Synergy and Symmetry in Deep Learning:

Interactions between the Data, Model, and Inference Algorithm

Lechao Xiao & Jeffrey Pennington
Google Research, Brain Team

Motivation: Curse of Dimensionality (CoD)

- E.g., dim(degree r polynomials in d variables) ~ d^r,

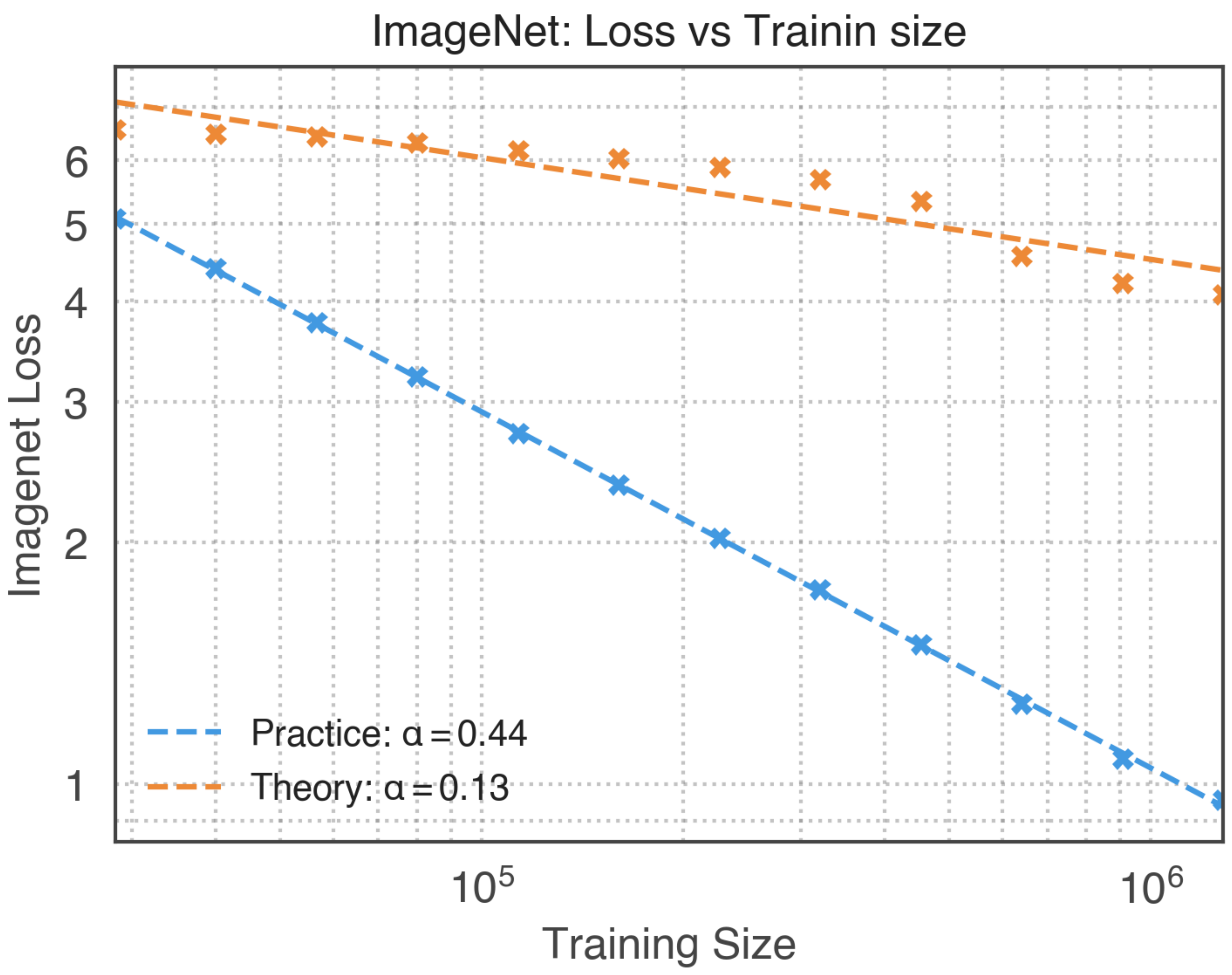
- Quickly become infeasible when d is large in practice (e.g., $d \sim 10^5$ for ImageNet)

- Require a huge number samples to learn!

