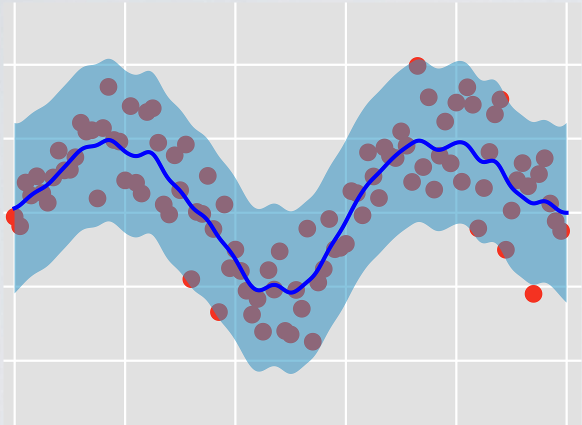
BRIDGING THE GAP BETWEEN DEEP LEARNING THEORY AND PRACTICE

Micah Goldblum

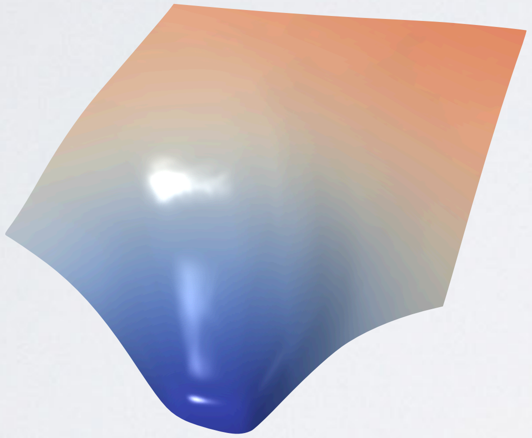
NEW YORK UNIVERSITY

Bayesian ML



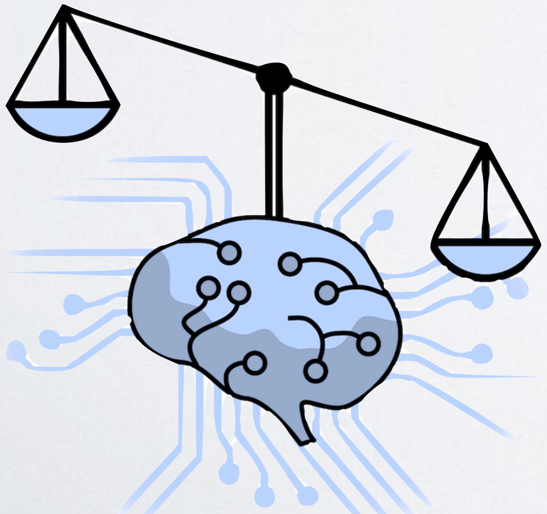
AI Security and Privacy

Generalization Theory

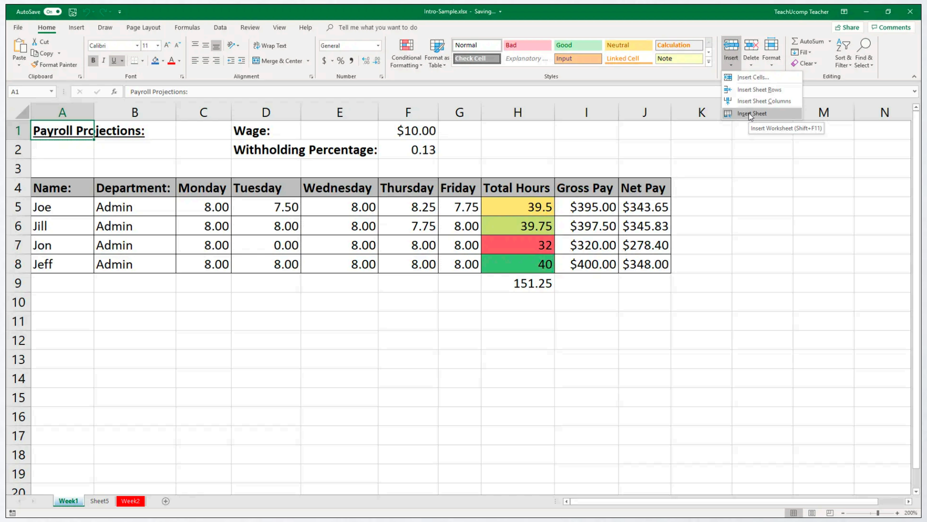


Generative Modeling

Algorithmic Fairness



ML for Tabular Data

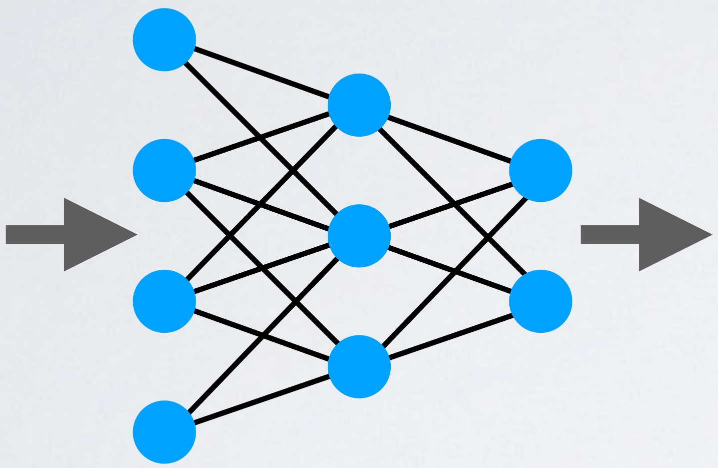


Payroll Projections:			**Wage:**		$10.00				
			Withholding Percentage:		0.13				
Name:	**Department:**	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Total Hours**	**Gross Pay**	**Net Pay**
Joe	Admin	8.00	7.50	8.00	8.25	7.75	39.5	$395.00	$343.65
Jill	Admin	8.00	8.00	8.00	7.75	8.00	39.75	$397.50	$345.83
Jon	Admin	8.00	0.00	8.00	8.00	8.00	32	$320.00	$278.40
Jeff	Admin	8.00	8.00	8.00	8.00	8.00	40	$400.00	$348.00
							151.25		

What is generalization?

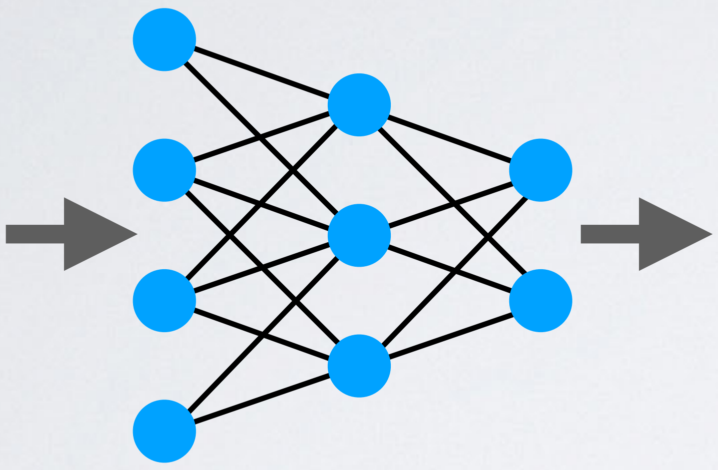
What is generalization?

Flexible model



What is generalization?

Flexible model

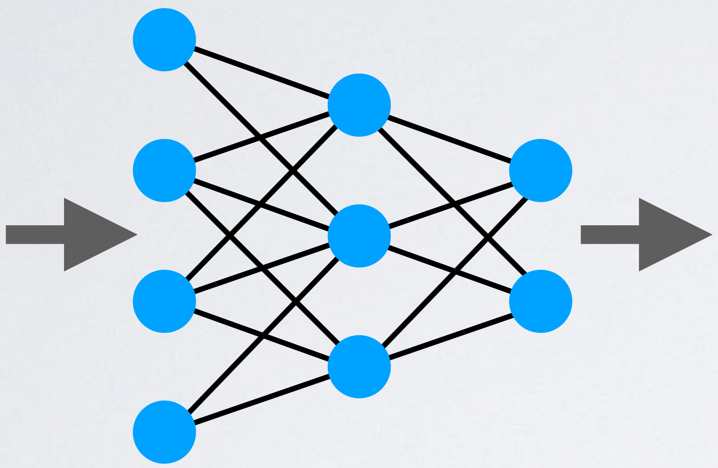


Training loss

$$L(w) = \sum_i \|f(\mathbf{x}_i; w) - y_i\|^2$$

What is generalization?

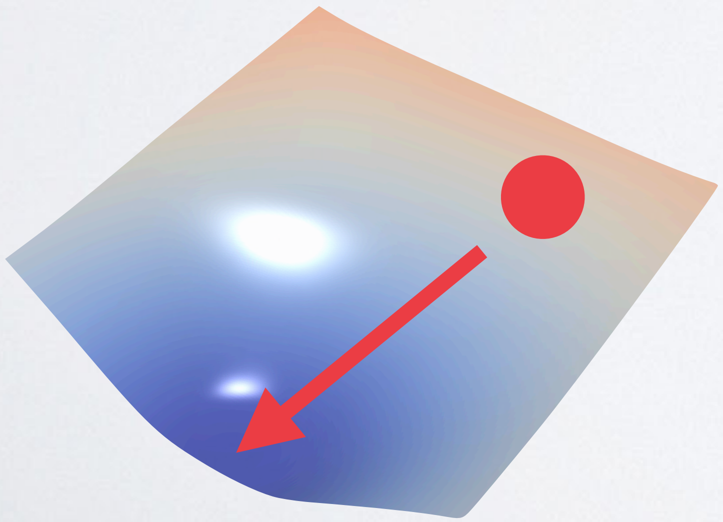
Flexible model



Training loss

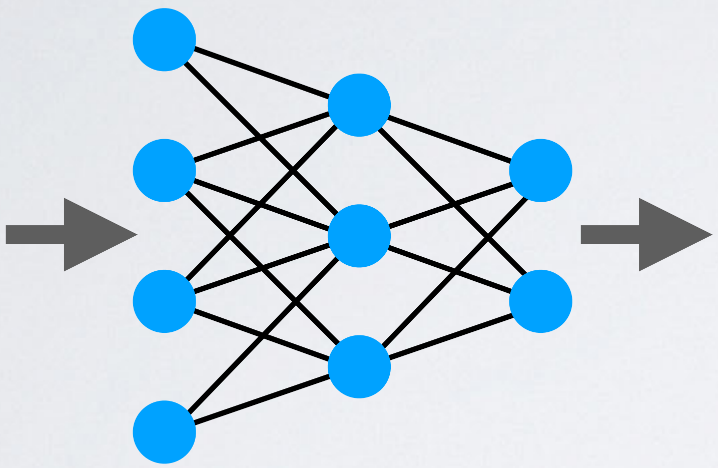
$$L(w) = \sum_i \|f(\mathbf{x}_i; w) - y_i\|^2$$

Minimize training loss



What is generalization?

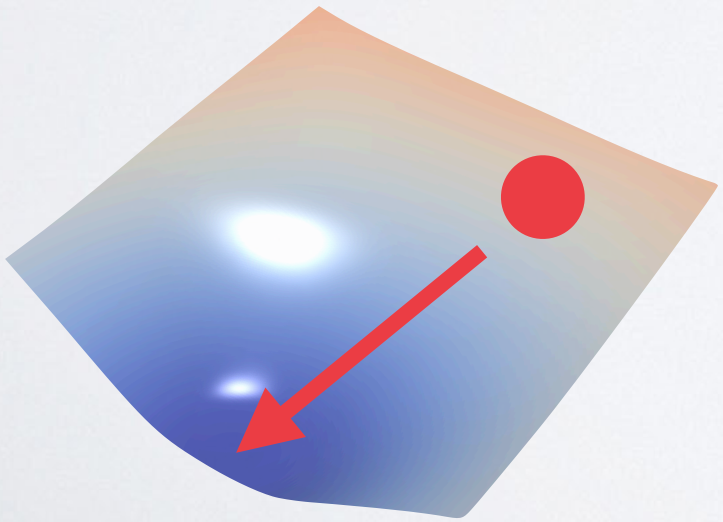
Flexible model



Training loss

$$L(w) = \sum_i \|f(\mathbf{x}_i; w) - y_i\|^2$$

Minimize training loss



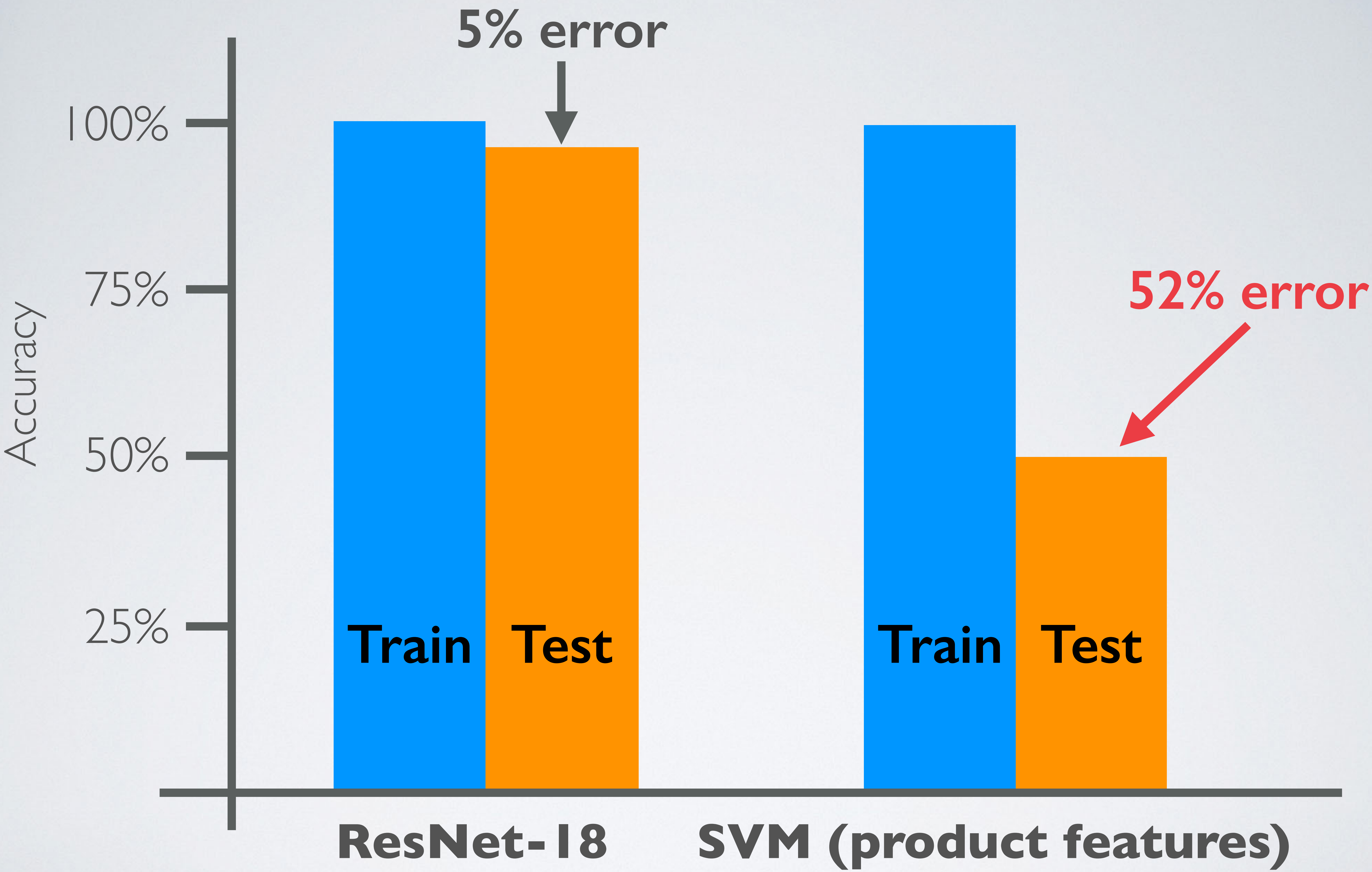
Test accuracy?

