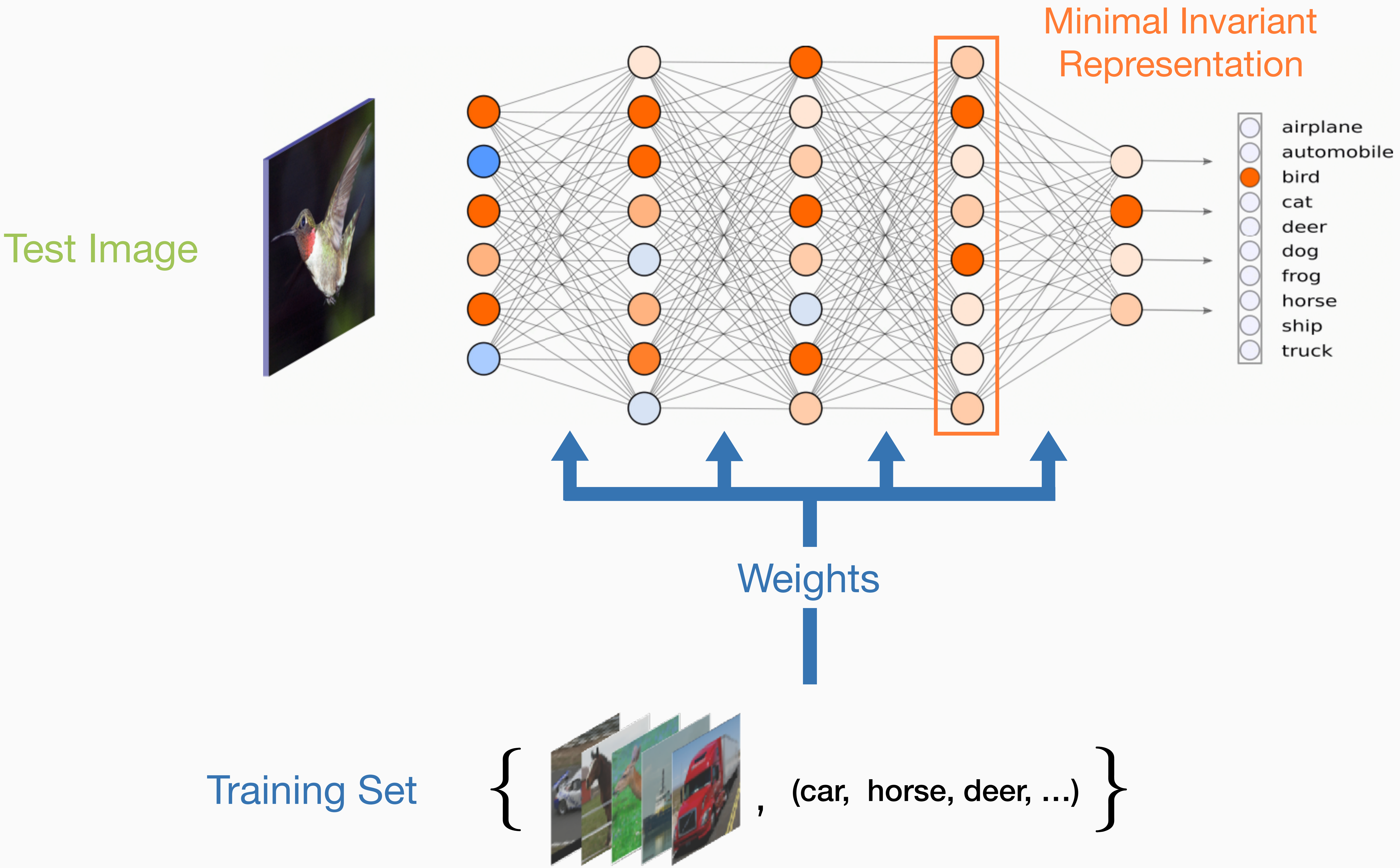
Invariance
Activations
Future

$$x \longrightarrow z \longrightarrow y$$

data — activations — label

Information Bottleneck $\mathcal{L} \doteq H_{p,q}(y|z) + \beta I(z;x)$

ANY RELATION BETWEEN THESE TWO?

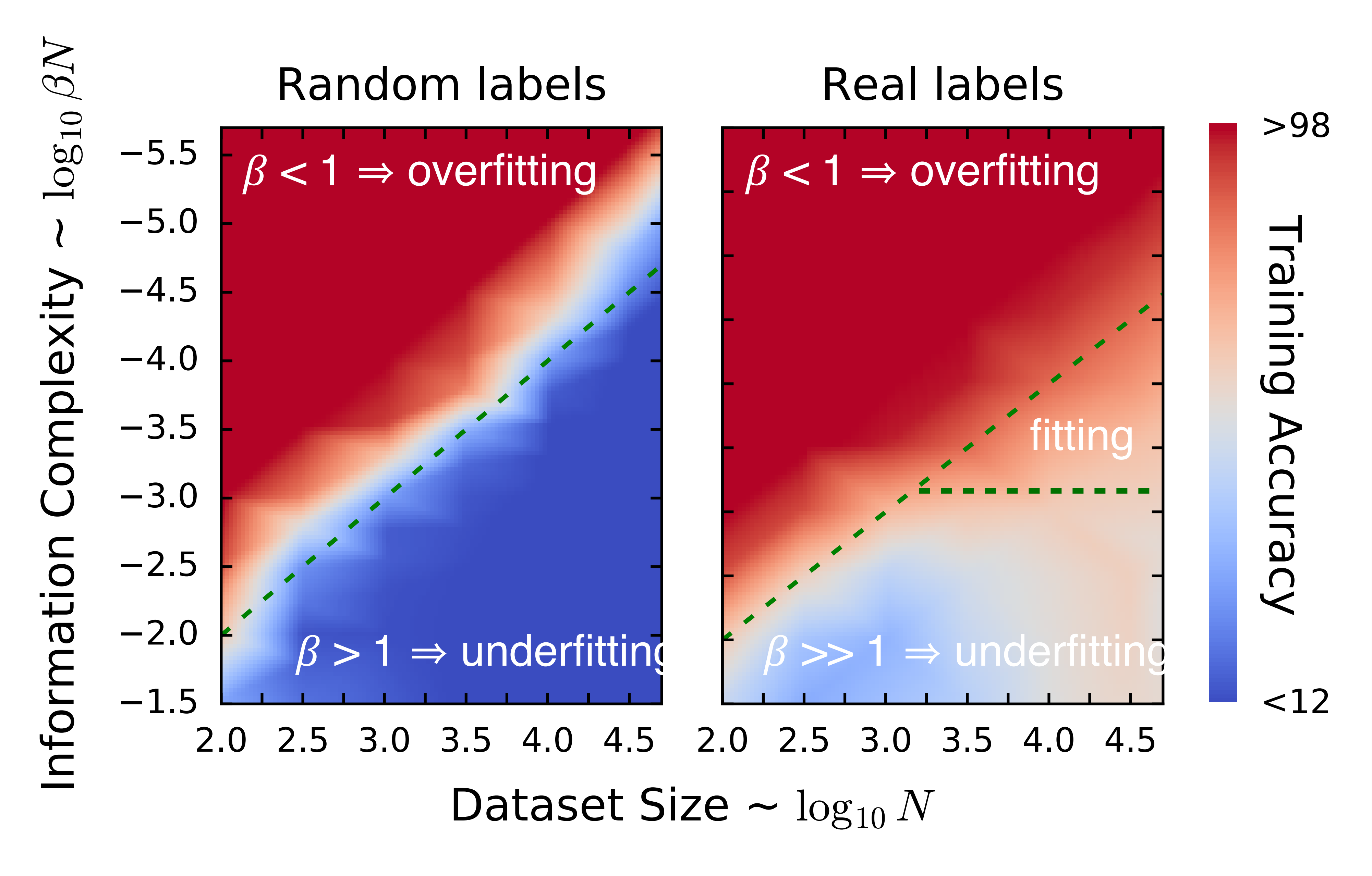

Minimal Invariant Representation
Test Image
airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck
Weights
Training Set
, (car, horse, deer, ...)