- Indicate "Poor" scaling law: **loss** $\sim m^{-\alpha}$, with α being tiny (e.g. $\alpha \sim 1/d$)

Neural Networks Can Overcome the CoD. Why?

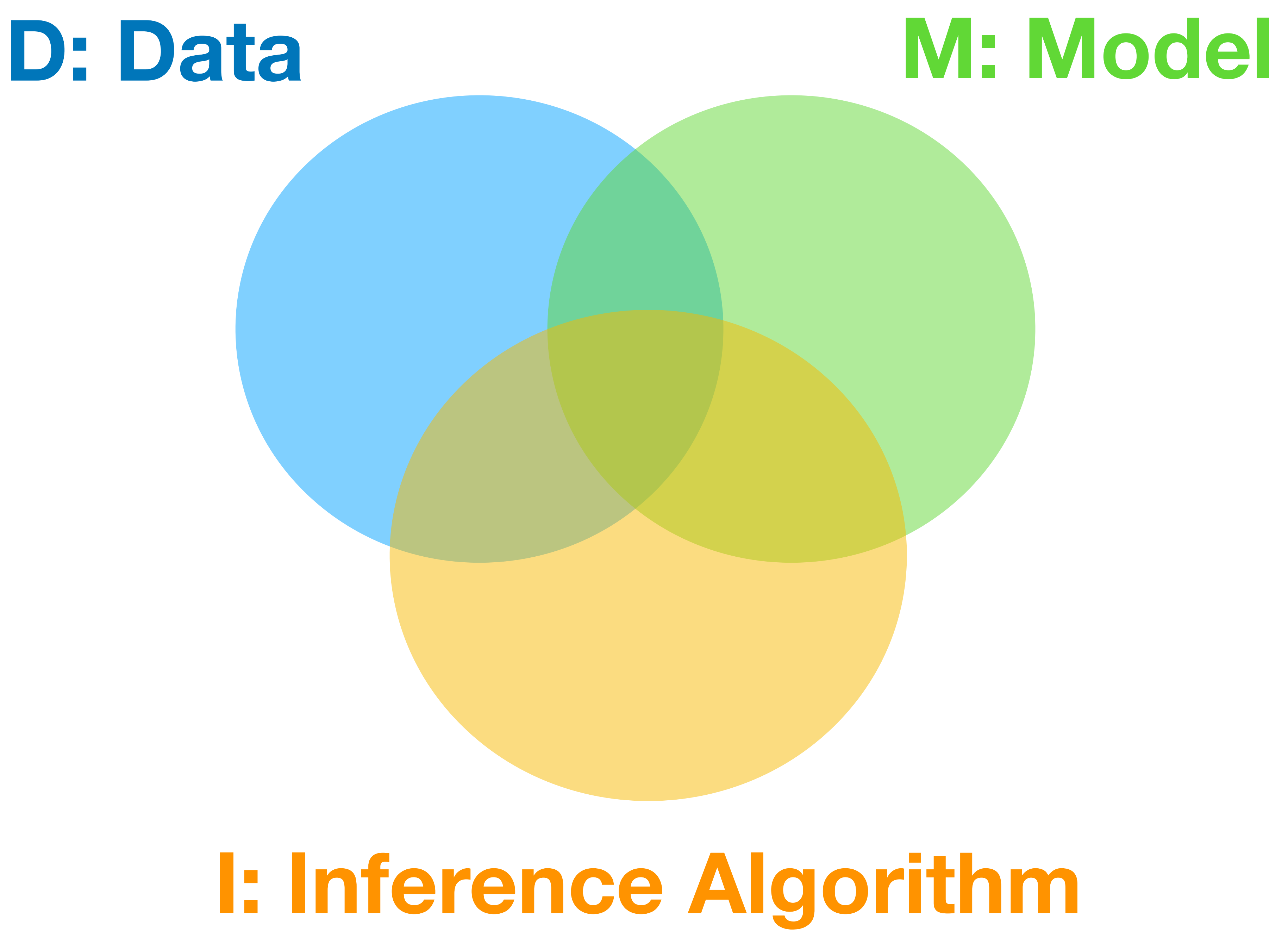
- Good scaling law observed in practice (Blue Curve) rather than
- poor scaling law indicated by theory (Orange Curve)