CIFAR-10

 # ResNet params ≈ # SVM params

CIFAR-10

 # ResNet params ≈ # SVM params



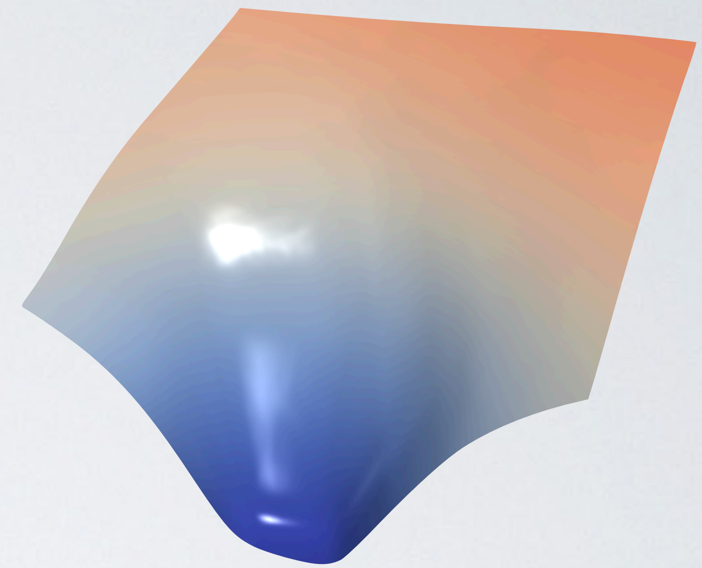
CIFAR-10

ResNet params ≈ # SVM params

Why do neural networks work?

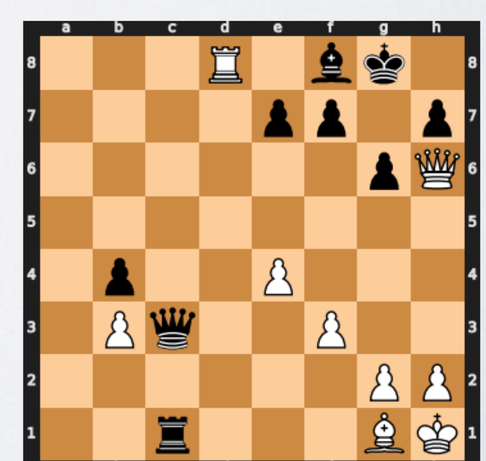
Why do neural networks work?

What are the properties of good minima and why do optimizers find them?



Theories that predict generalization

Observing generalization in reasoning problems



Are all neural network minima good?

Are all neural network minima good?

1. **Suboptimal local minima**

Truth or Backpropaganda, ICLR '20

Are all neural network minima good?

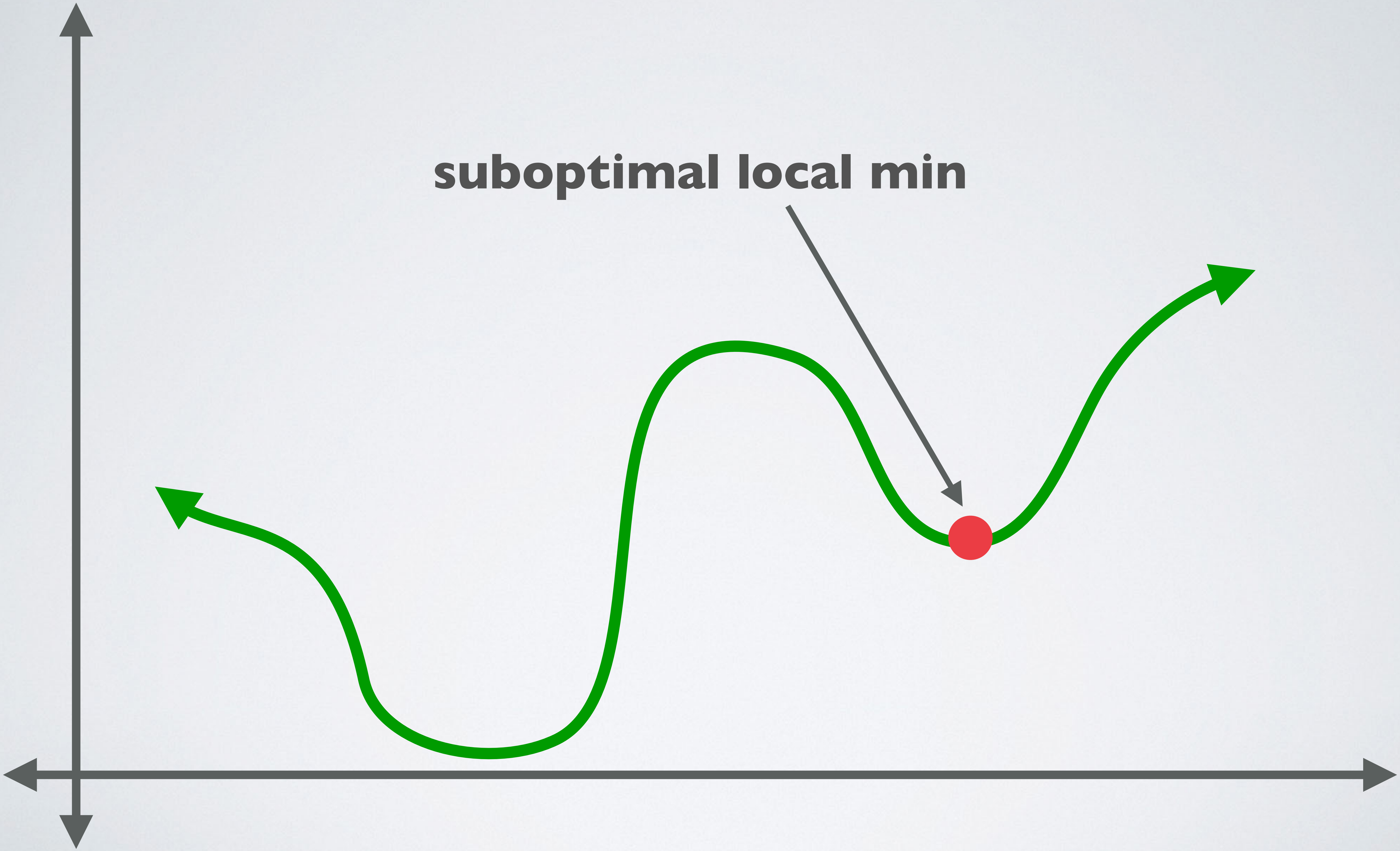
1. **Suboptimal local minima**

Truth or Backpropaganda, ICLR '20

2. **Global minima that generalize poorly**

Understanding generalization through visualizations, Under Review

Are all neural network minima good?



Truth or Backpropaganda, ICLR '20

Suboptimal Local Minima

Assumptions:

Suboptimal Local Minima

Assumptions:

- Continuous loss function $\mathscr{L}$

Suboptimal Local Minima

Assumptions:

- Continuous loss function $\mathscr{L}$
- MLP with ReLUs, minimum features per layer m

Suboptimal Local Minima

Assumptions:

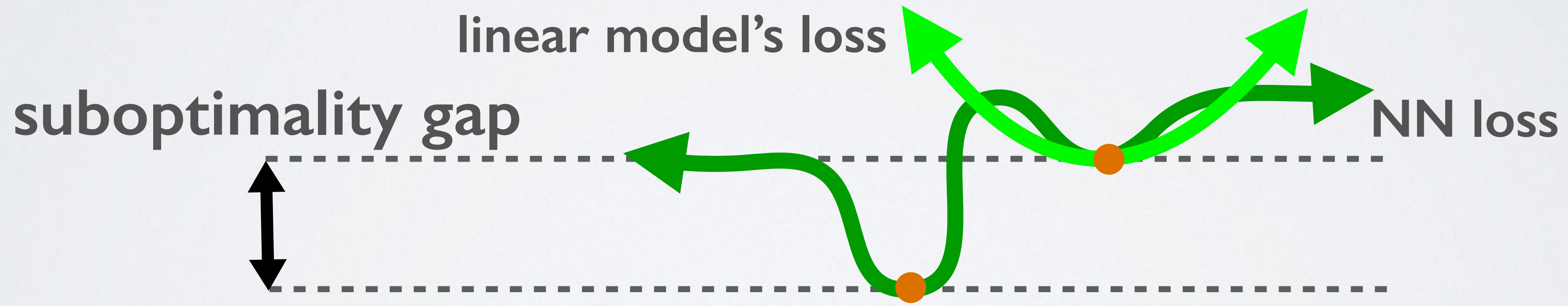
- Continuous loss function $\mathscr{L}$
- MLP with ReLUs, minimum features per layer m
- Linear model with rank$(W) \leq m$

Suboptimal Local Minima

Assumptions:

- Continuous loss function $\mathscr{L}$
- MLP with ReLUs, minimum features per layer m
- Linear model with rank$(W) \leq m$

Theorem (informal): if the NN can achieve lower training loss than the linear model, it has a suboptimal local minimum.



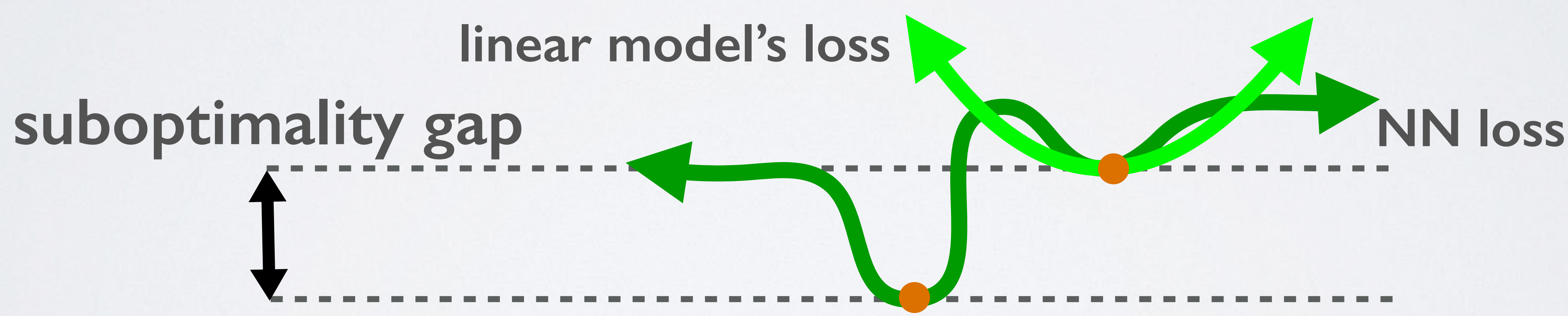
Truth or Backpropaganda, ICLR '20

Suboptimal Local Minima

Assumptions:

- Continuous loss function $\mathscr{L}$
- MLP with ReLUs, minimum features per layer m
- Linear model with rank$(W) \leq m$

Theorem (informal): if the NN can achieve lower training loss than the linear model, it has a suboptimal local minimum.



Extensions:

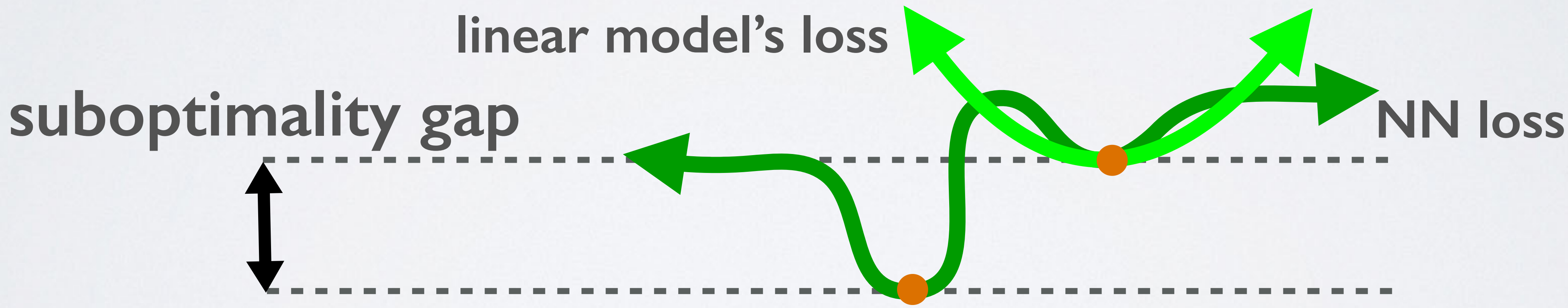
Truth or Backpropaganda, ICLR '20

Suboptimal Local Minima

Assumptions:

- Continuous loss function $\mathscr{L}$
- MLP with ReLUs, minimum features per layer m
- Linear model with rank$(W) \leq m$

Theorem (informal): if the NN can achieve lower training loss than the linear model, it has a suboptimal local minimum.