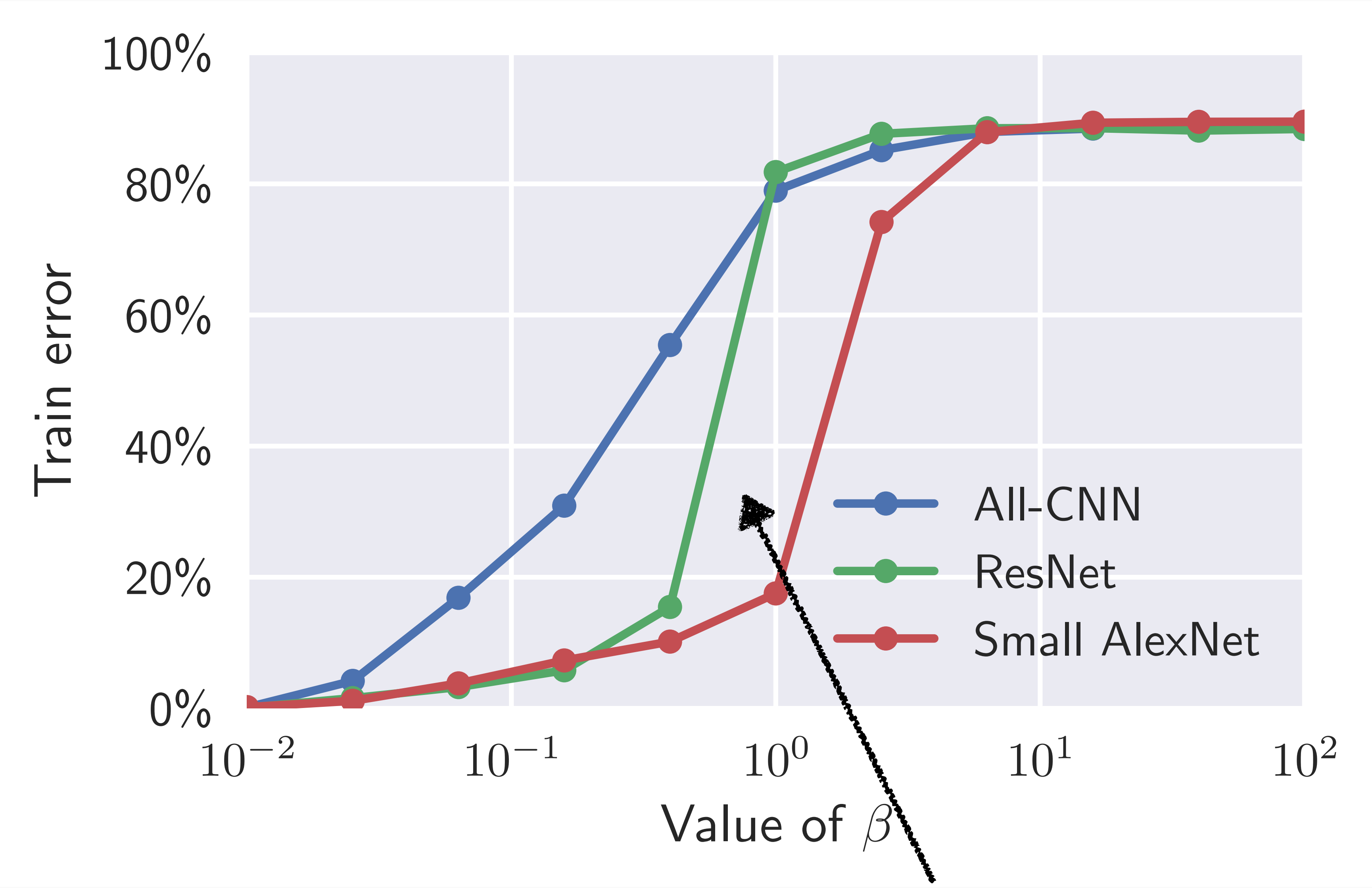
Emergence Bound

- A sufficient representation that **minimizes the information** the weights contain **about past data**, **maximizes invariance** of the representation of **future data.**

- **Pertains to the combination of DNNs (sufficient capacity to overfit) and SGD (inductive bias)**

Phase transition





Phase transition

Using the regularized loss:

$$L(w) = H_{p,q}(\mathcal{D}|w) + \beta\, KL(q(w|\mathcal{D}) \| p(w))$$

For random labels there is a transition between over- and under-fitting at $\beta = 1$.

Achille and Soatto, *Emergence of Invariance and Disentanglement in Deep Representations*, JMLR 2018

What's next?

1. This addresses what is an optimal representation *for a given task*
2. Even an optimal representation may be useless (garbage-in/garbage-out)
3. What if the task is not known ahead of time?
4. When are two tasks close? What is the distance between two tasks?
5. Can one predict if a model pre-trained on a task will perform well on another?

A Topology on the Space of Tasks

Distance between tasks:

$$d(\mathscr{D}_1 \to \mathscr{D}_2) = I(\mathscr{D}_1\mathscr{D}_2; w) - I(\mathscr{D}_1; w)$$

Complexity of learning together

Complexity of learning one

Notice that this is an asymmetric distance

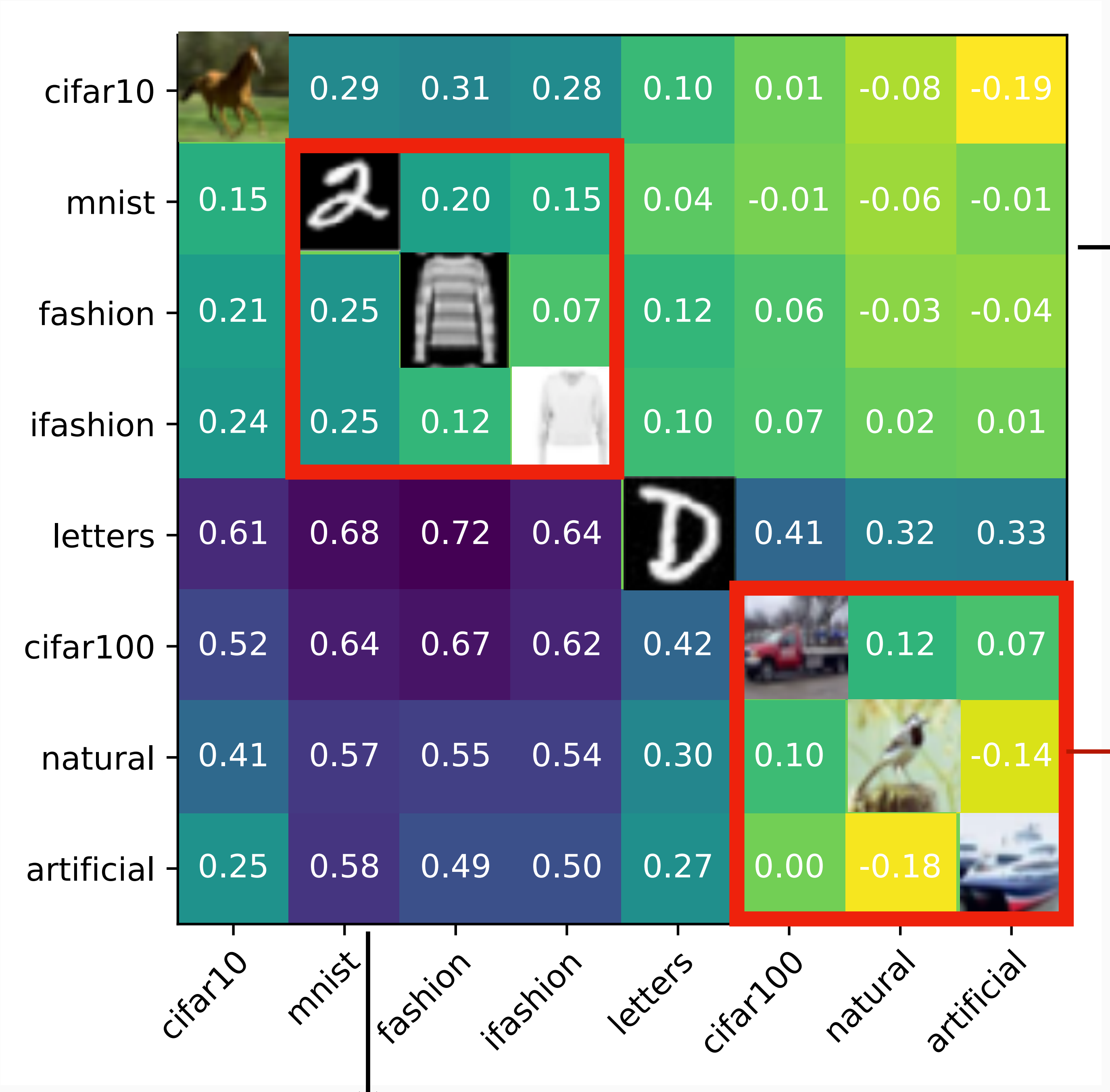
A., Paolini, Mbeng, Soatto, *The Information Complexity of Learning Tasks, their Structure and their Distance*, 2019