ImageNet Scaling Plot (ResNet50)

Methodologies

- Consider the Triple (D, M, I) as an integrated system
- Study basic symmetries associated to this system (**algorithmic symmetry**)
- Examine relation between symmetry and performance



Algorithmic Symmetry

- **Algorithmic Symmetry :** invariance of ***the learning procedures*** to certain group transformation of the data, namely, changing the coordinate system of the data.

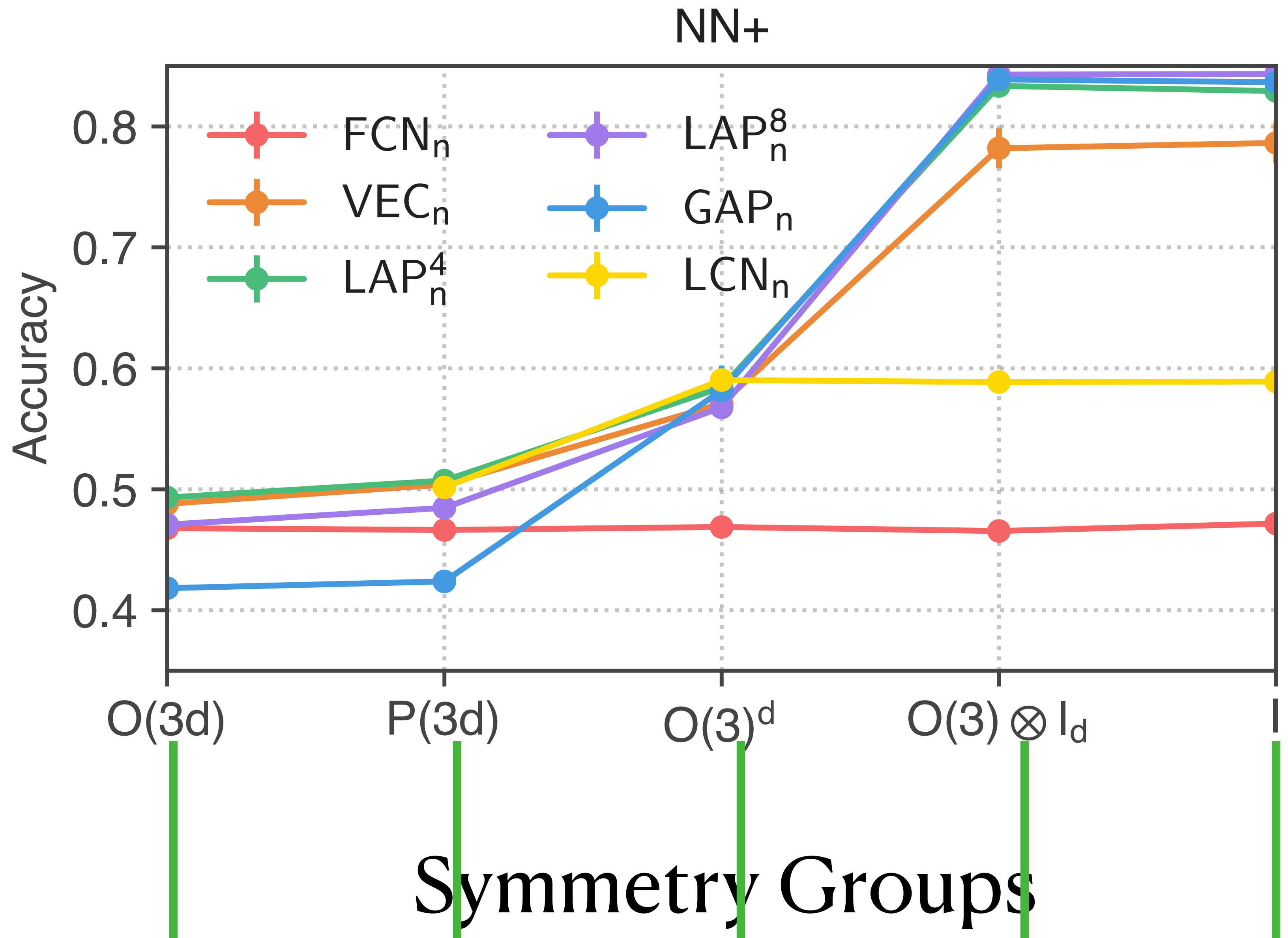
- **Functional Symmetry** $\subsetneq$ **Algorithmic Symmetry**