Extensions:

- Convolutional networks

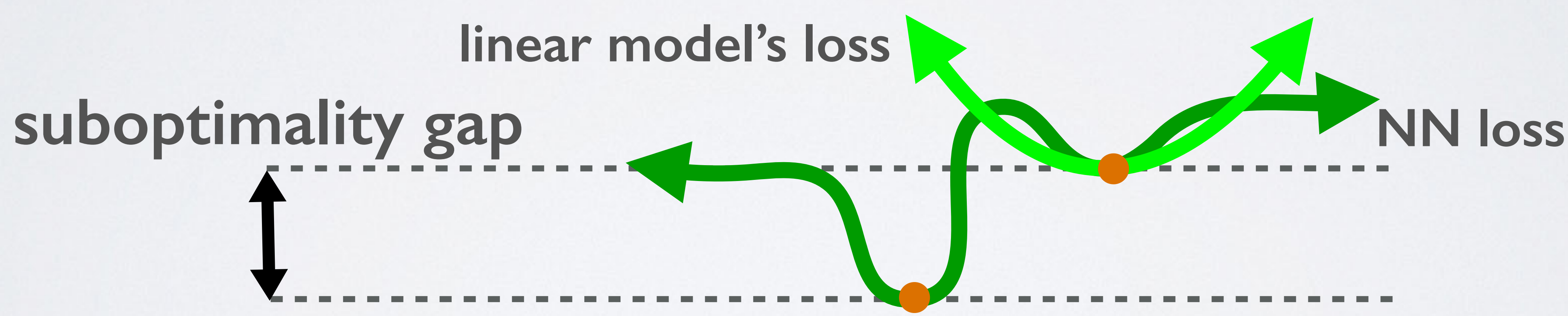
Truth or Backpropaganda, ICLR '20

Suboptimal Local Minima

Assumptions:

- Continuous loss function $\mathscr{L}$
- MLP with ReLUs, minimum features per layer m
- Linear model with rank$(W) \leq m$

Theorem (informal): if the NN can achieve lower training loss than the linear model, it has a suboptimal local minimum.



Extensions:

- Convolutional networks
- Replace linear models with smaller neural nets

Truth or Backpropaganda, ICLR '20

Are all global minima good?

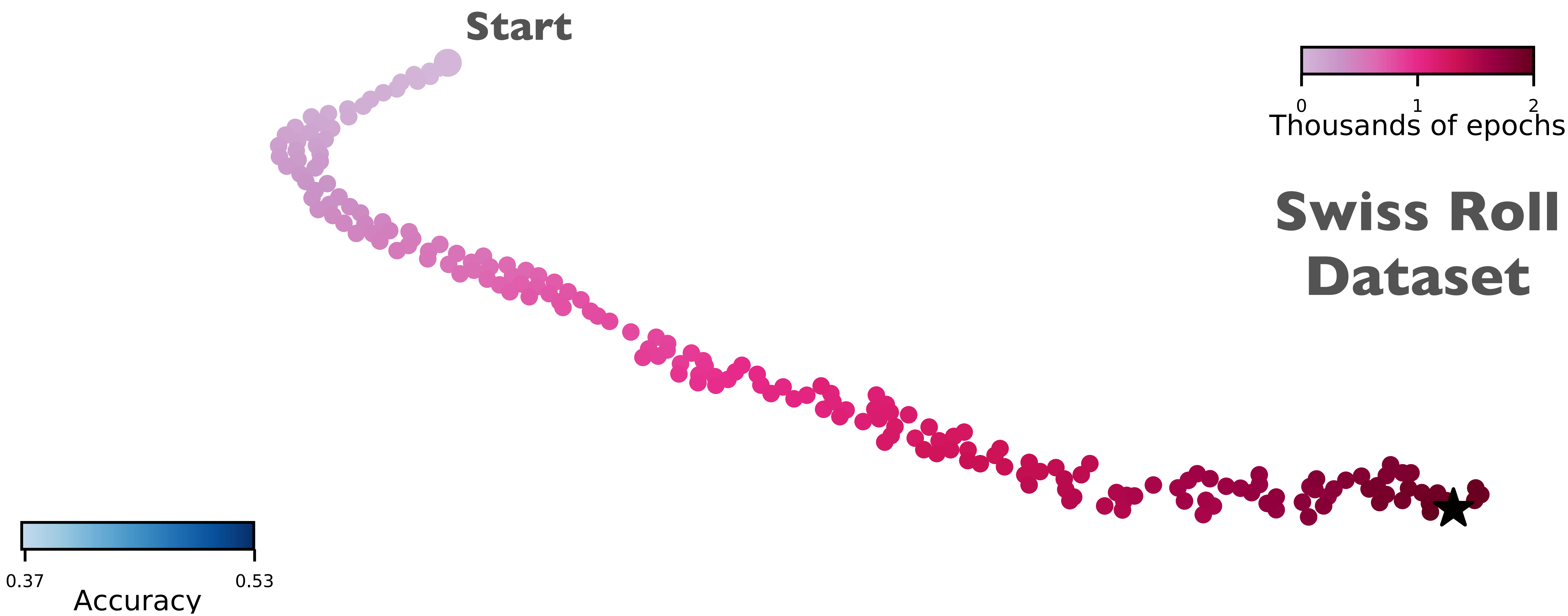
Global minima that generalize poorly

Understanding generalization through visualizations, Under Review

Are all global minima good?

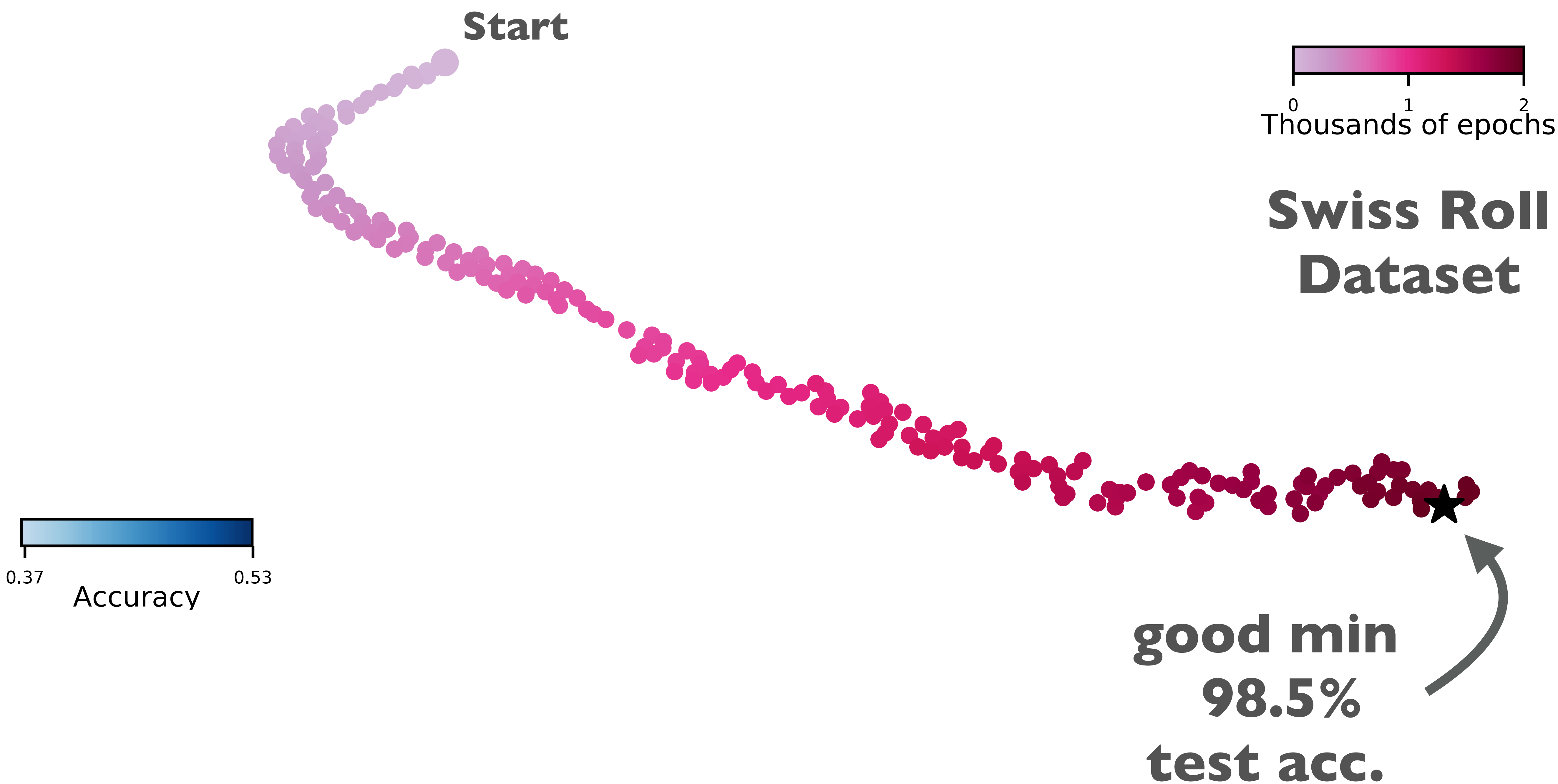
Understanding generalization through visualizations, Under Review

Are all global minima good?



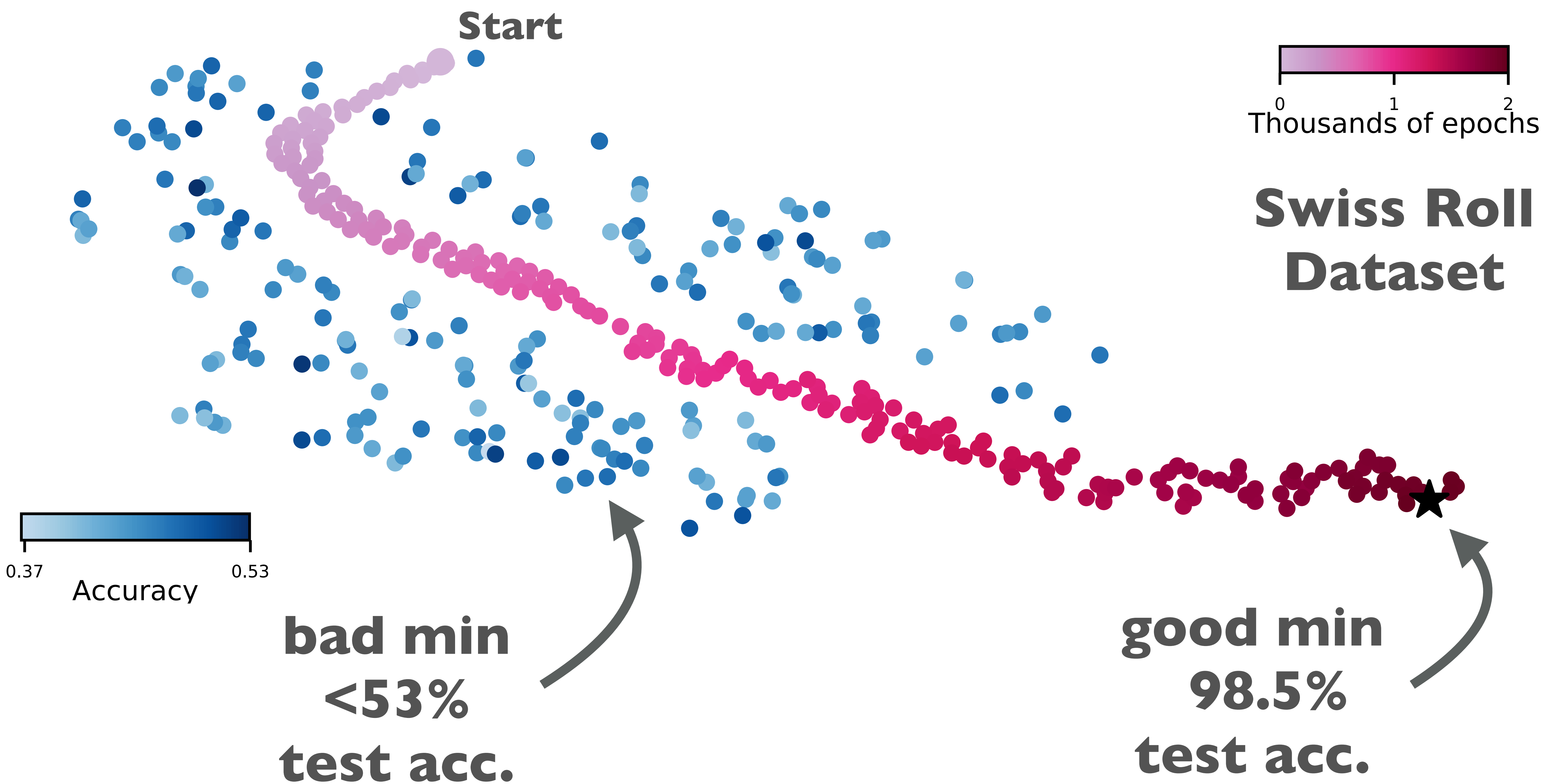
Understanding generalization through visualizations, Under Review

Are all global minima good?



Understanding generalization through visualizations, Under Review

Are all global minima good?