A Topology on the Space of Tasks

Kolmogorov (asymmetric) distance between tasks:

$$d(\mathcal{D}_1 \to \mathcal{D}_2) = K(\mathcal{D}_2 | \mathcal{D}_1)$$

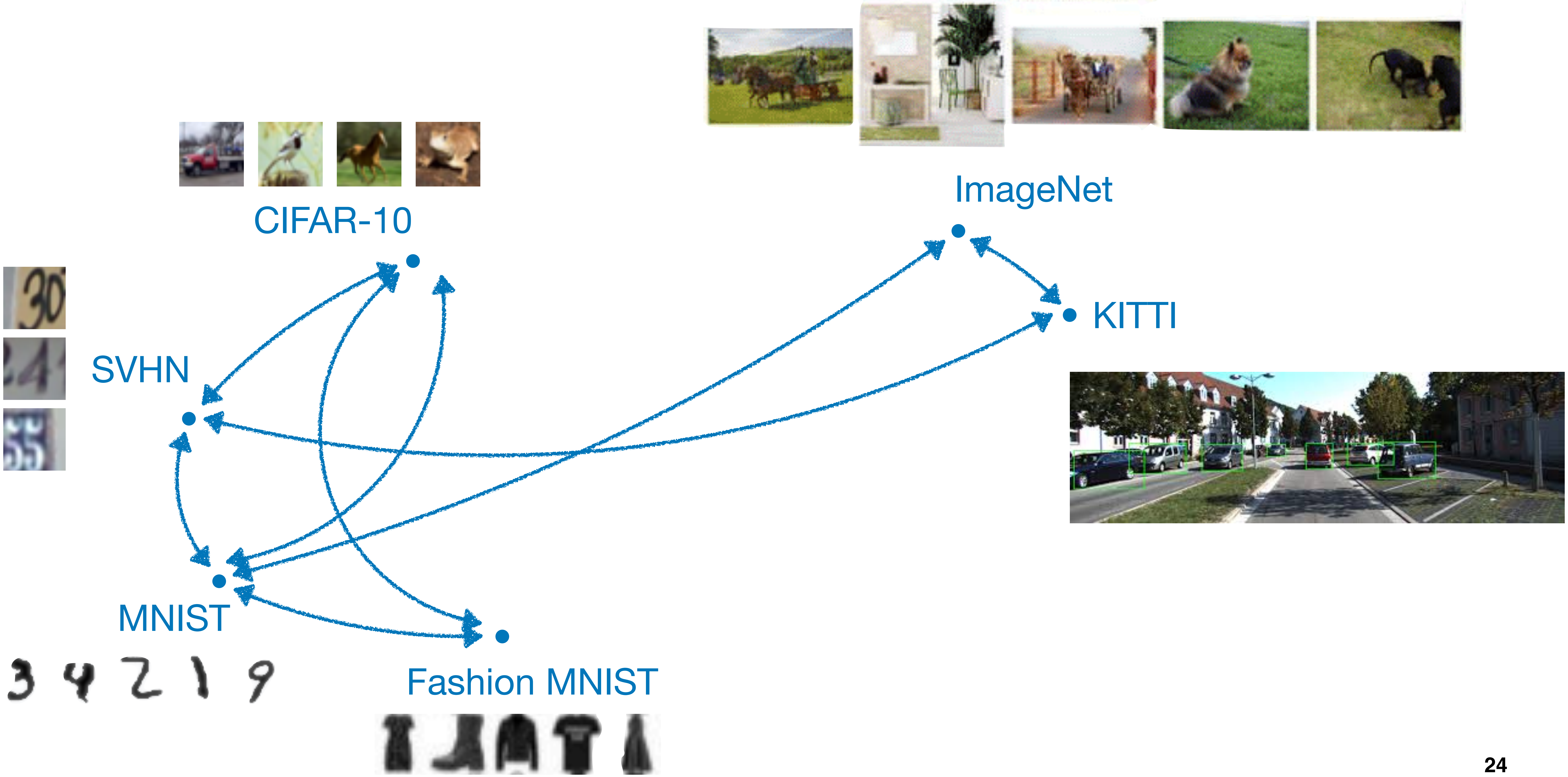
How much more structure do we need to learn?