- **Example**: Kernel regression with an inner product kernel is algorithmic but **not** functionally invariant to rotation because for any rotation τ and all inputs x, x'

$$K(\tau x, \tau x') = K(x, x') \quad \text{but} \quad K(x, \tau x') \neq K(x, x')$$

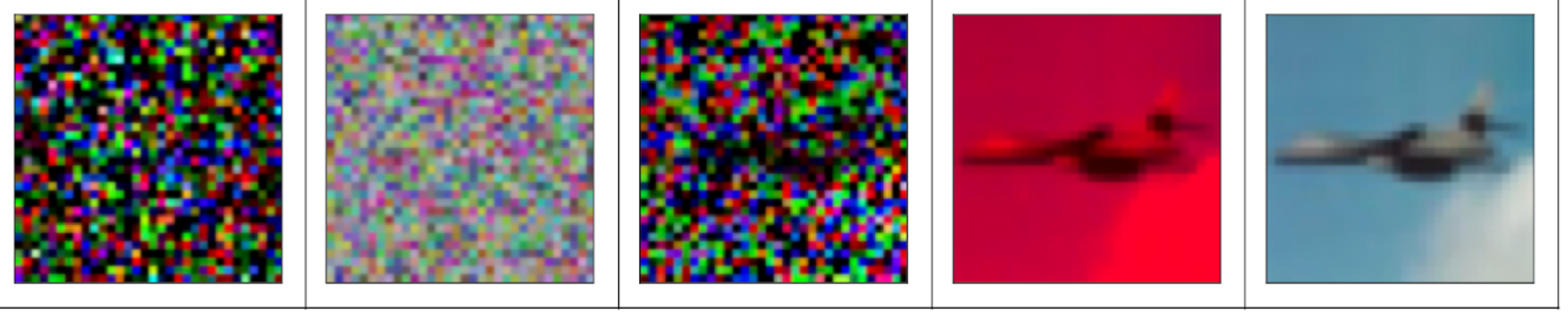
Better Architectures Break Spurious Symmetries

- Identify basic symmetries associated to
  1. **Baseline**: Fully-connected networks (**FCN**)
  2. **+Locality**: Locally-connected Networks (**LCN**)
  3. **+Weight-sharing**: Convolutional networks with a vectorization readout layer (**VEC**)
  4. **+Translation-invariance**: Convolutional networks with a global average pooling (**GAP**)
- Better architectures break spurious symmetries and lead to better performance.