Understanding generalization through visualizations, Under Review

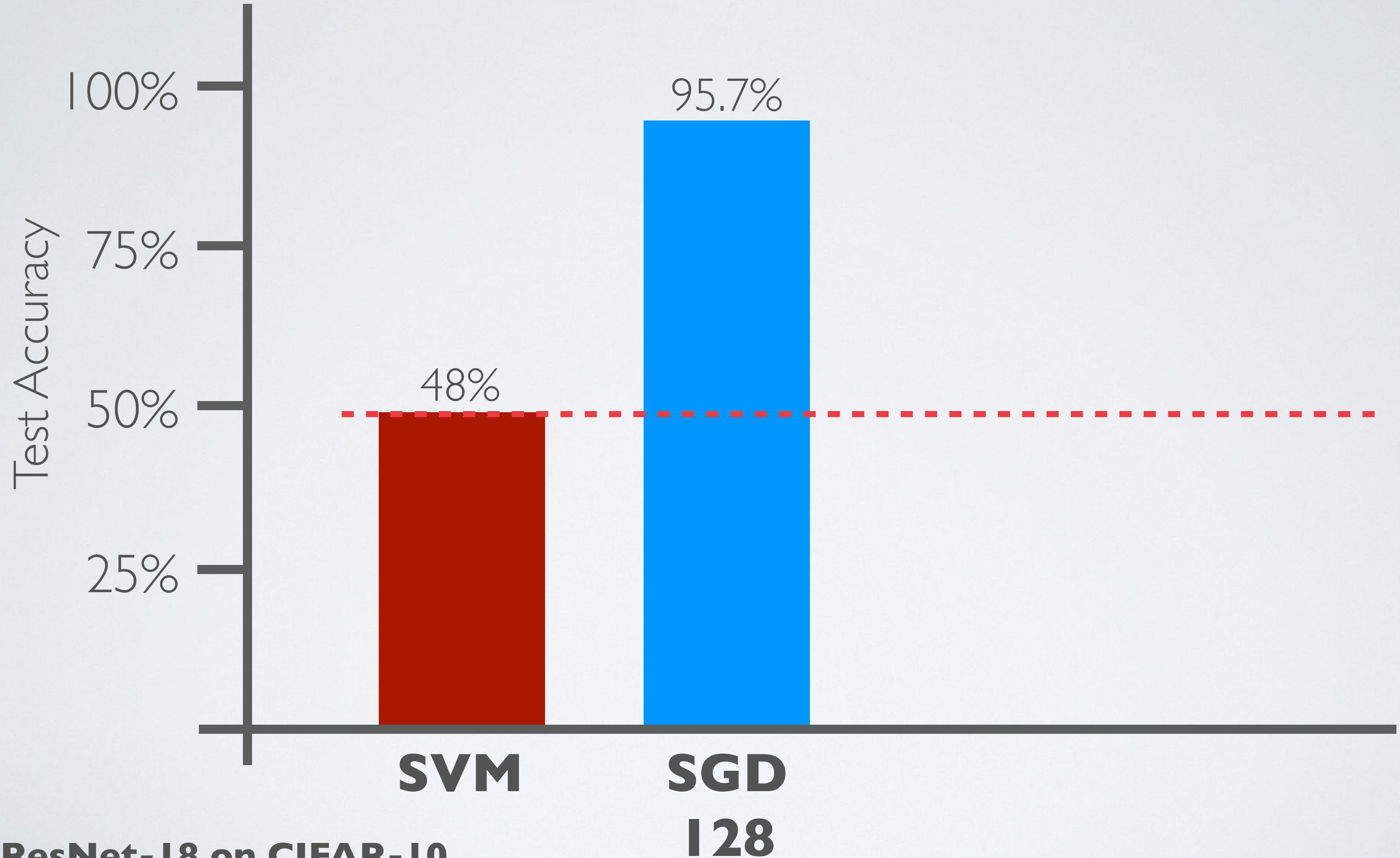
So why do we find good minima?

The optimizer?