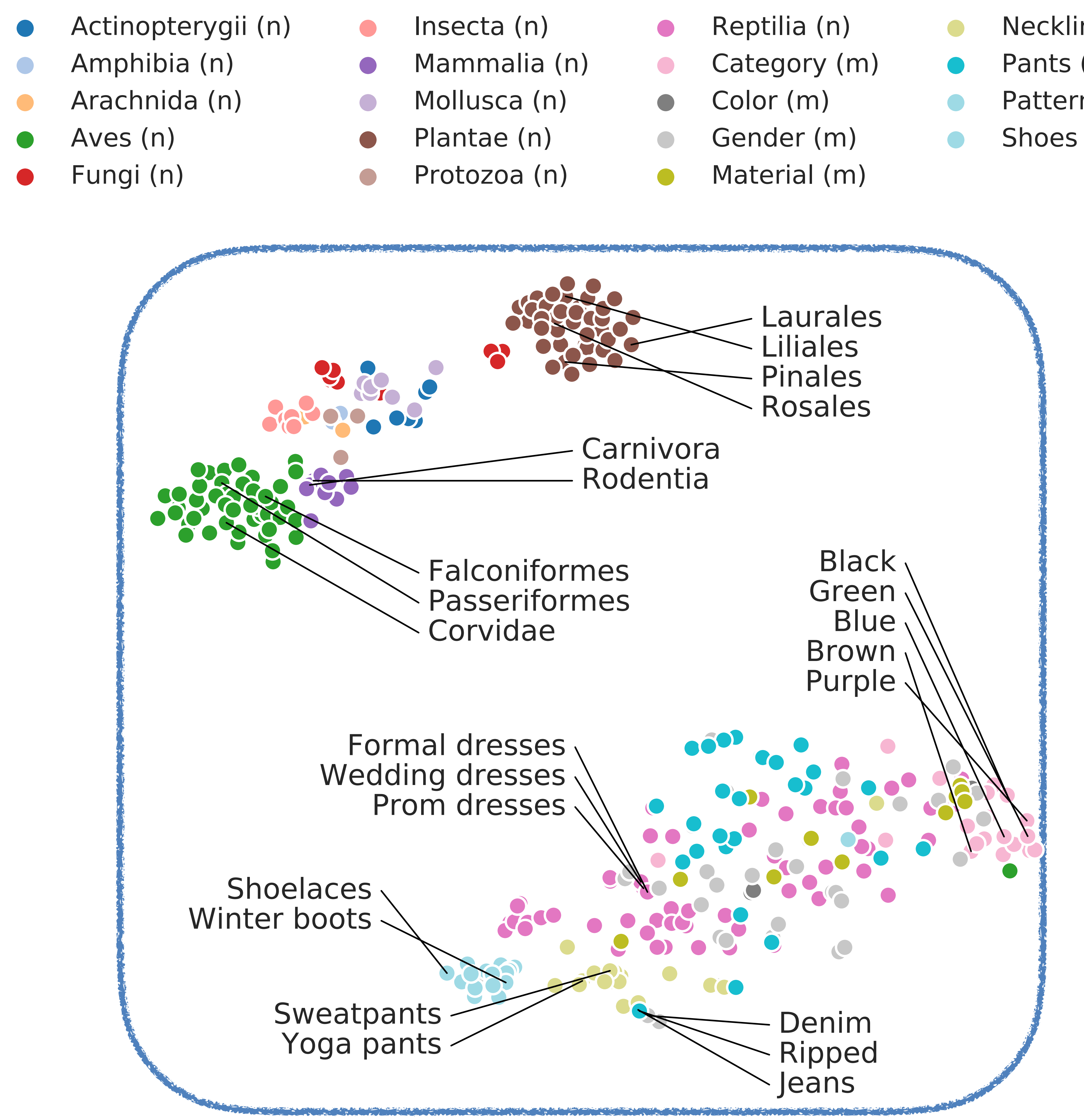
Difficult task to easy task

Similar tasks cluster together

Easy task to difficult task

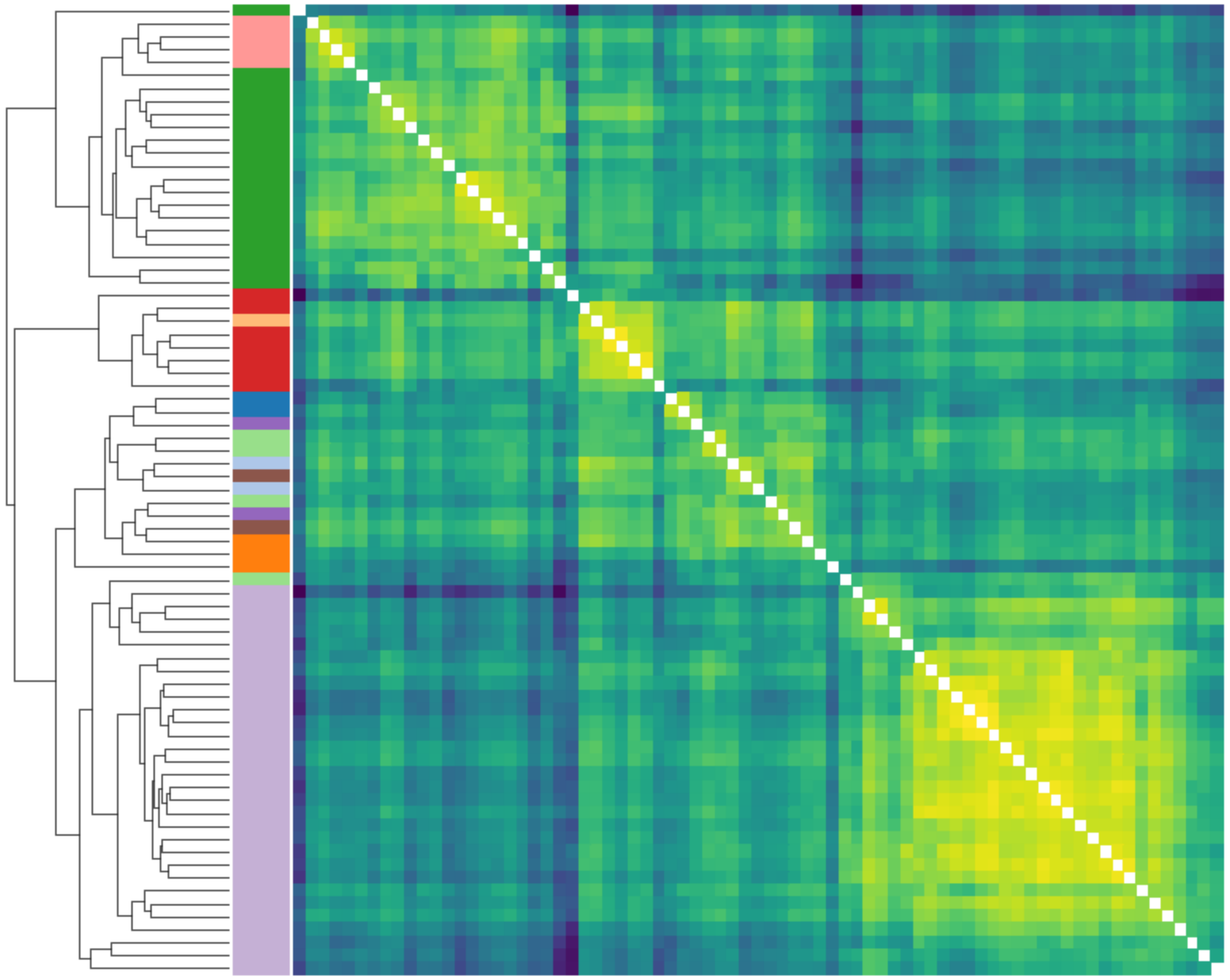

CIFAR-10
ImageNet
KITTI
SVHN
MNIST
Fashion MNIST

TASK2VEC: Embedding tasks in a metric space



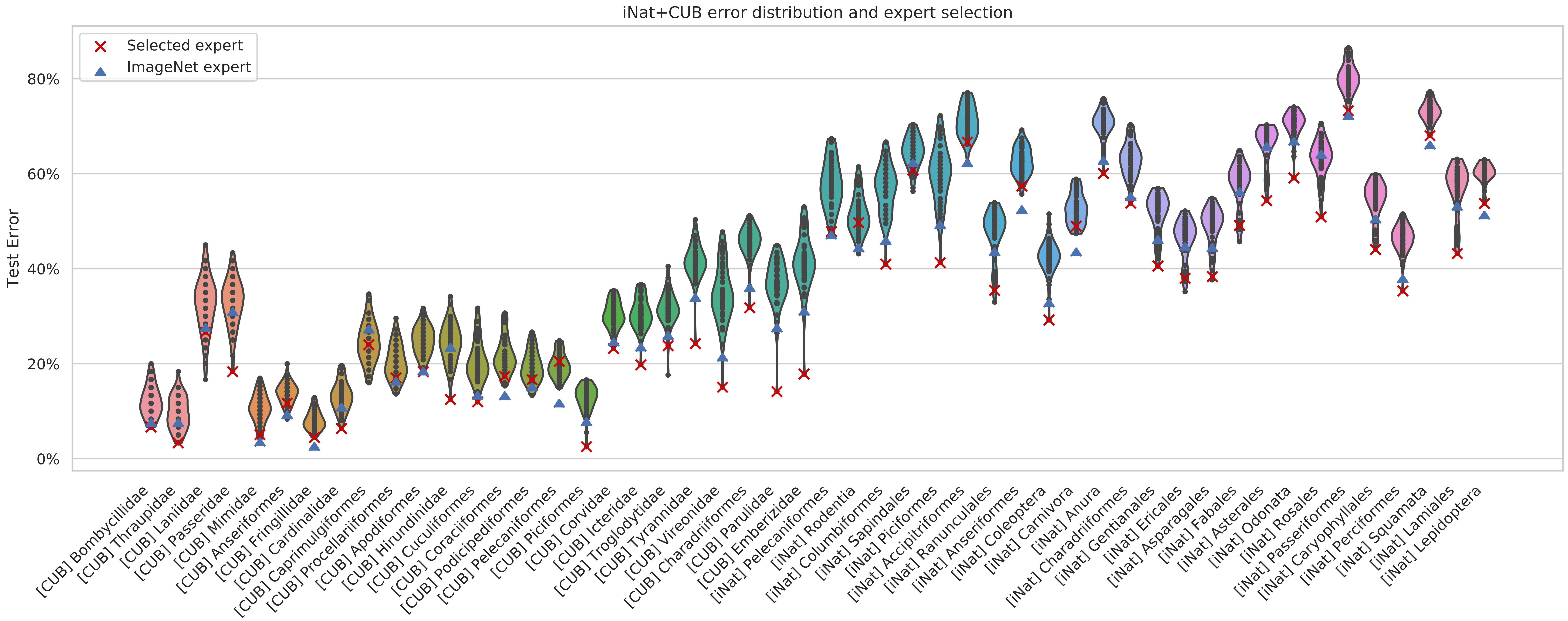
Recovers a meaningful topology on hundred of tasks

Recovers species taxonomy on iNaturalist



A. et al., TASK2VEC: Task embedding for meta-learning, 2019

Proposing an optimal expert for the task



Allows to select the best expert to solve a task and substantially reduce error and training time.

A snag: Critical Periods

Two almost identical tasks, yet it is not possible to fine-tune from one to the other.

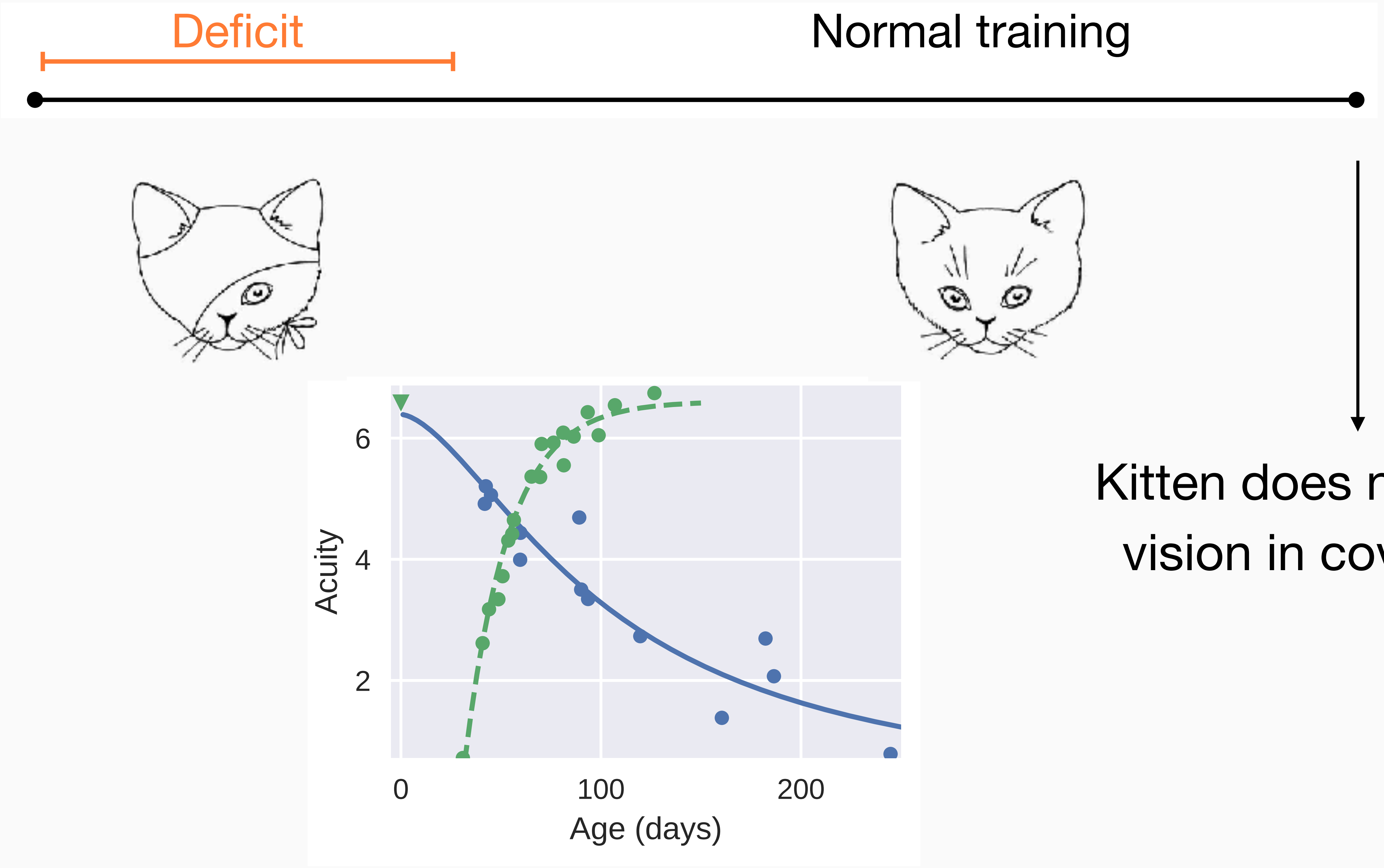
Excursus: Critical Periods for learning

Follow-up: Task reachability. Complexity is physical.