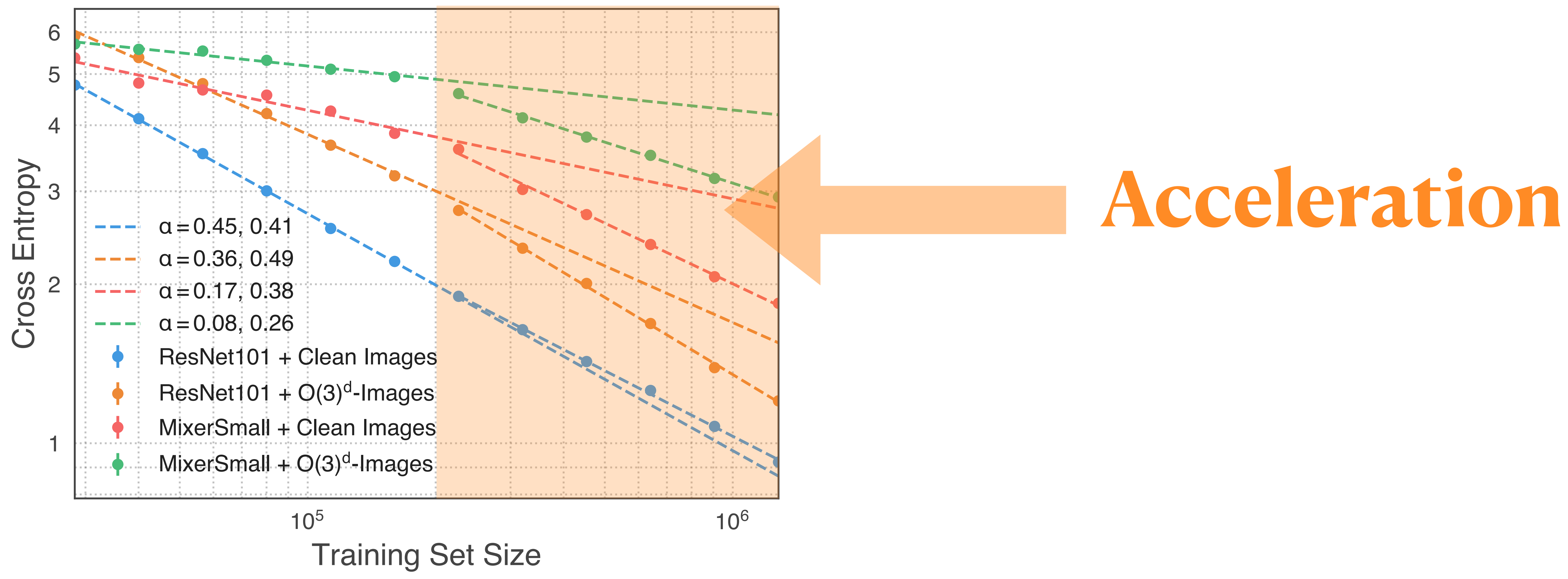
Images for training/test

Models	$FCN_{n/\infty}$ (iid Gaussian)	FCN_n (iid Non-Gaussian)	VEC_∞/$LCN_{n/\infty}$ (iid Gaussian)	VEC_n/$GAP_{n/\infty}$ (iid Gaussian)	
Symmetry Groups	$O(3d)$	$P(3d)$	$O(3)^d$	$O(3) \otimes I_d$	I_{3d}
Rotated Images					



Data Improves Data Efficiency (DIDE)