Stochastic training is not necessary for generalization, ICLR '22

Gradient-based optimization is not necessary for generalization, ICLR '23

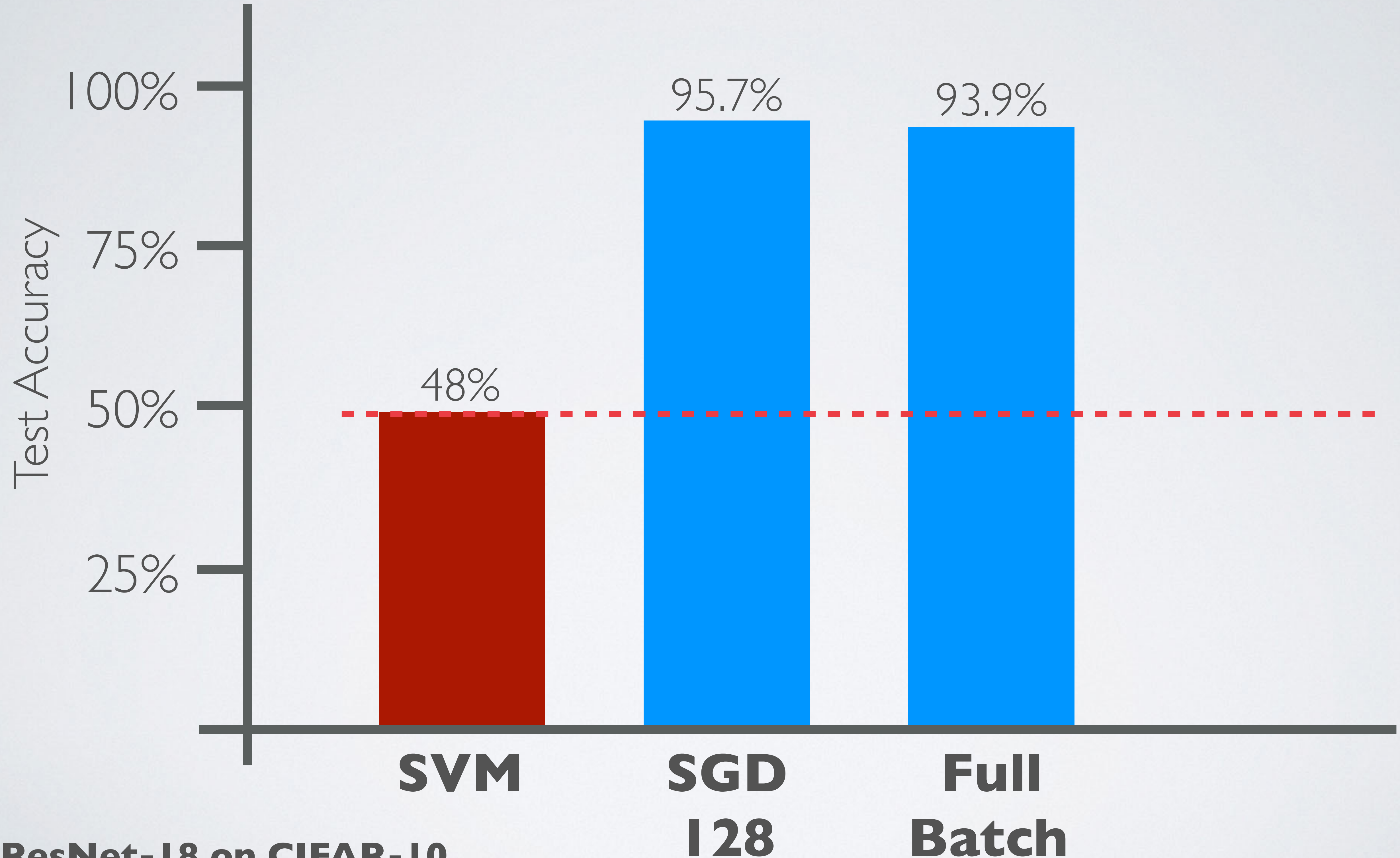
The implicit regularization of SGD



ResNet-18 on CIFAR-10

Stochastic training is not necessary for generalization, ICLR '22

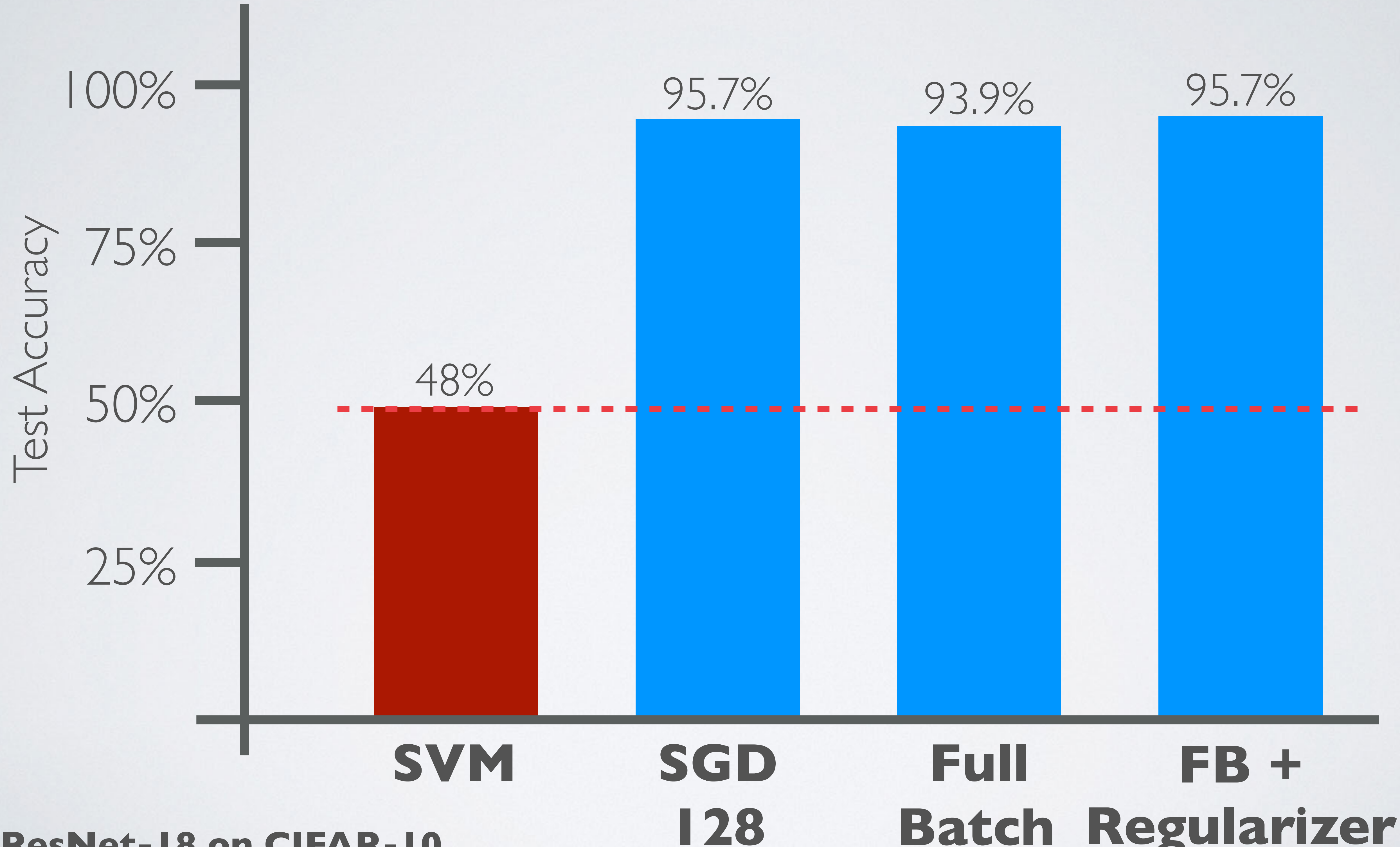
The implicit regularization of SGD



ResNet-18 on CIFAR-10

Stochastic training is not necessary for generalization, ICLR '22

The implicit regularization of ~~S~~GD

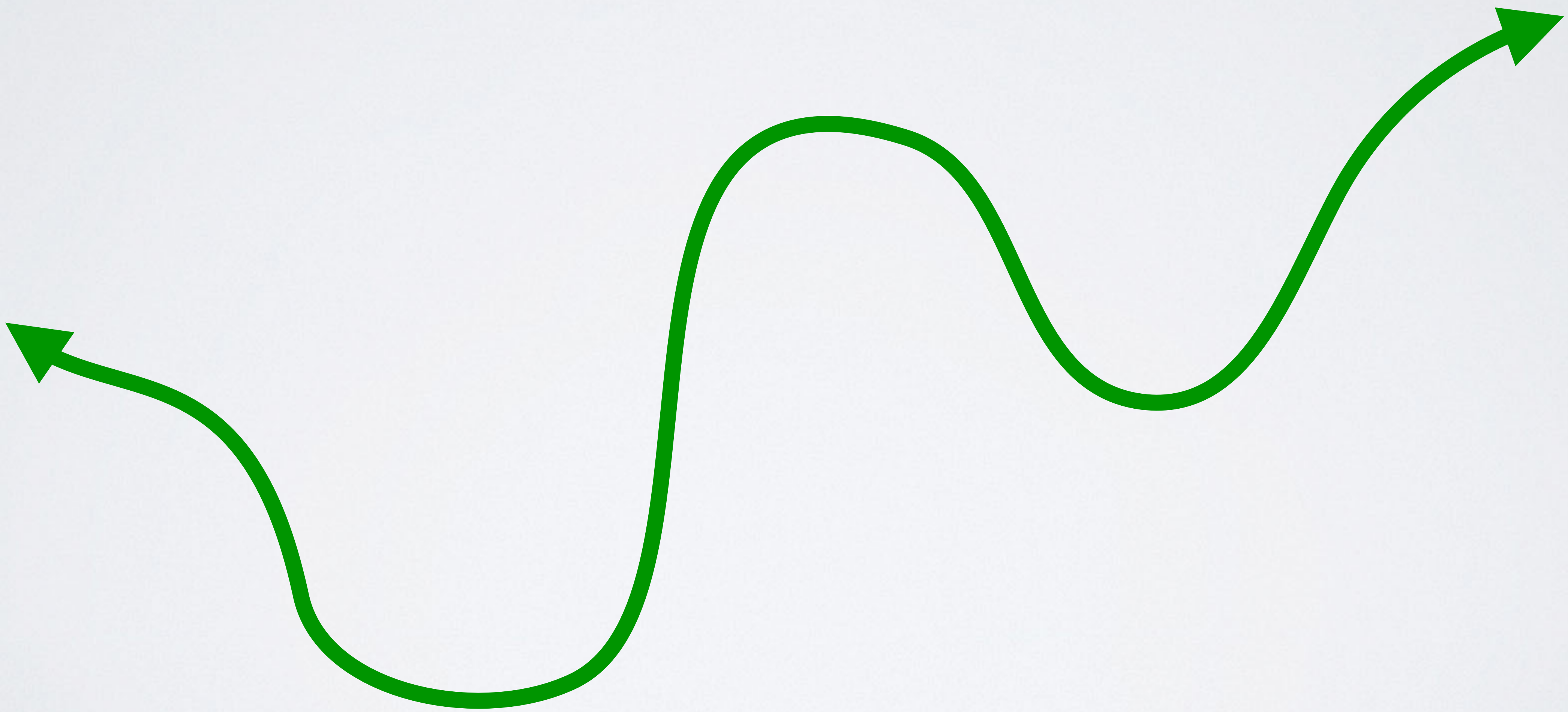


ResNet-18 on CIFAR-10

Stochastic training is not necessary for generalization, ICLR '22

The implicit regularization of ~~SG~~D

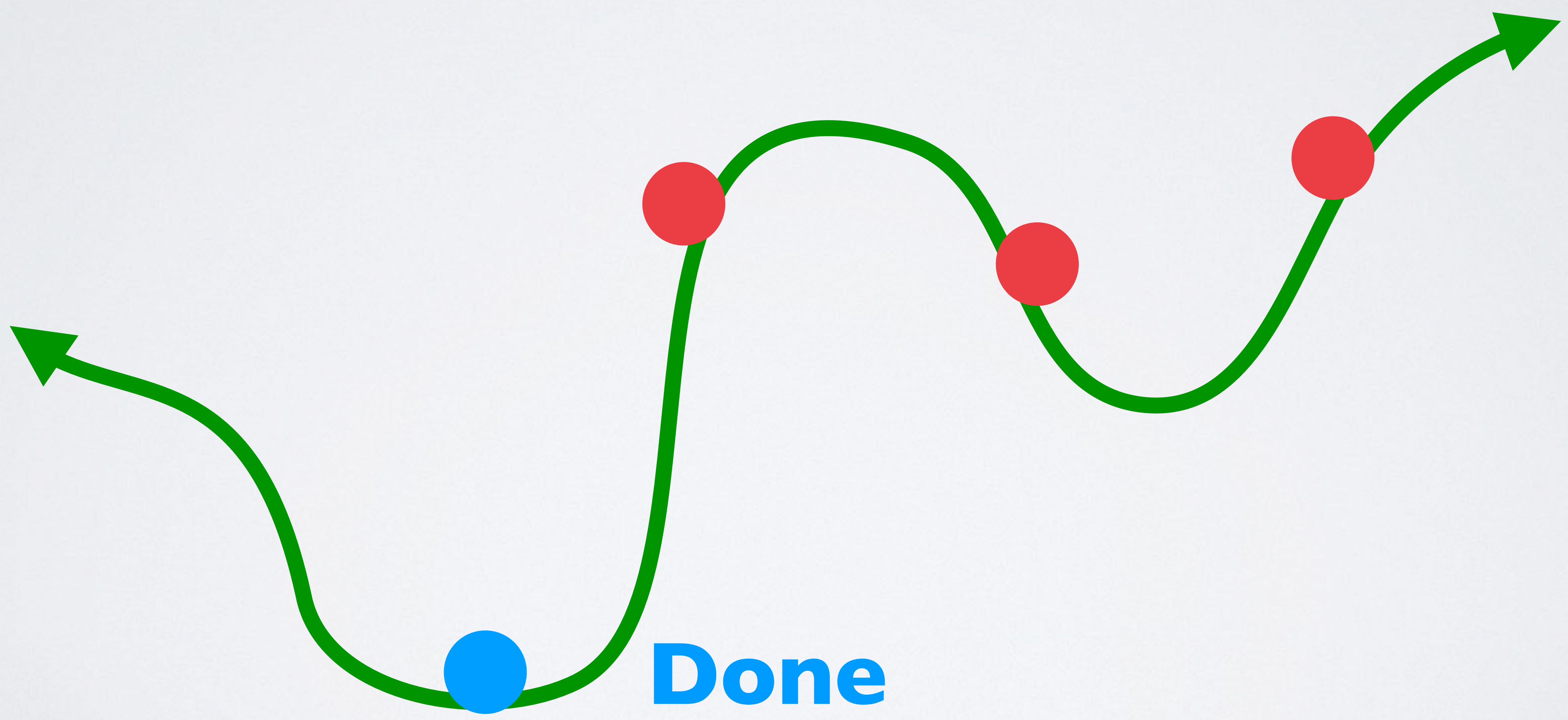
Guess and Check!



Gradient-based optimization is not necessary for generalization, ICLR '23

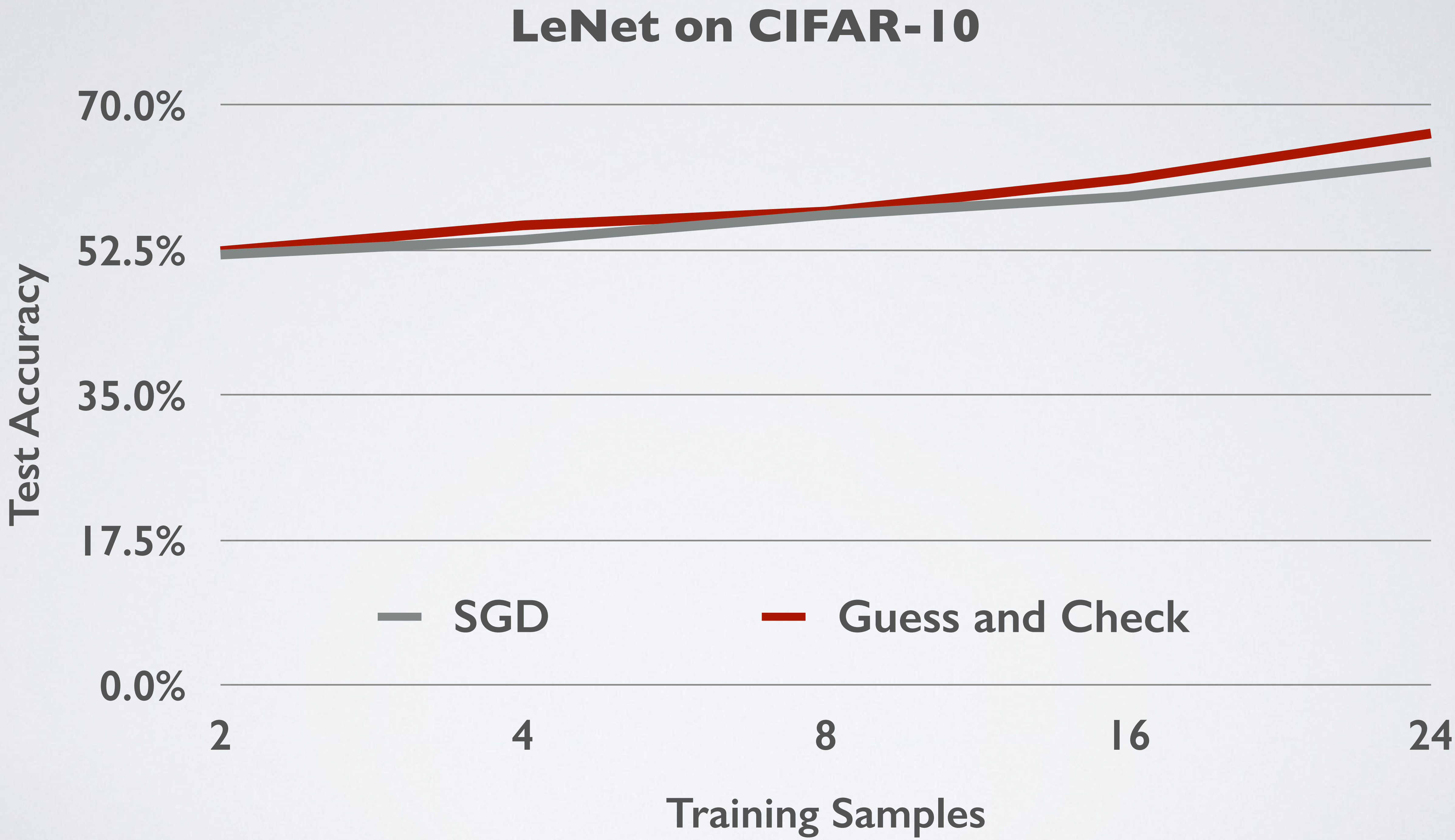
The implicit regularization of SGD

Guess and Check!



Gradient-based optimization is not necessary for generalization, ICLR '23

The implicit regularization of ~~SGD~~



Gradient-based optimization is not necessary for generalization, ICLR '23

What's the difference between good and bad minima?

The sharp vs. flat dilemma

The sharp vs. flat dilemma

“Good” minima are “flat”

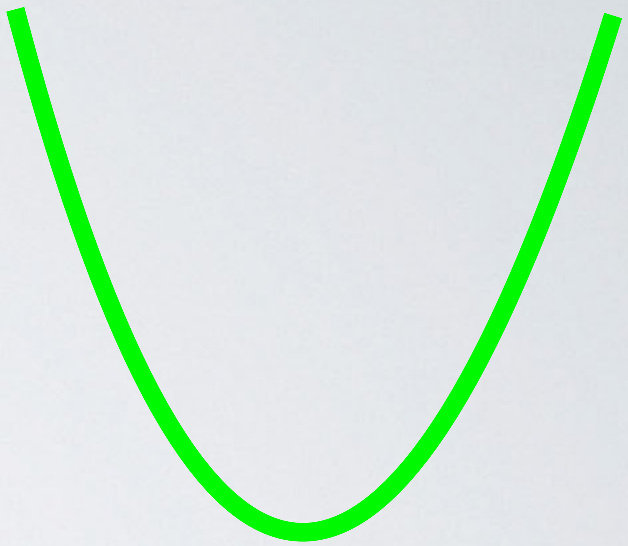
Hochreiter & Schmidhuber, Flat Minima ‘97

Chaudhari et al, Entropy SGD ‘17

Keskar et al, On large batch training ‘17

Li et al, Visualizing the loss landscape ‘18

Flat



Sharp

The sharp vs. flat dilemma

“Good” minima are “flat”

Hochreiter & Schmidhuber, Flat Minima ‘97

Chaudhari et al, Entropy SGD ‘17

Keskar et al, On large batch training ‘17

Li et al, Visualizing the loss landscape ‘18

“Flat” minima are “good”

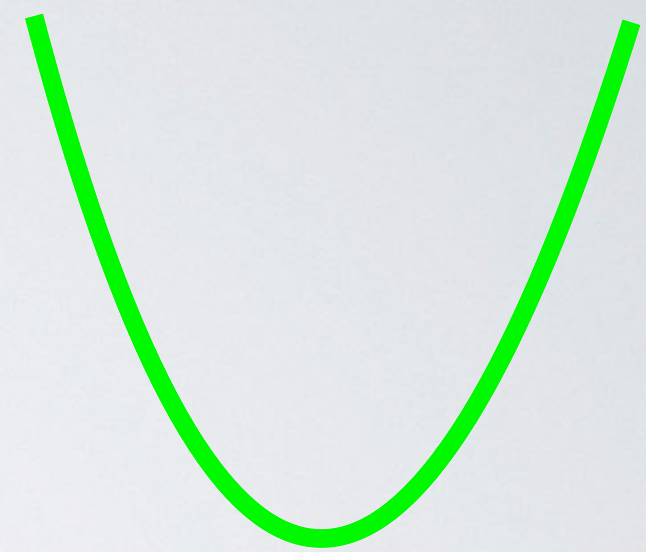
Dziugaite & Roy, Computing non-vacuous ‘17

Izmailov et al, Averaging weights ‘18

Foret et al, Sharpness aware minimization ‘21

Geiping et al, Stochastic training is not necessary ‘21

Flat



Sharp

The sharp vs. flat dilemma

“Good” minima are “flat”

Hochreiter & Schmidhuber, Flat Minima ‘97

Chaudhari et al, Entropy SGD ‘17

Keskar et al, On large batch training ‘17

Li et al, Visualizing the loss landscape ‘18

“Flat” minima are “good”

Dziugaite & Roy, Computing non-vacuous ‘17

Izmailov et al, Averaging weights ‘18

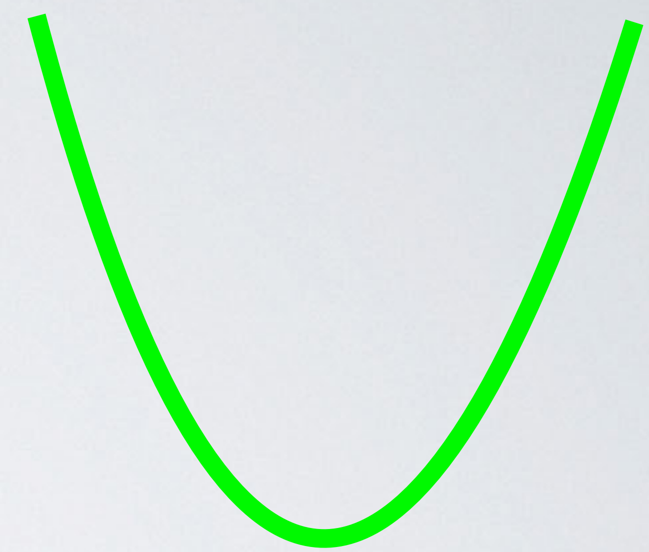
Foret et al, Sharpness aware minimization ‘21

Geiping et al, Stochastic training is not necessary ‘21

...but you have to define “sharp” carefully

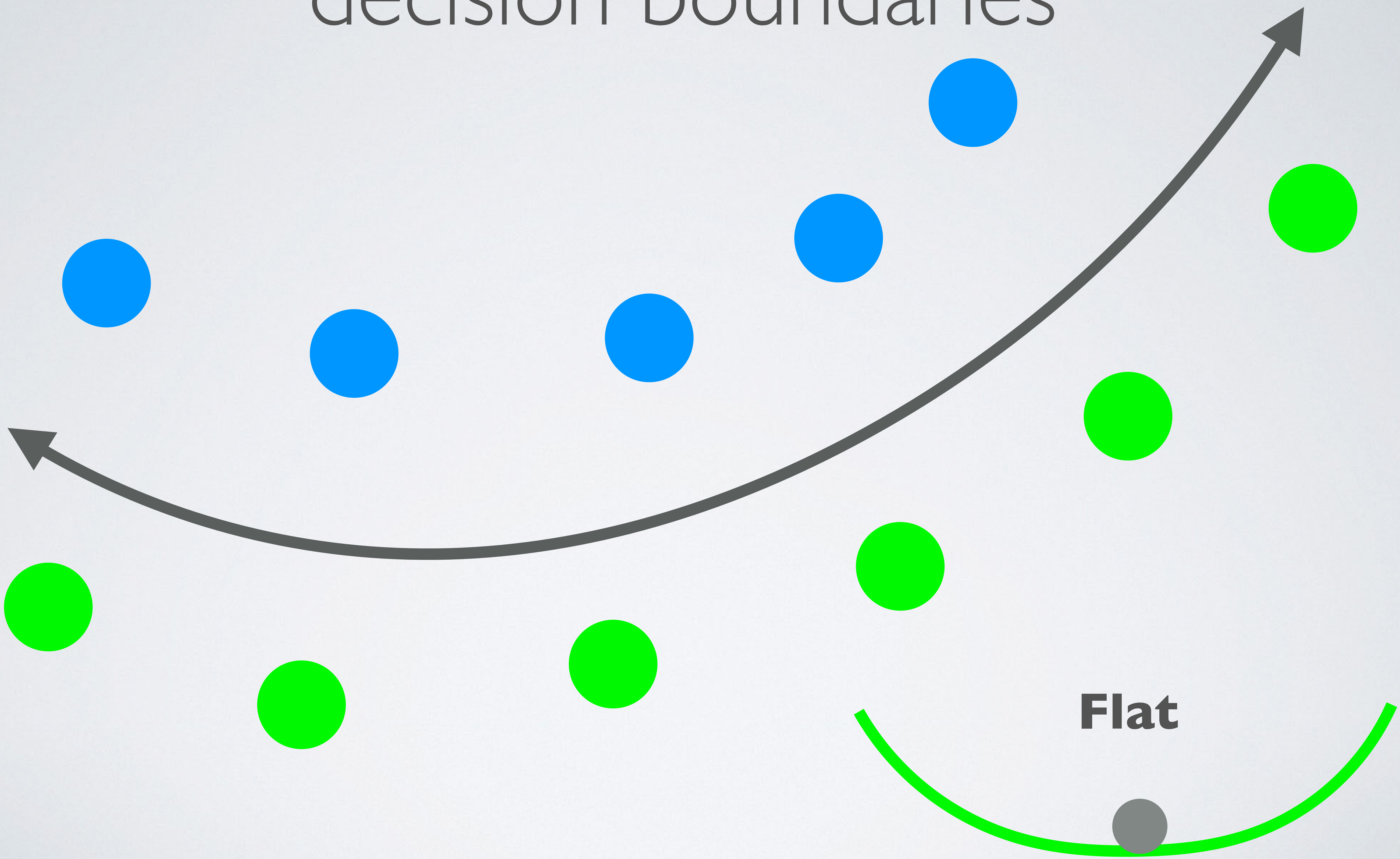
Dinh, Pascanu, Bengio & Bengio,
Sharp minima can generalize for deep nets ‘17