Critical periods

Critical periods: A time-period in early development where sensory deficits can permanently impair the acquisition of a skill

Examples: monocular deprivation, cataracts, imprinting, language acquisition



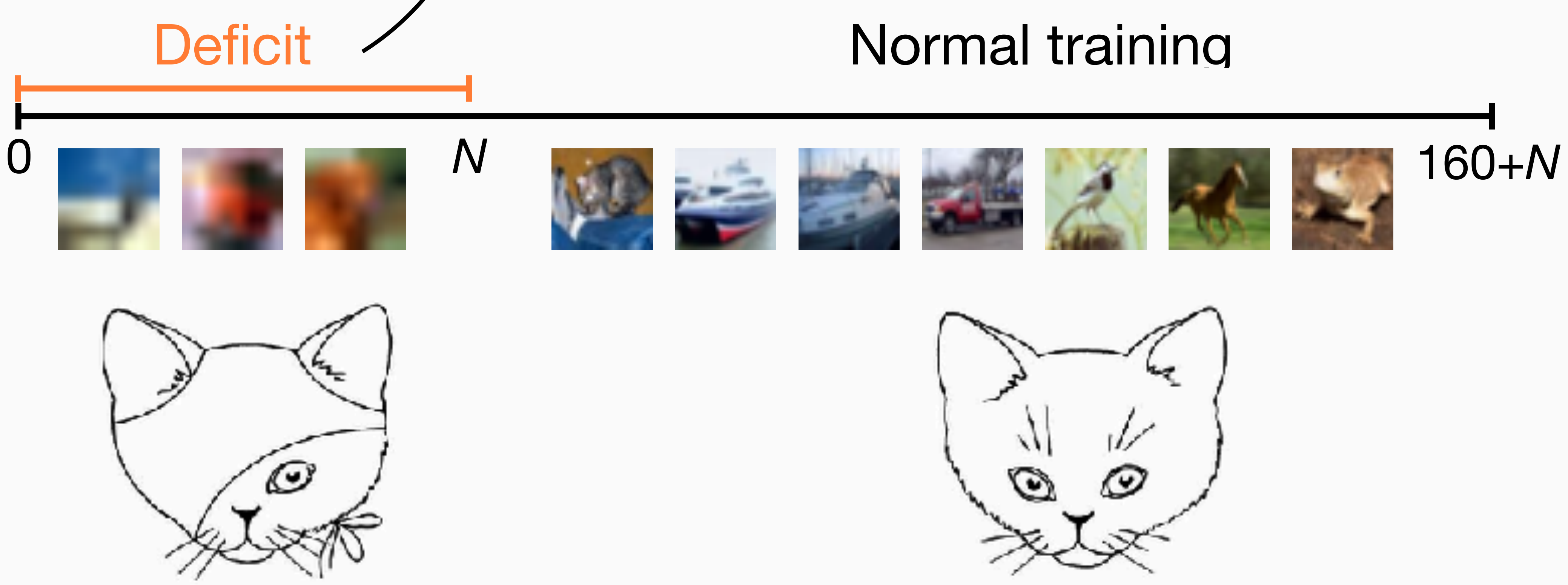
Kitten does not recover vision in covered eye