- With more data, models can overcome spurious symmetries, improving scaling law

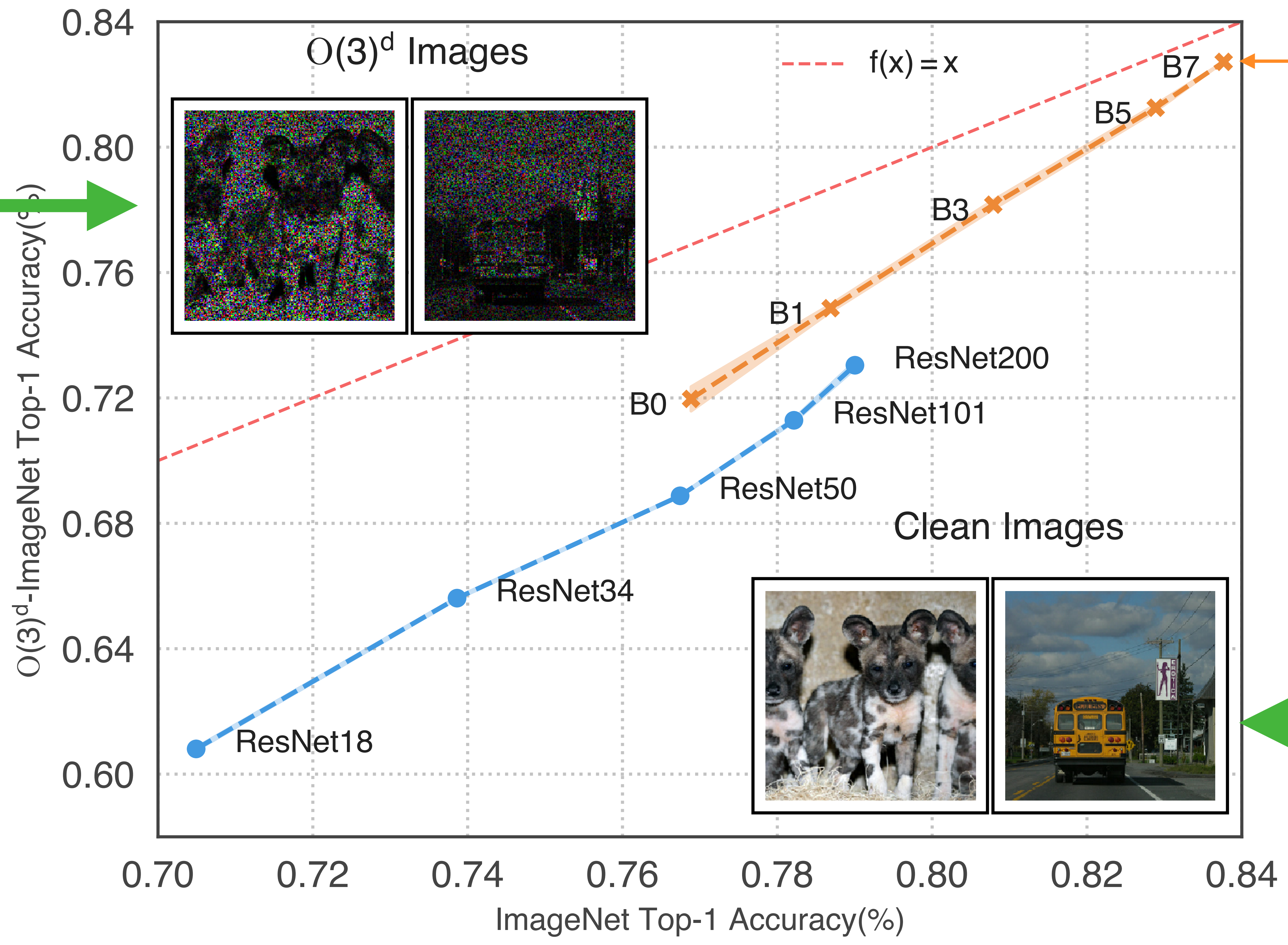


Change of Scaling

Data Improves Data Efficiency (DIDE)

- Larger models (ResNets, EfficientNets) exhibit even more impressive power

Use **Rotated Images** as training/test sets



Using **Rotated Images,** EfficientNet B7 obtains on par performance

Use **Original Images** as training/test sets

Other Contributions

- Performance is degraded in the same way when applying spurious symmetries to the **model** or the **data**.

- Spurious symmetries eliminate the benefits of SGD

- Finite-width VEC breaks some spurious symmetries from infinite-width networks, leading to better performance.

Conclusion

To understand deep learning, we need to understand the interactions between

(**Data**, **Model, Inference algorithm**)