Flat



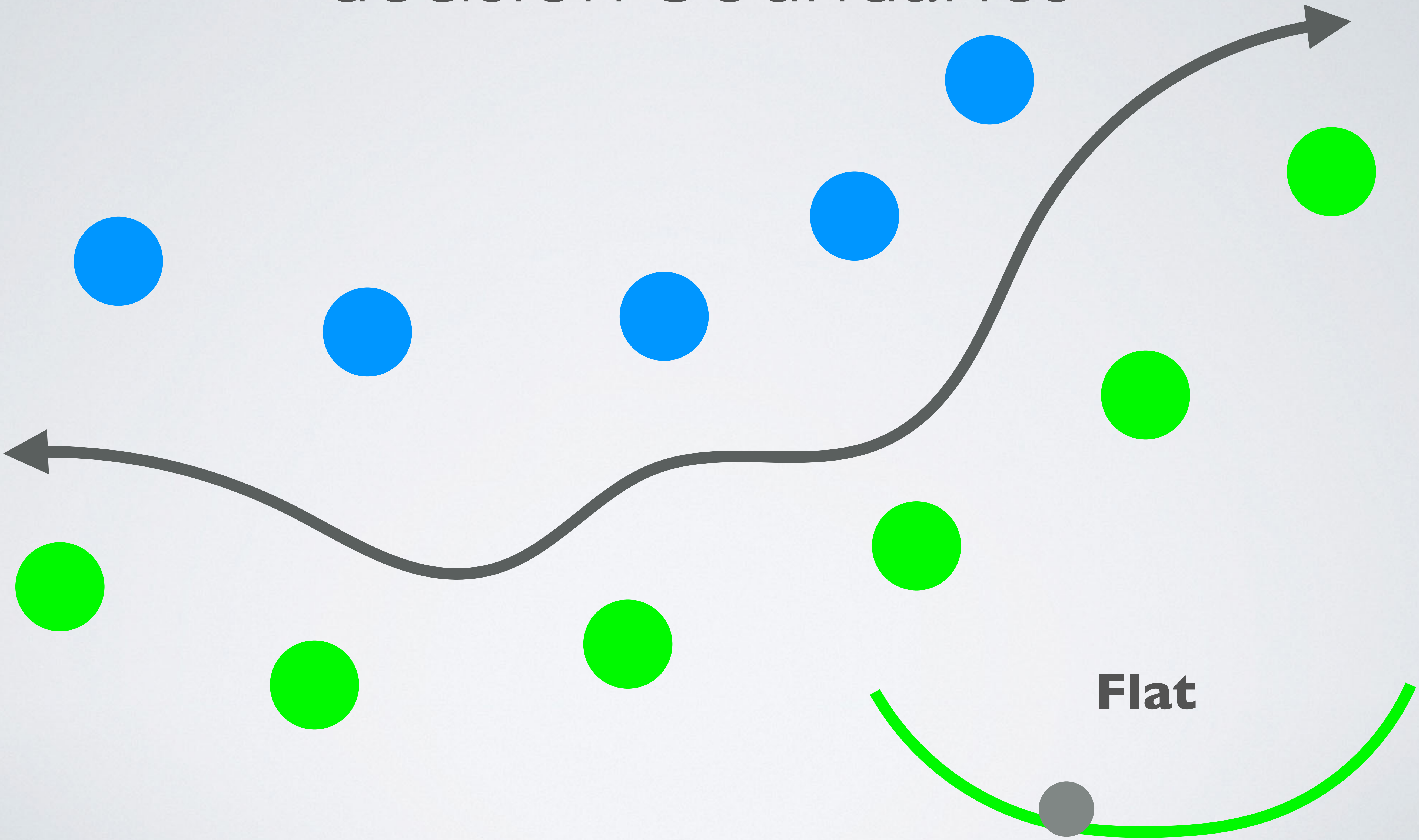
Sharp

Flatness is a wide margin criterion for decision boundaries



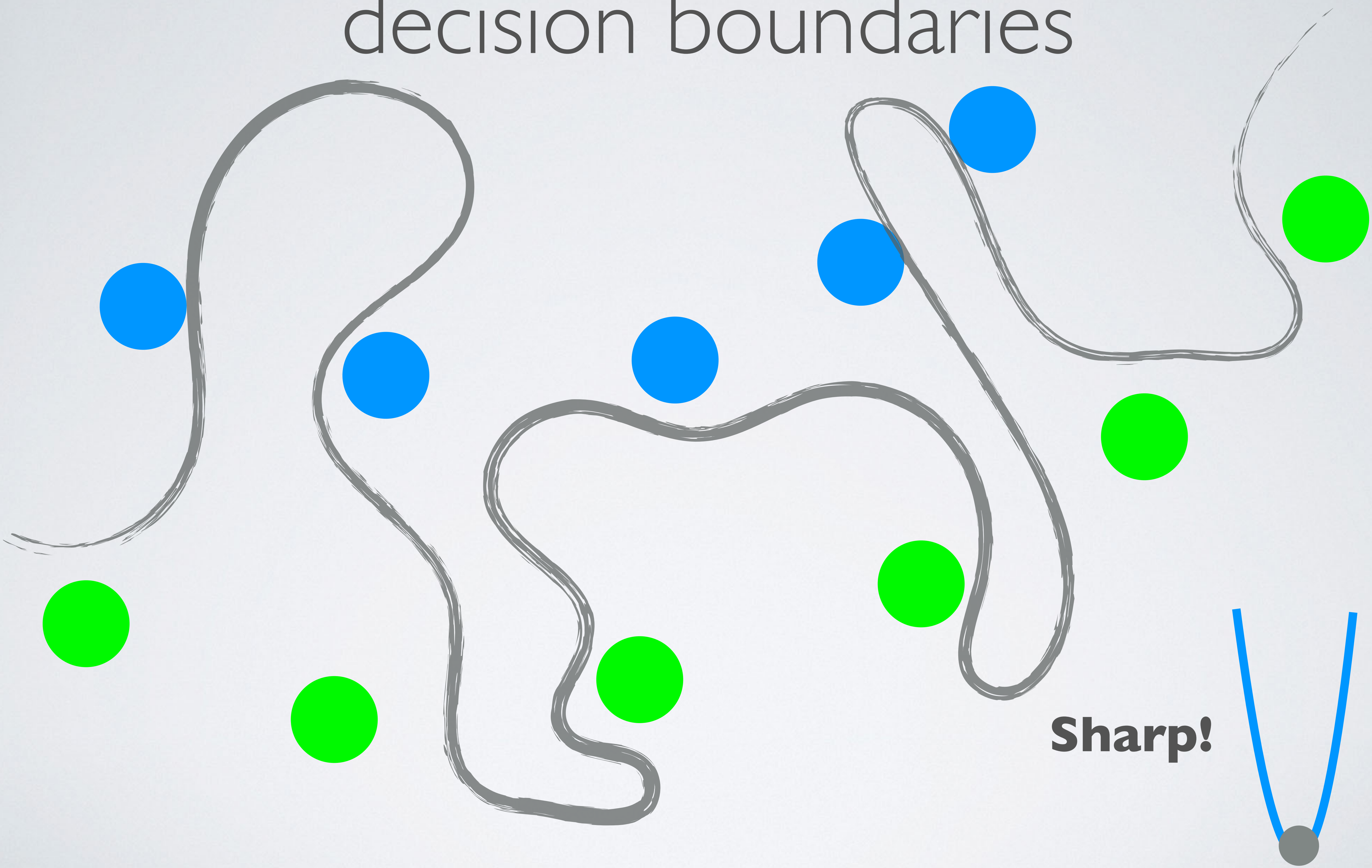
Understanding generalization through visualizations, Under Review

Flatness is a wide margin criterion for decision boundaries



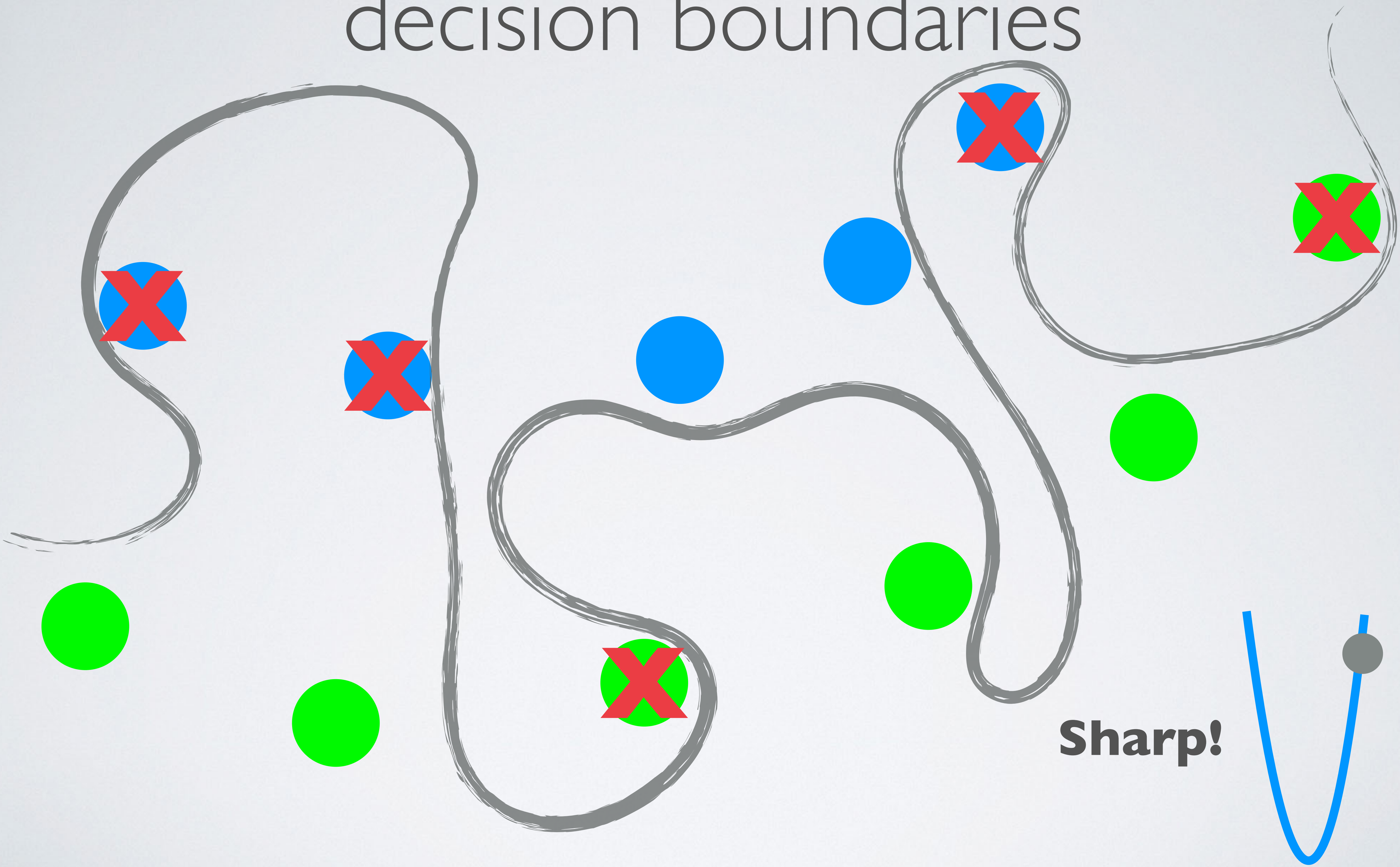
Understanding generalization through visualizations, Under Review