Hubel and Wiesel

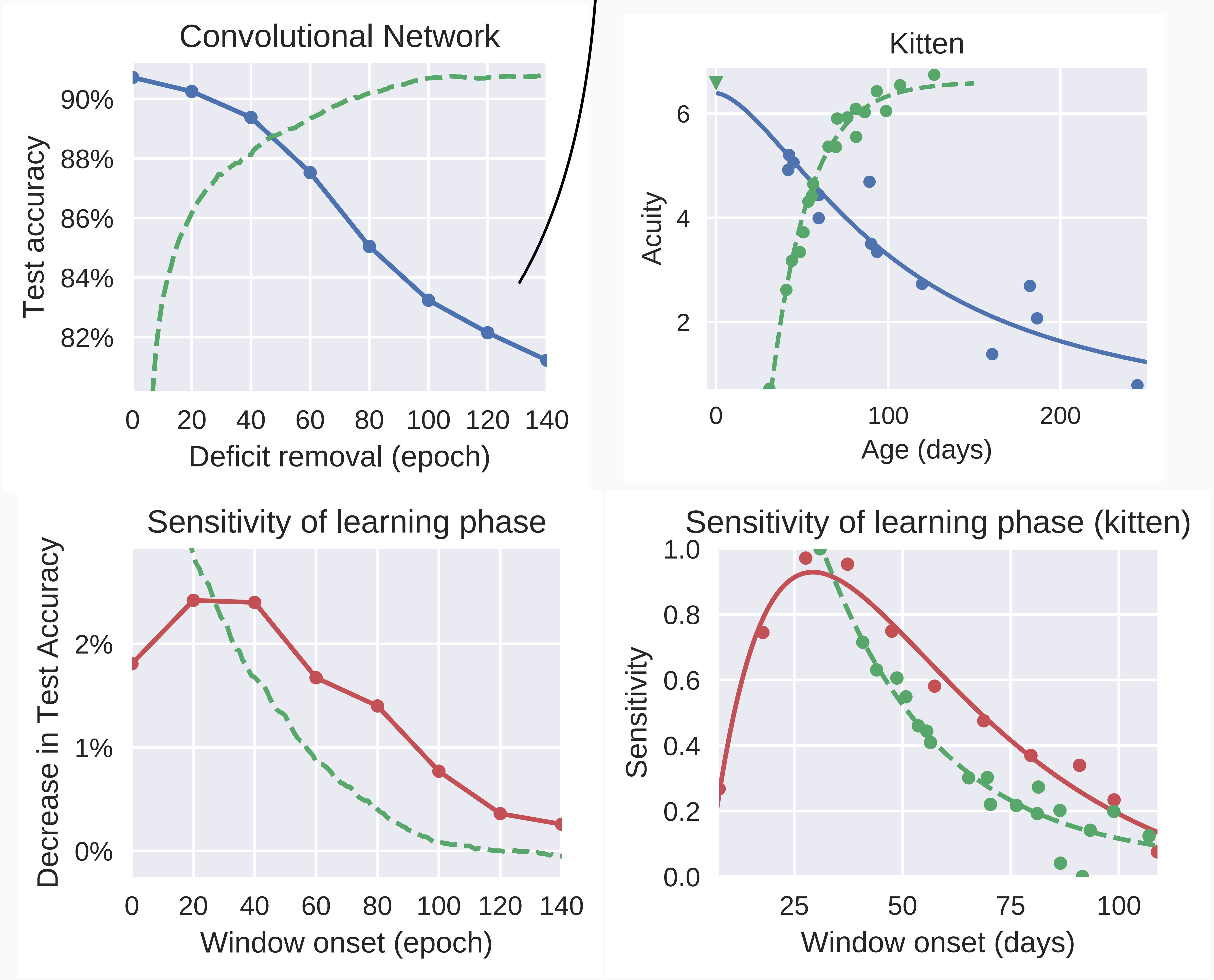
Image from *Cnops et al.*, 2008

Critical periods in Deep Networks

Show network blurred images to simulate cataract



The network does not classify correctly if the deficit is removed to late



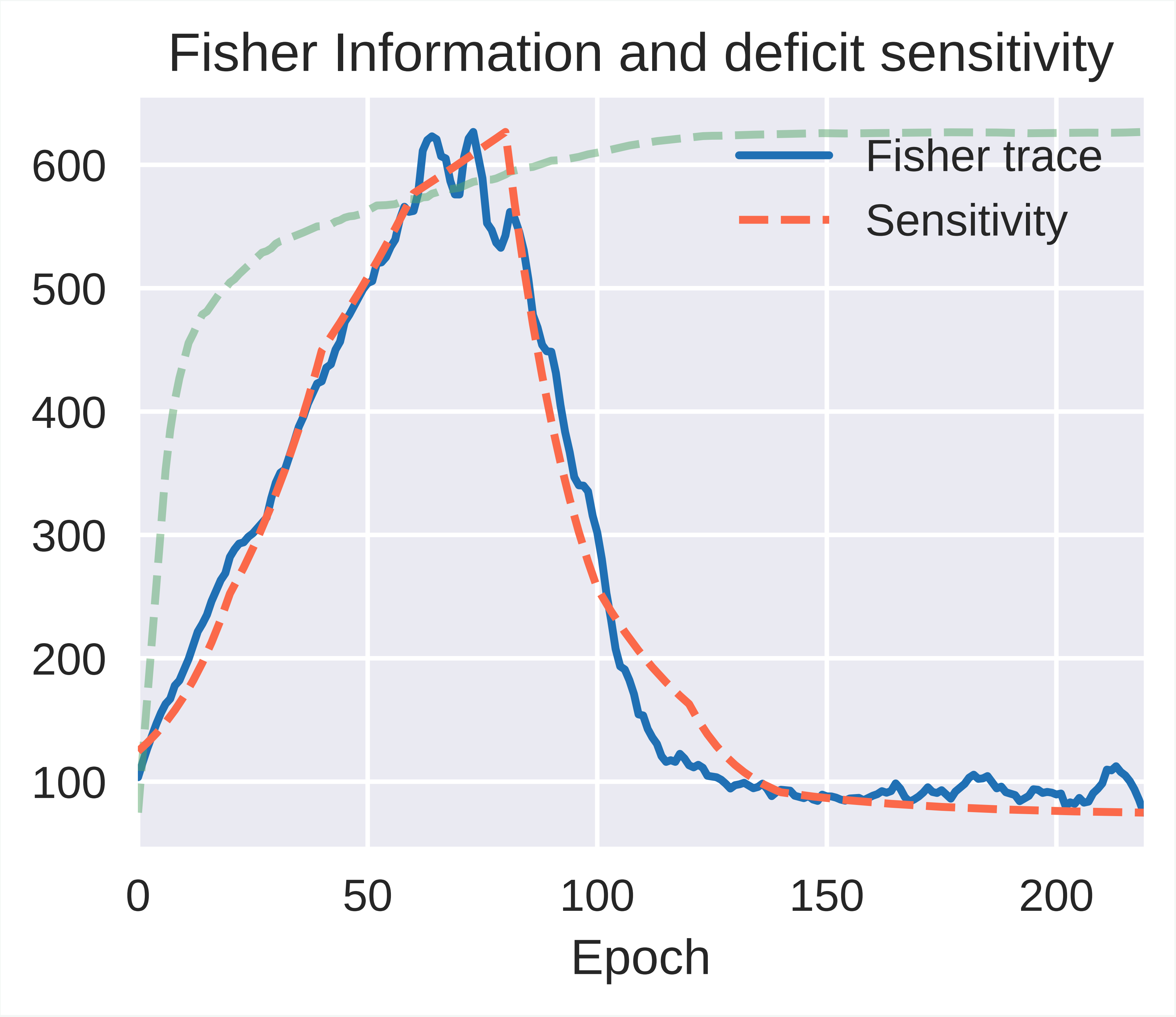
A short deficit at epoch ~40 is enough to permanently damage the network!

Achille, Rovere, Soatto, *Critical Learning Periods in Deep Networks, 2018*

Critical learning periods and Information in Weights

Sensitivity to deficits peaks when network is absorbing information.

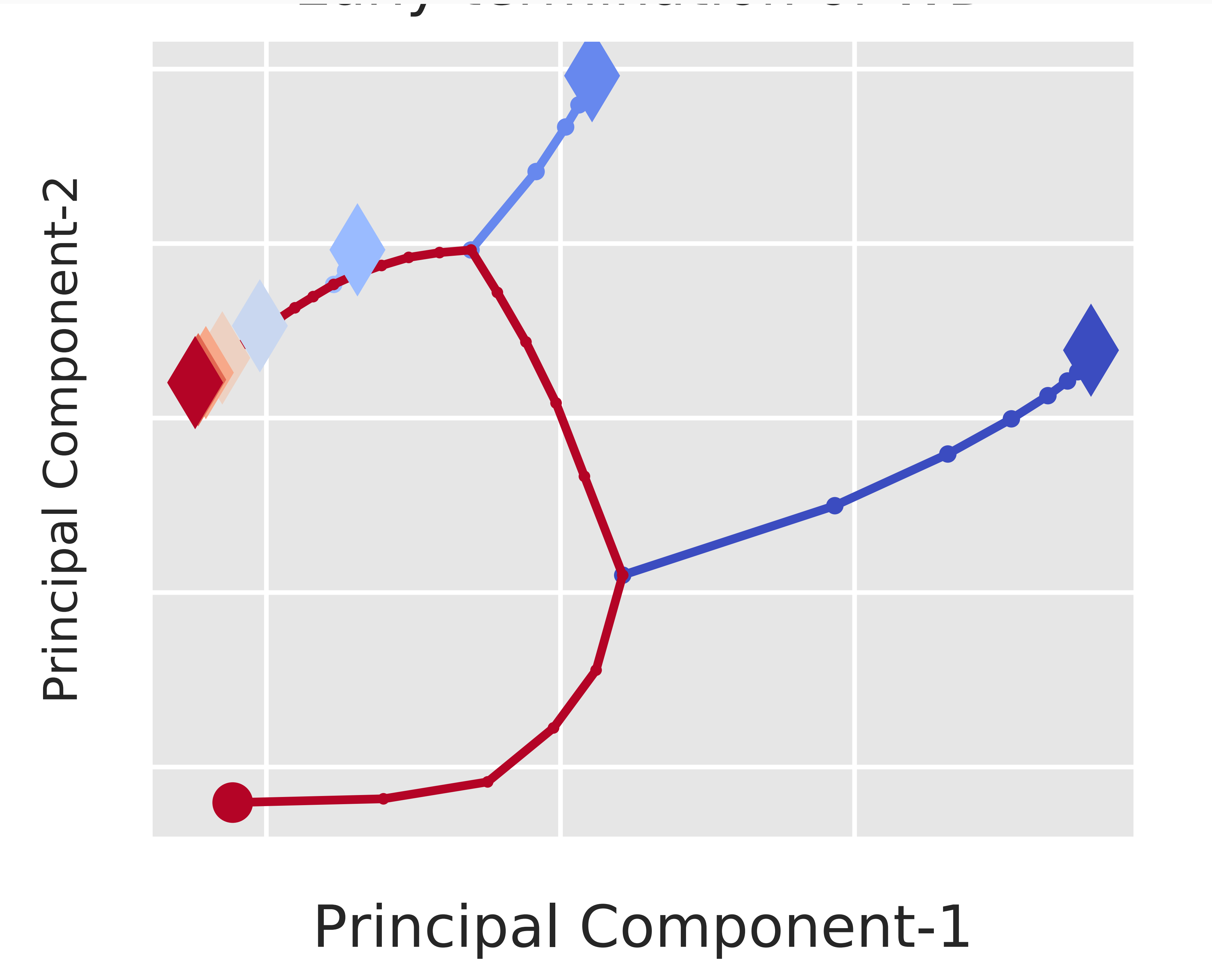
Is minimal when the network is consolidating information.