Flatness is a wide margin criterion for decision boundaries

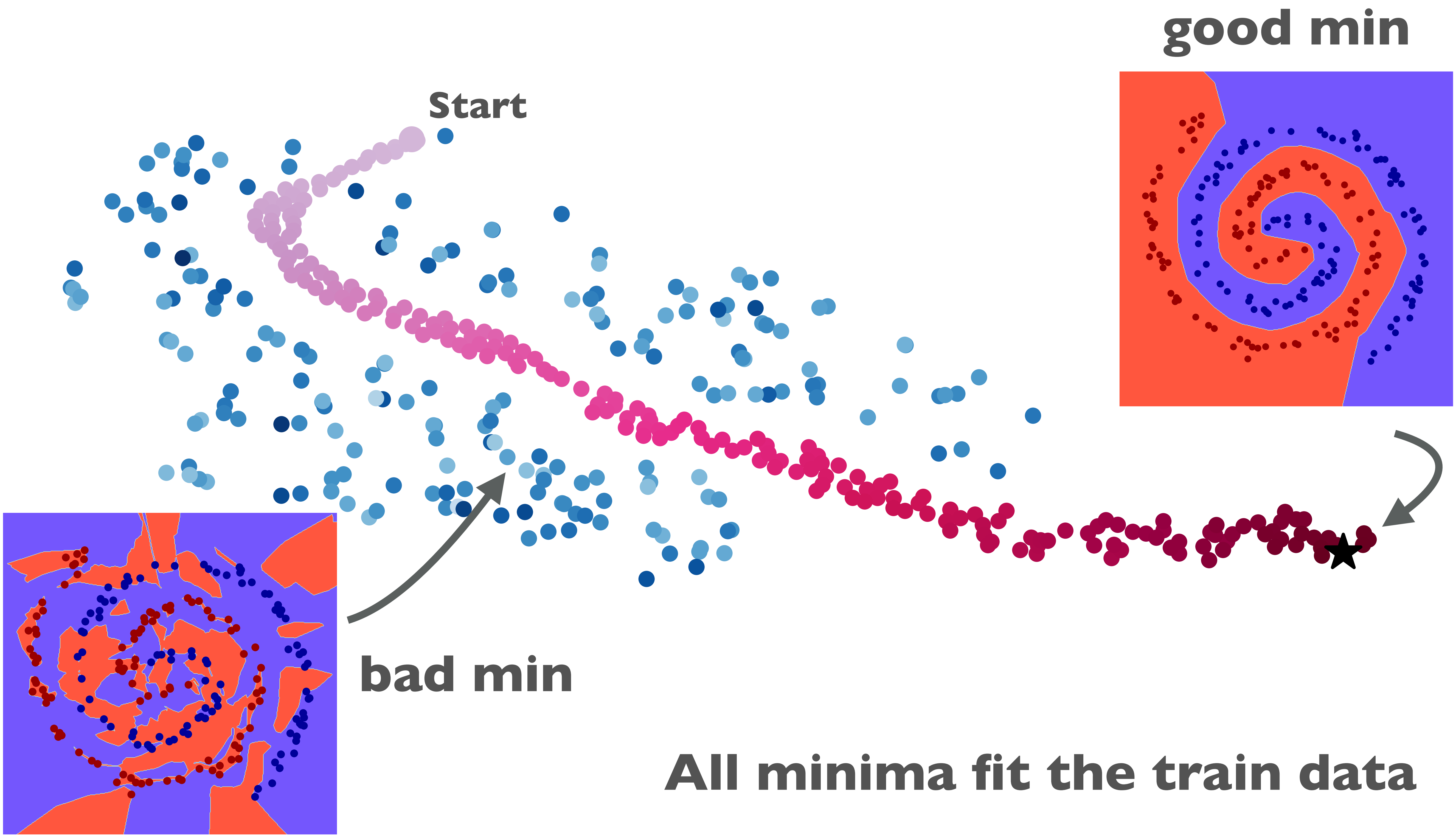


Understanding generalization through visualizations, Under Review

Flatness is a wide margin criterion for decision boundaries

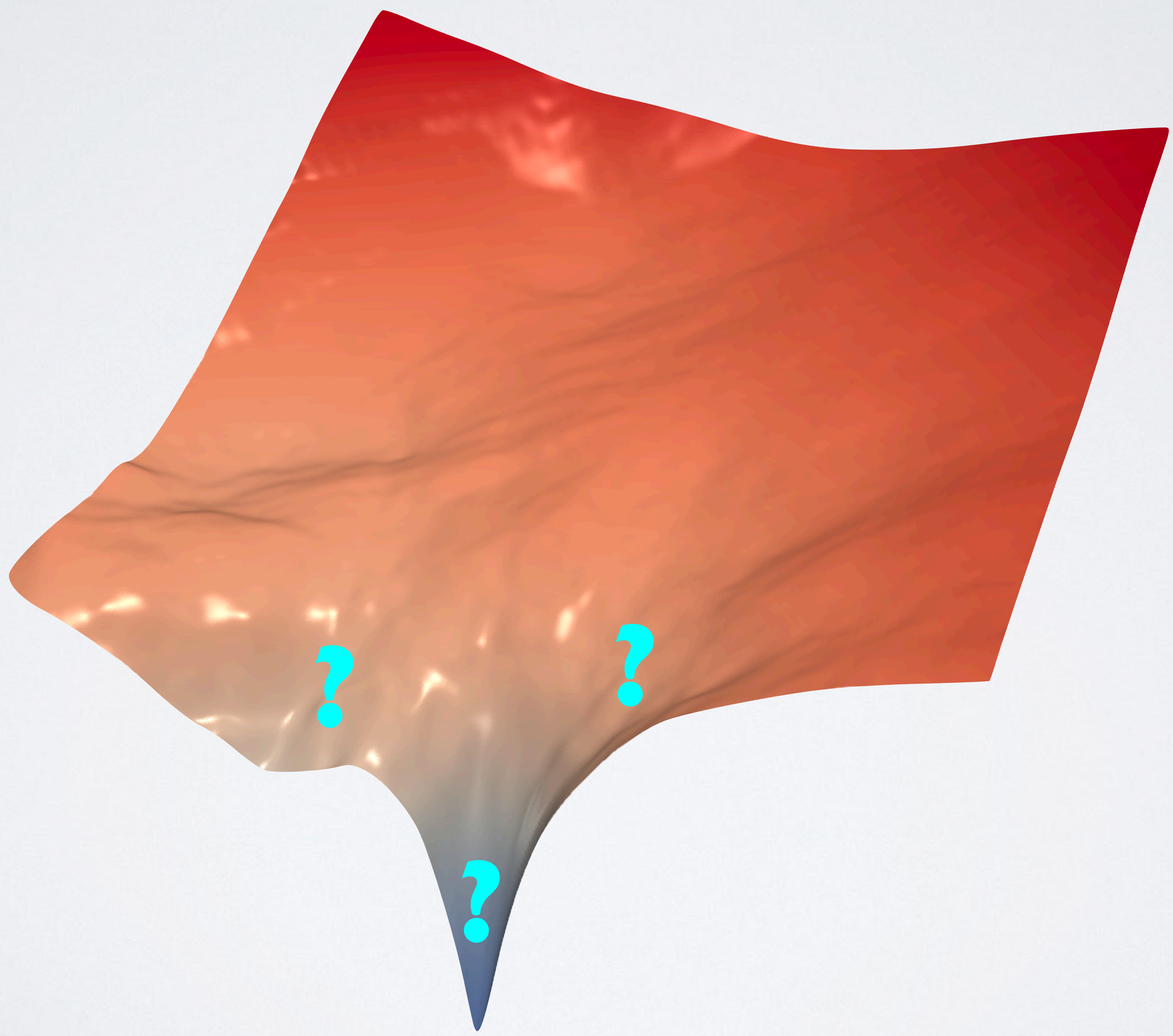


Understanding generalization through visualizations, Under Review

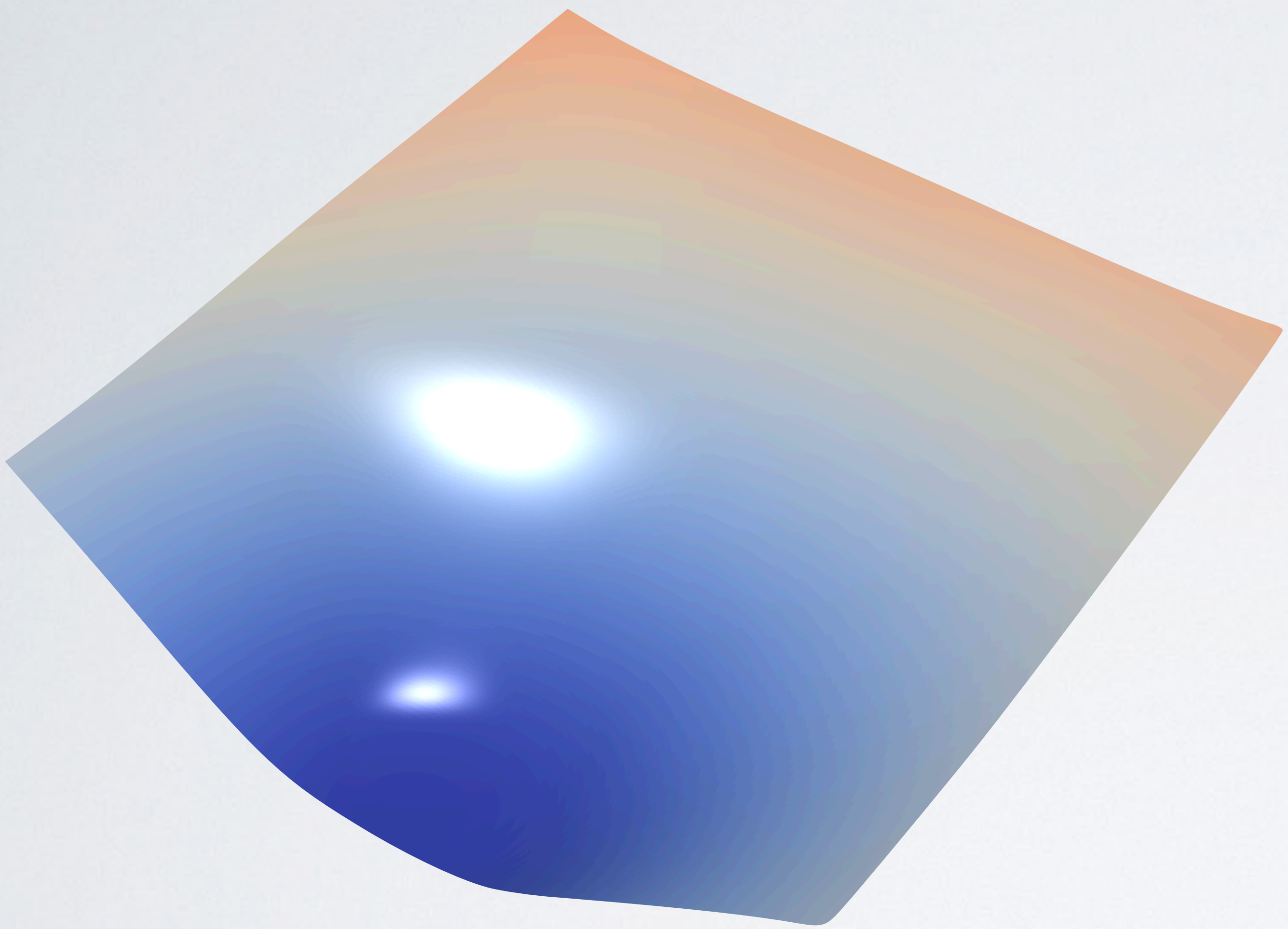


Understanding generalization through visualizations, Under Review

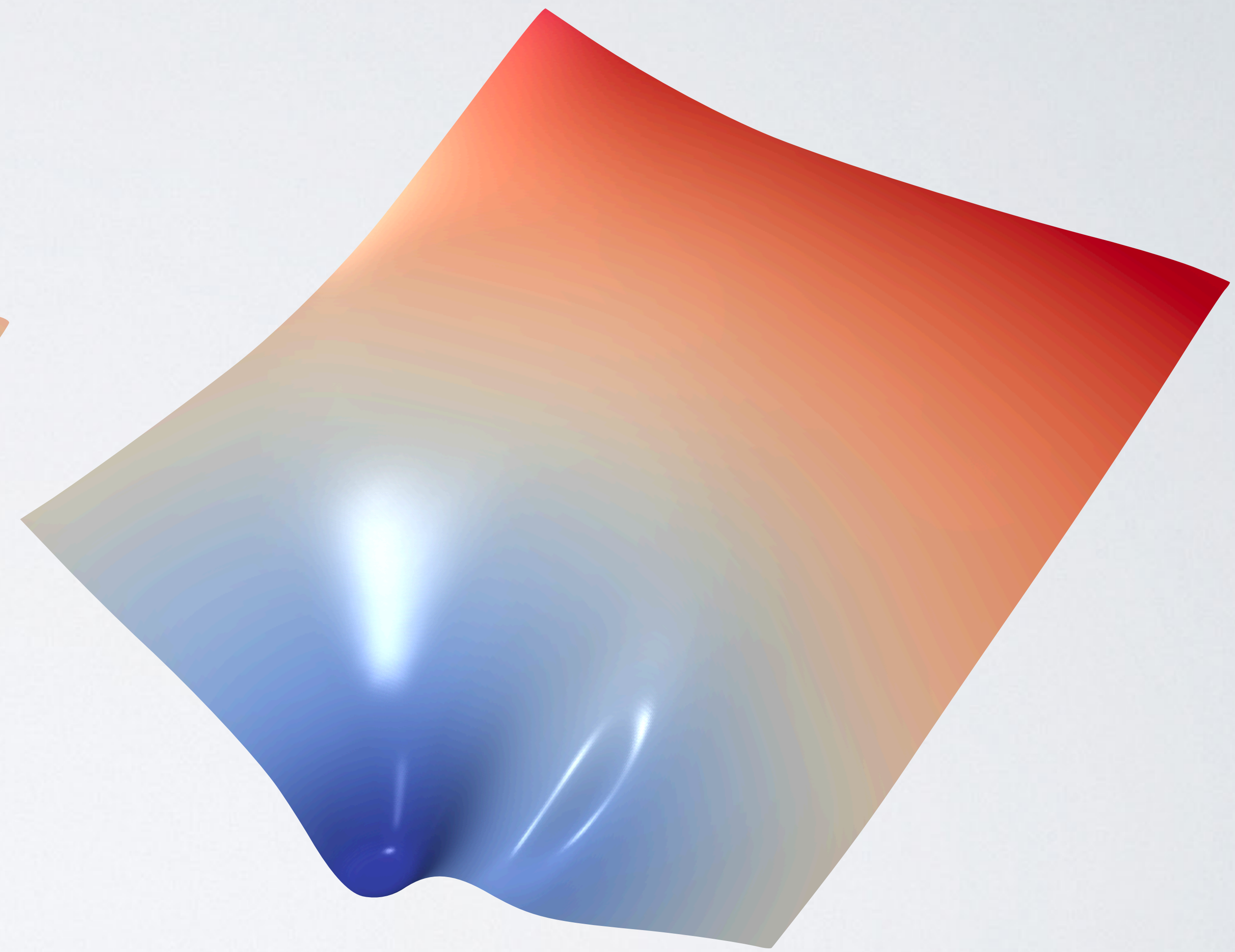
Will incompressible solutions generalize?



A good minimum
100% train 97% test



A bad minimum
100% train 28% test



Street View House Numbers

Understanding generalization through visualizations, Under Review

Why does generalization happen?