Achille, Rovere, Soatto, *Critical Learning Periods in Deep Networks, 2018*

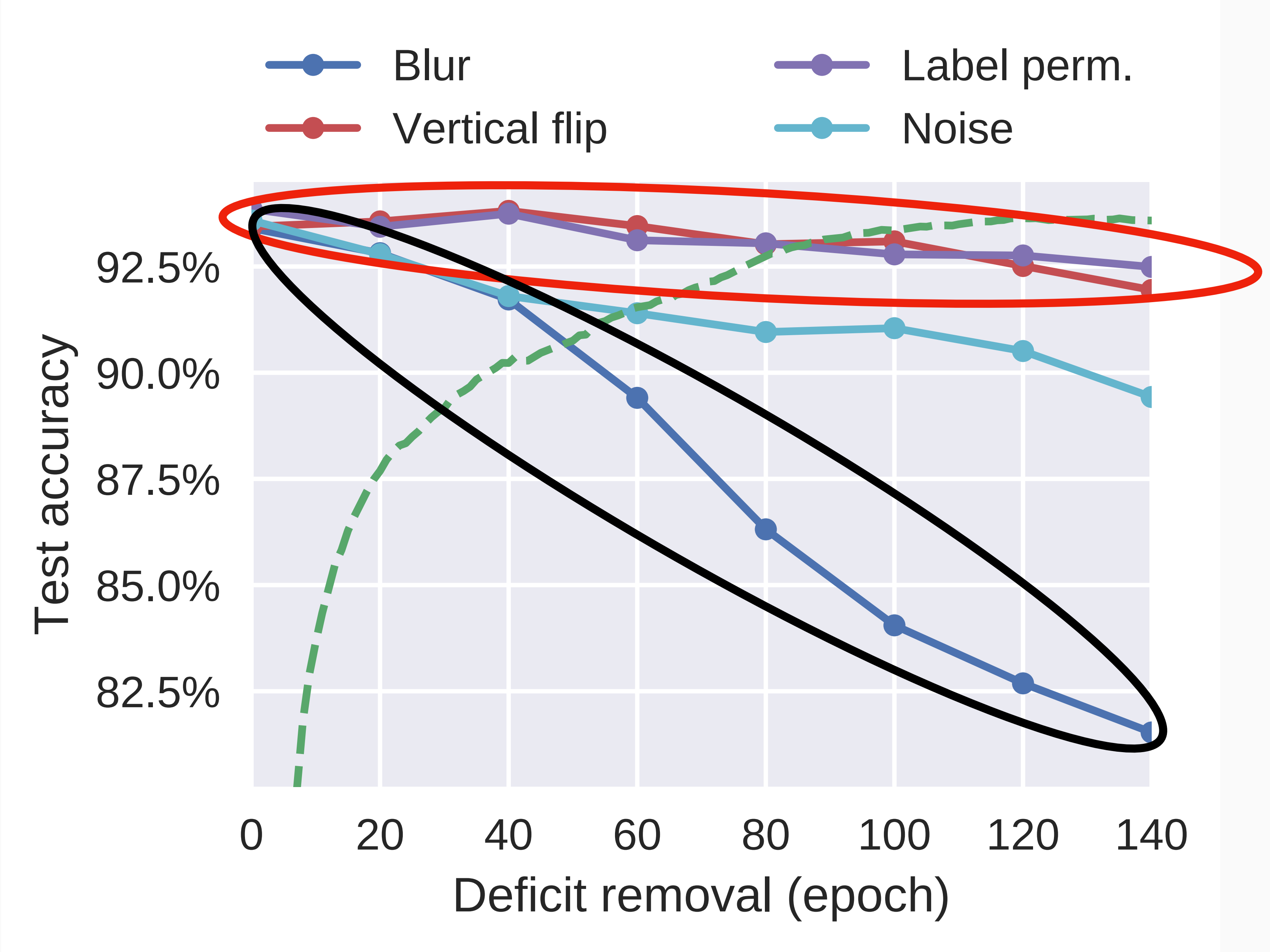
Solutions Move after Critical Periods

The moving is not for the better; does not affect performance.



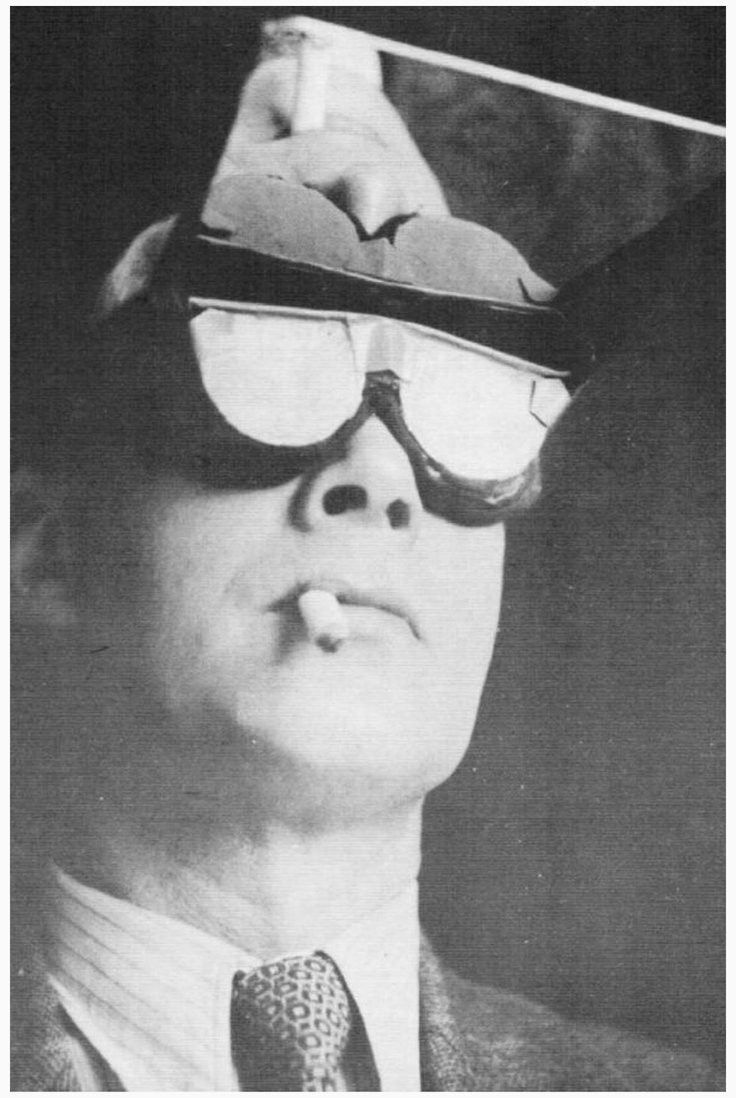
High-level deficits do not have a critical period

Deficits that only change high-level statistics of the data do not show a critical period.



High-level deficits do not exhibit a critical period

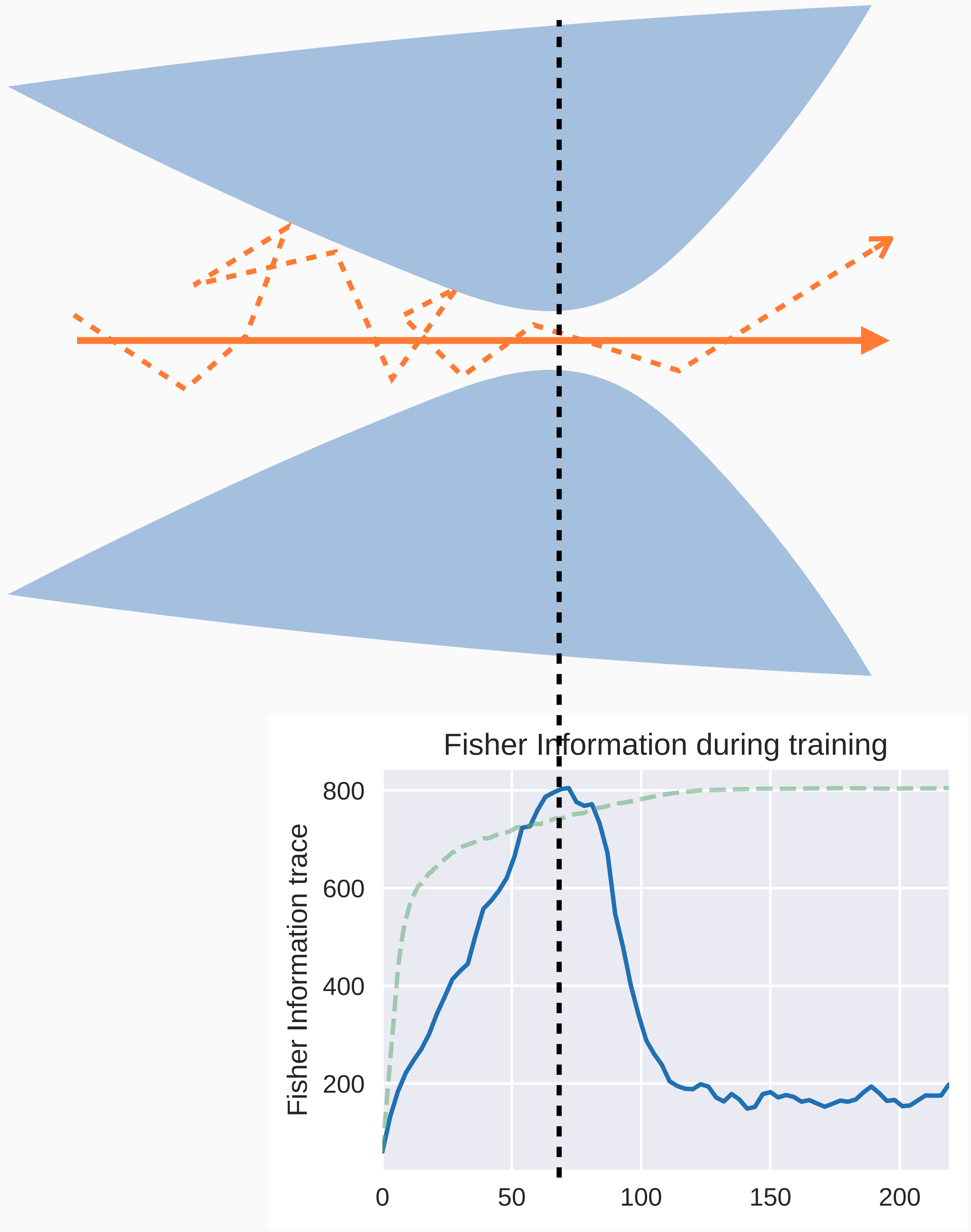
Low-level deficit exhibit a critical period



Achille, Rovere, Soatto, *Critical Learning Periods in Deep Networks, 2018*
Picture from *"The world is upside down" — The Innsbruck Goggle Experiments of Theodor Erismann and Ivo Kohler,* Sachse et al.

Information is physical

How can the Fisher Information affect the learning dynamics?



Idea: When using SGD, the Fisher Information adds a drag term controlled by the batch size

$$V_{\text{eff}} = \underbrace{U}_{\text{Real loss}} + \underbrace{\frac{k}{B}\log|F|}_{\text{Drag term}}$$

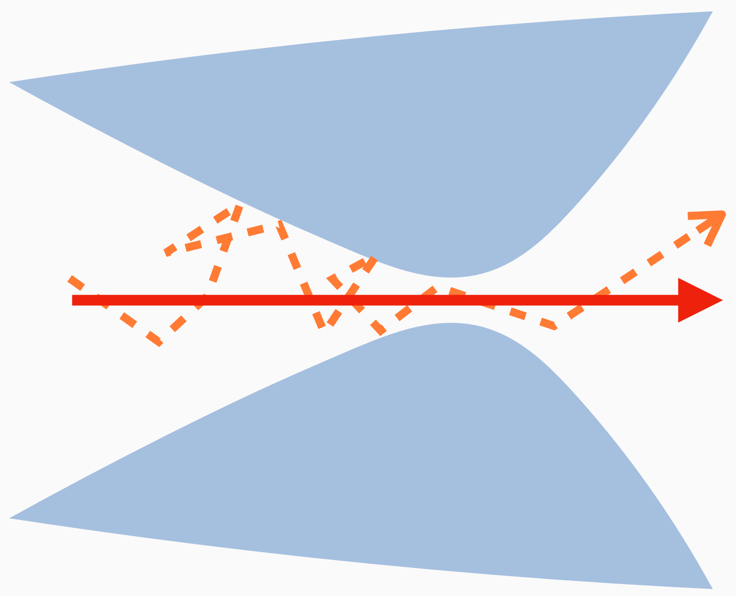
SGD MINIMIZES THE FISHER INFORMATION OF THE WIGHTS (INDUCTIVE BIAS OF SGD)

A path-integral approximation

1) Approximate SGD with gradient descent + white noise. Use MSR formalism to obtain probability of following a path *w(t)*:

$$p(w(t)|w_0, t_0) = e^{\frac{1}{D}\int \mathcal{L}}$$

2) Assume most path are perturbations of distinct “critical” paths:



3) Approximate the loss function quadratically along critical paths, and integrate out the perturbations to find total probability of crossing bottleneck:

$$p(w_f, t_f | w_0, t_0) = e^{-[\ \cdots} \cdots {}^{u(t))dt} du(t)$$

SGD EFFECTIVELY MINIMIZES THE IBL FOR THE WEIGHTS

Static part

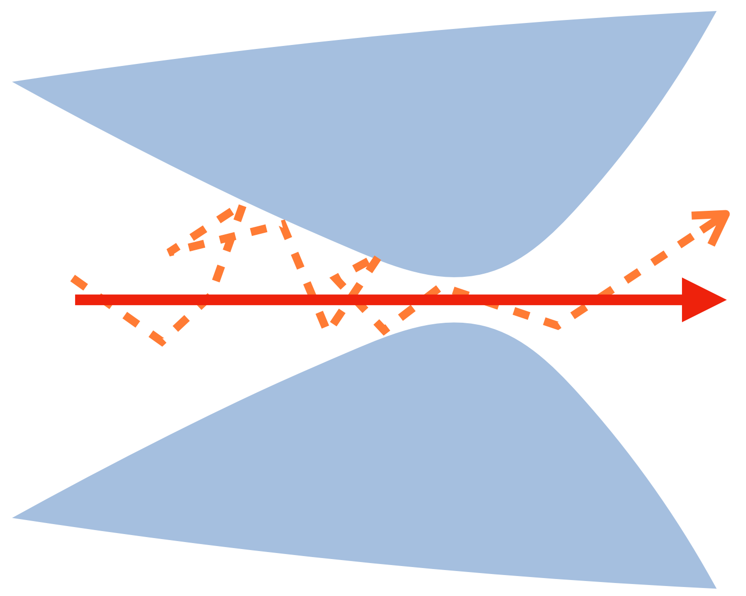
Depends only on the IBL at initial point and final point

Dynamic part

Depends on the existence of likely path between the two

Achille, Mbeng, Soatto, *Dynamics of learning, arXiv 2018*

Path Integral Approximation and Task Reachability



SGD EFFECTIVELY MINIMIZES THE IBL FOR THE WEIGHTS

$$p(w_f, t_f | w_0, t_0) = e^{-\boxed{\Delta \mathcal{L}(w; \mathcal{D})}} \int_{w_0}^{w_f} e^{-\frac{1}{2D} \int_{t_0}^{t_f} \frac{1}{2}\dot{u}(t)^2 + V(u(t)) dt} du(t)$$

Reachability

Static part

Dynamic part

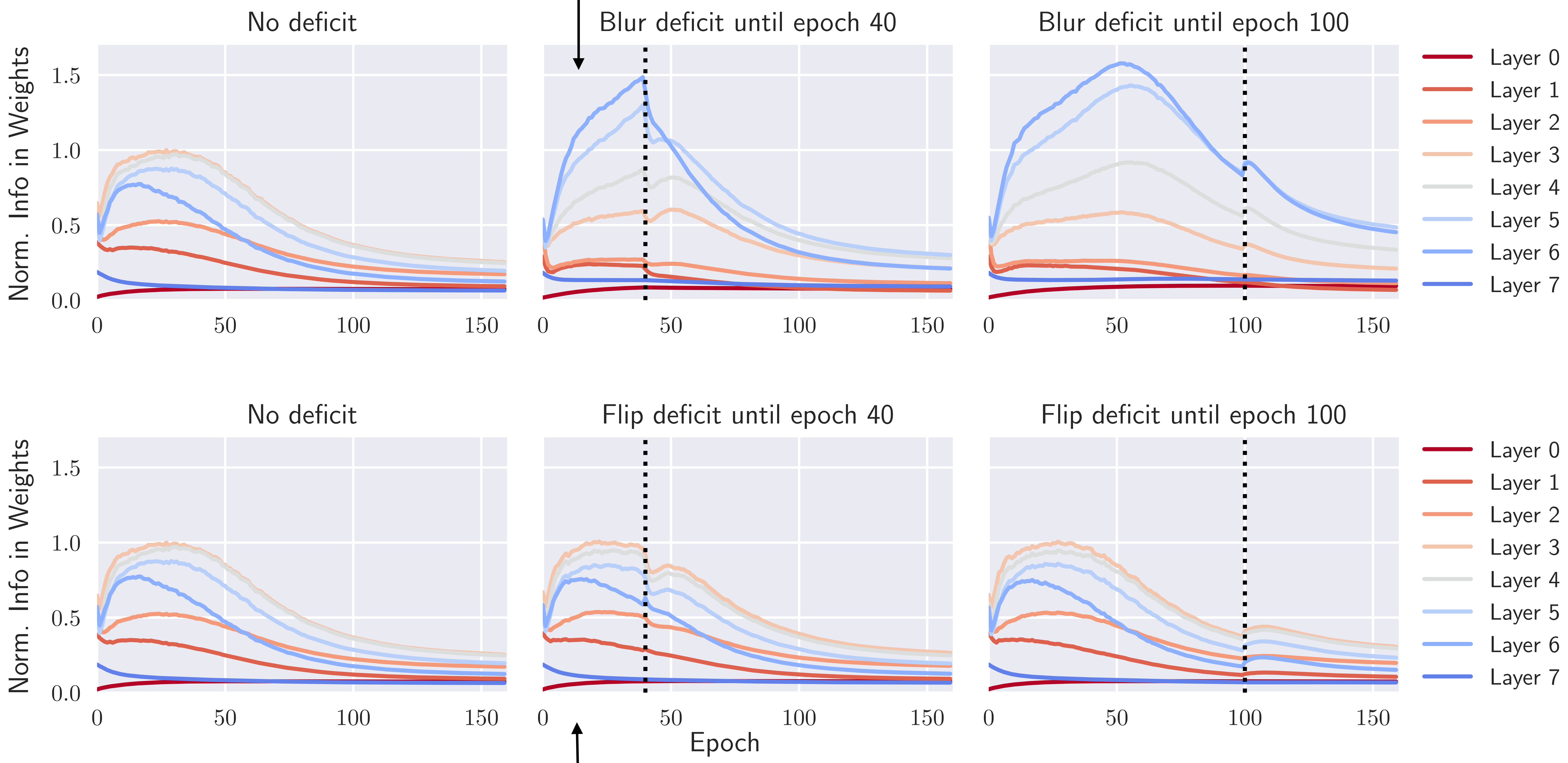
Information Lagrangian

Critical Periods

Achille, Mbeng, Soatto, The Dynamic Distance Between Tasks, NeurIPS Workshop 2018

Information Plasticity in Deep Networks

Introducing a blur deficit
changes layer organization



High-level deficit do not change
layer organization

A., Rovere, Soatto, *Critical Learning Periods in Deep Networks, 2018*

Summary

1. Emergence Theory addresses optimal representations for a given task.

2. Tasks live in a complex space, where “distances” depend not just on the geometry of the residual landscape (static component), but also on the direction of travel (asymmetric distance).

3. Critical Periods expose the importance of the transient of learning; introduced the notion of “Information Plasticity”

4. Learning Dynamics: Tasks may or may not be reachable depending on the dynamics of learning. Dynamic distance between tasks and reachability.

references

- A. Achille & Ss: On the Emergence of Invariance and Disentanglement in Deep Representations, JMLR 2018
- A. Achille & Ss: Information Dropout, PAMI 2018
- A. Achille & Ss: A Separation Principle for Control in the Age of Deep Learning, Annual Reviews, 2018 (also ArXiv)
- A. Achille, et al: Critical Learning Periods in Deep Neural Networks, ICLR 2019
- A. Achille et al: Where is the Information in a Deep Neural Network? ArXiv 2019