Flat minima in high dimensions

Understanding generalization through visualizations, Under Review

Flat minima in high dimensions

flat minima → higher volume

Understanding generalization through visualizations, Under Review

Flat minima in high dimensions

flat minima → higher volume

dimensionality amplifies volume differences

Flat minima in high dimensions

flat minima → higher volume

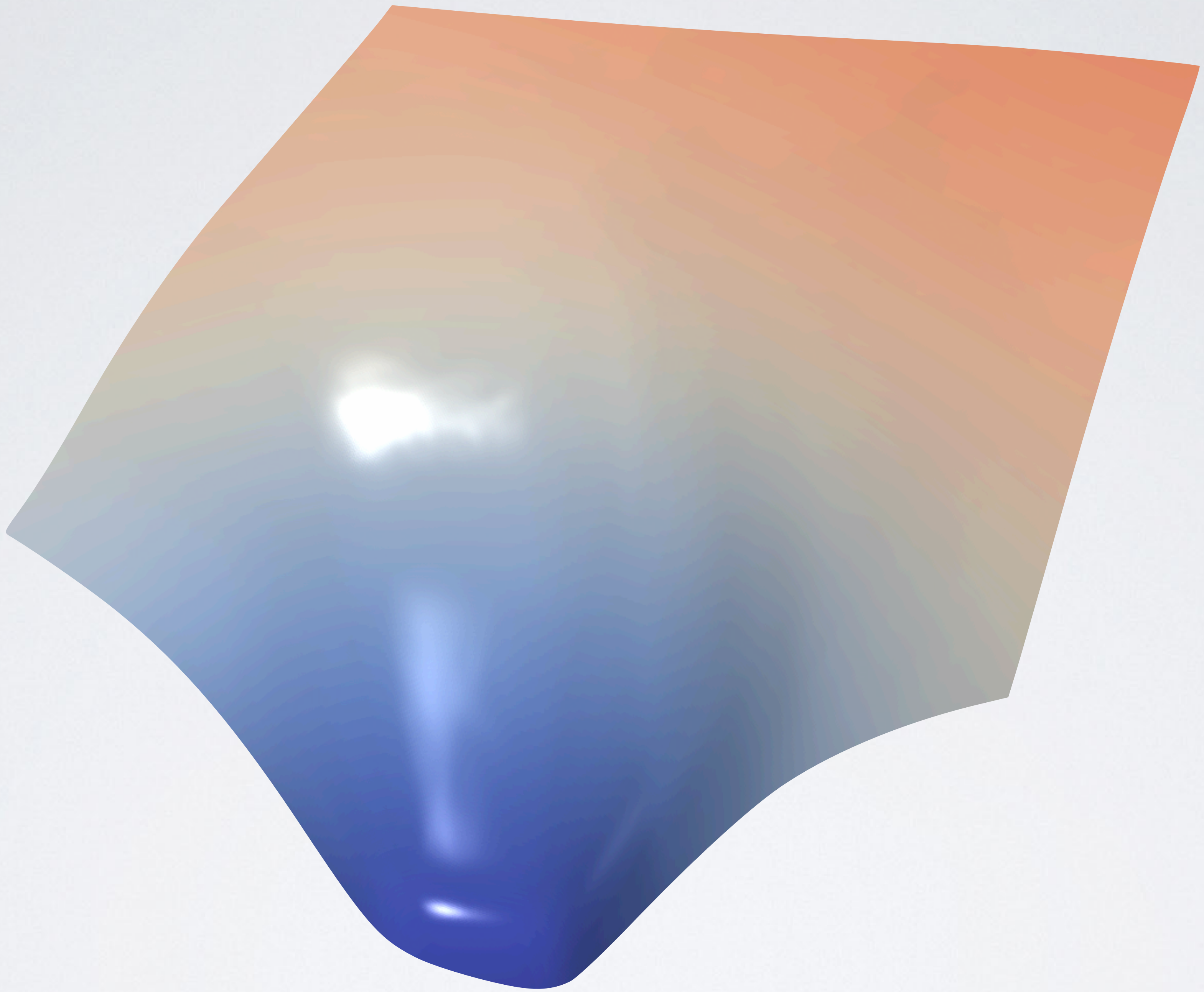
dimensionality amplifies volume differences

easy to find big targets



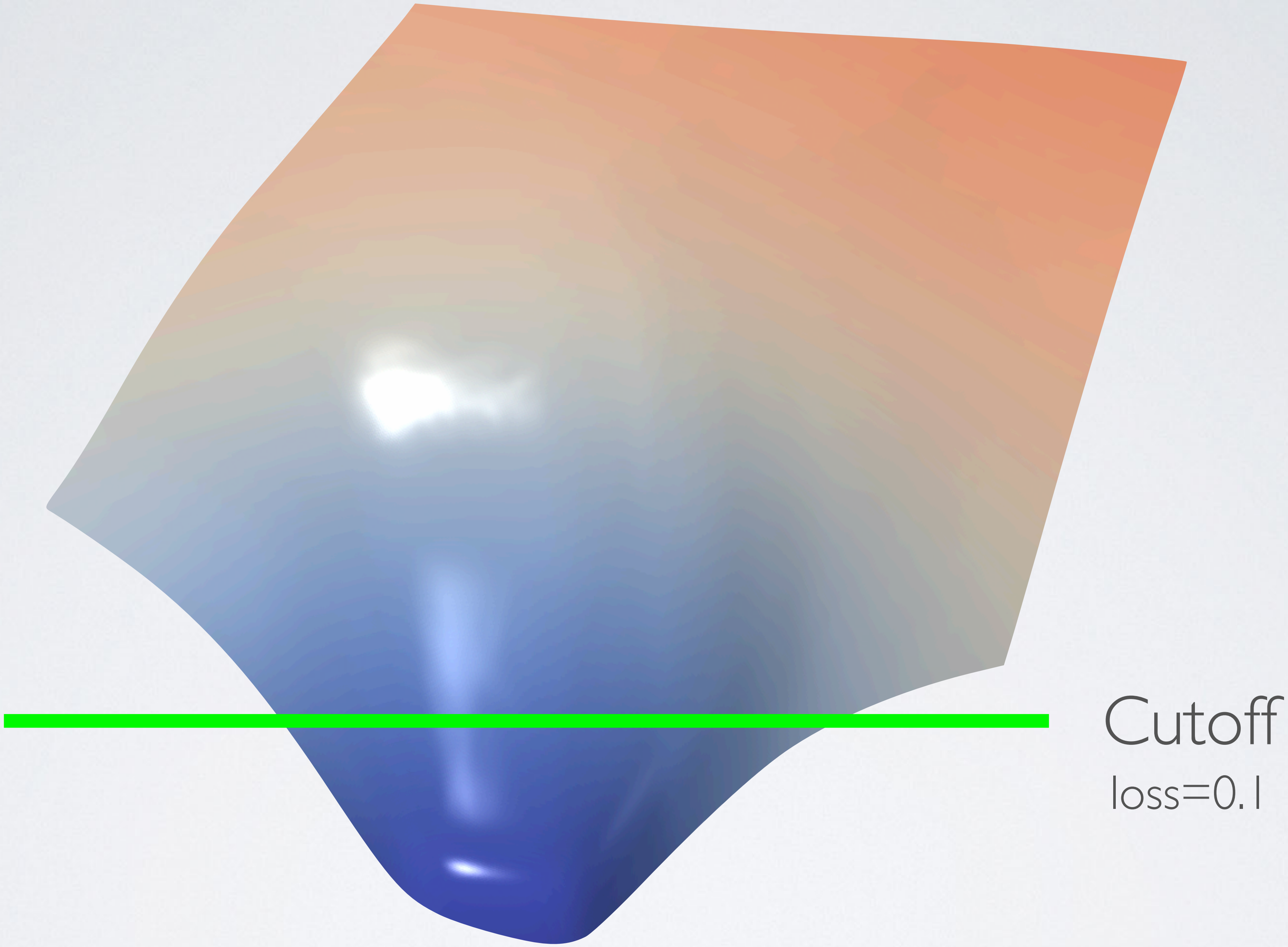
Understanding generalization through visualizations, Under Review

How to quantify the volume of basins around minima?



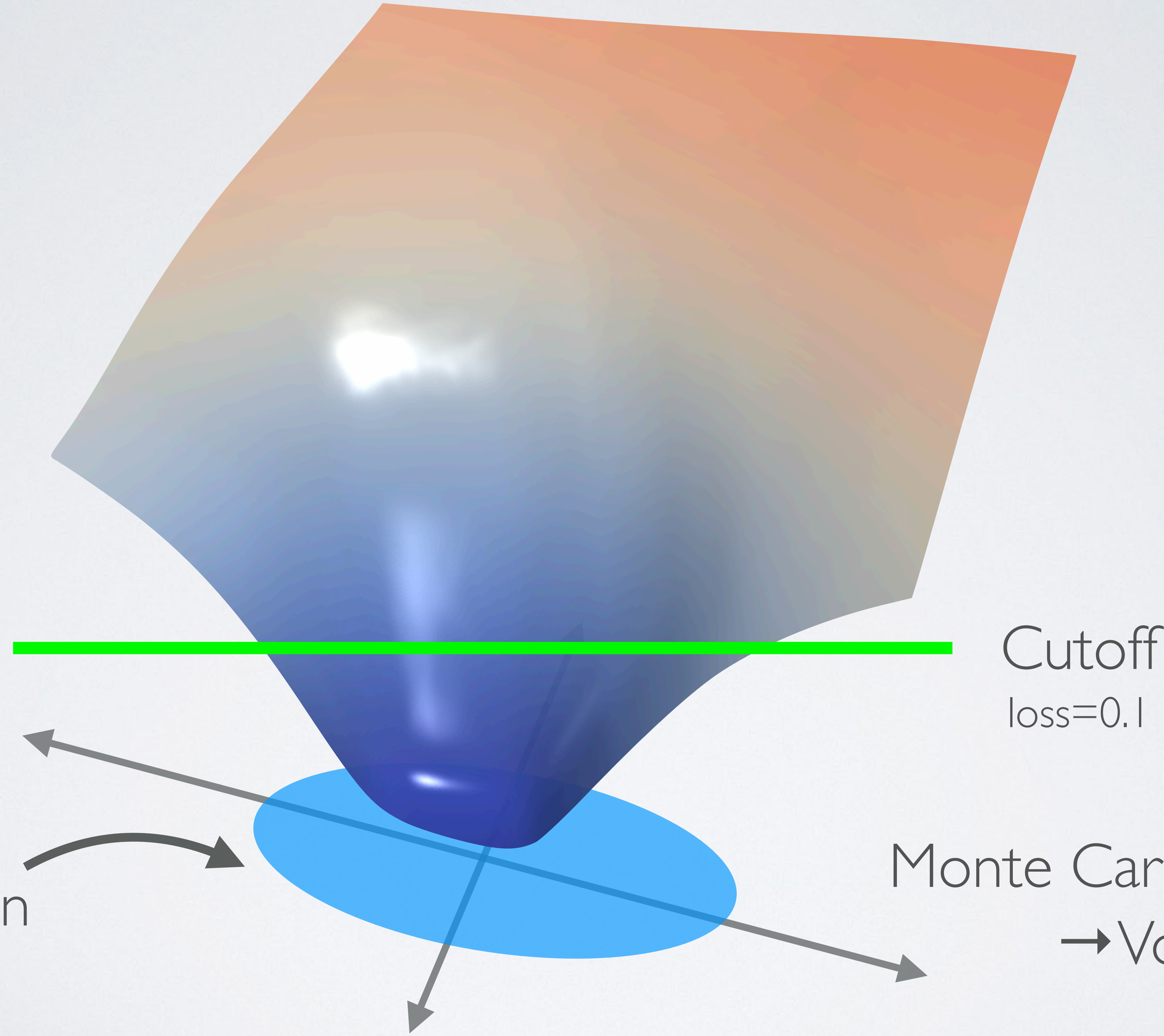
Understanding generalization through visualizations, Under Review

How to quantify the volume of basins around minima?



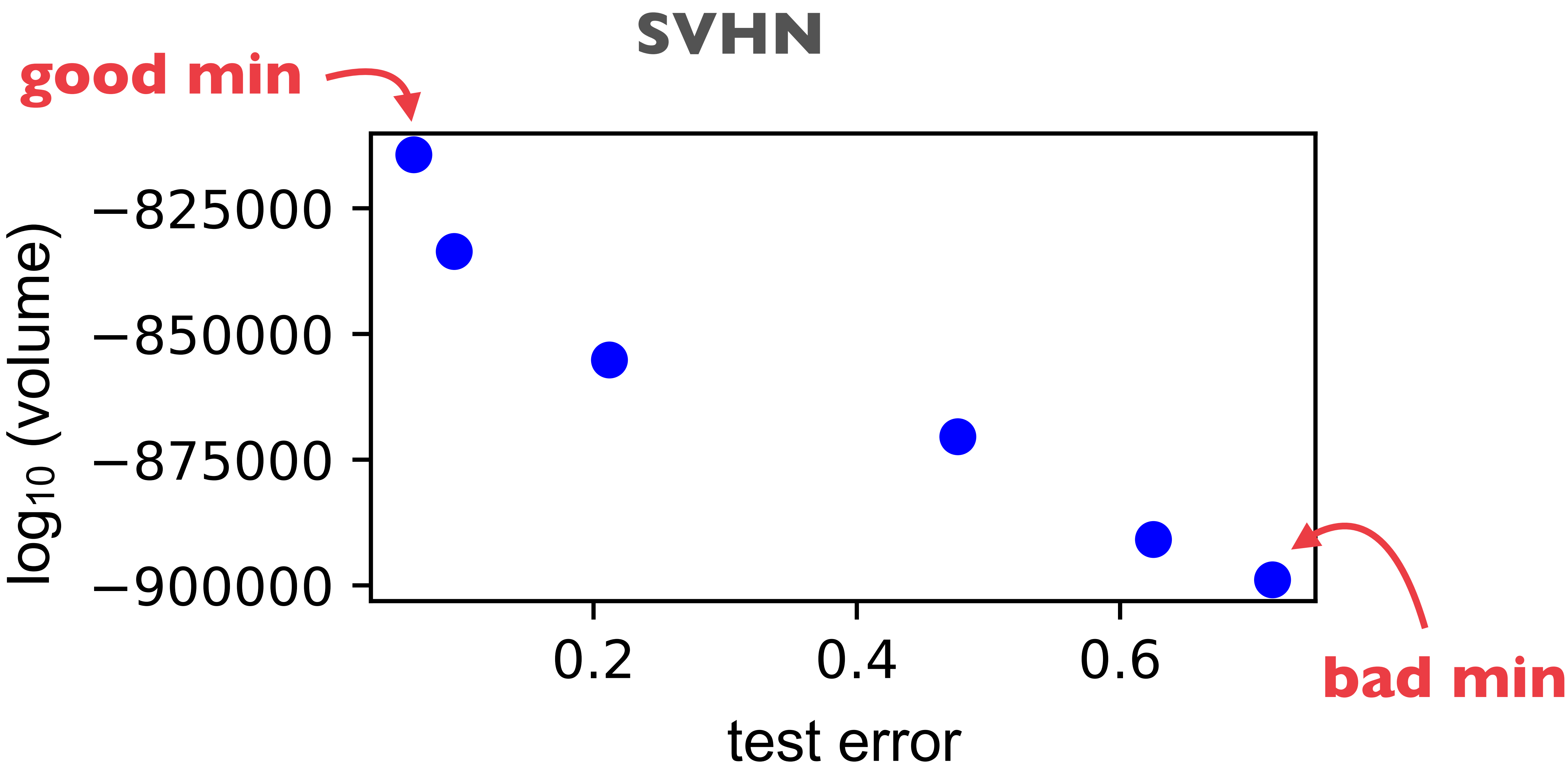
Understanding generalization through visualizations, Under Review

How to quantify the volume of basins around minima?



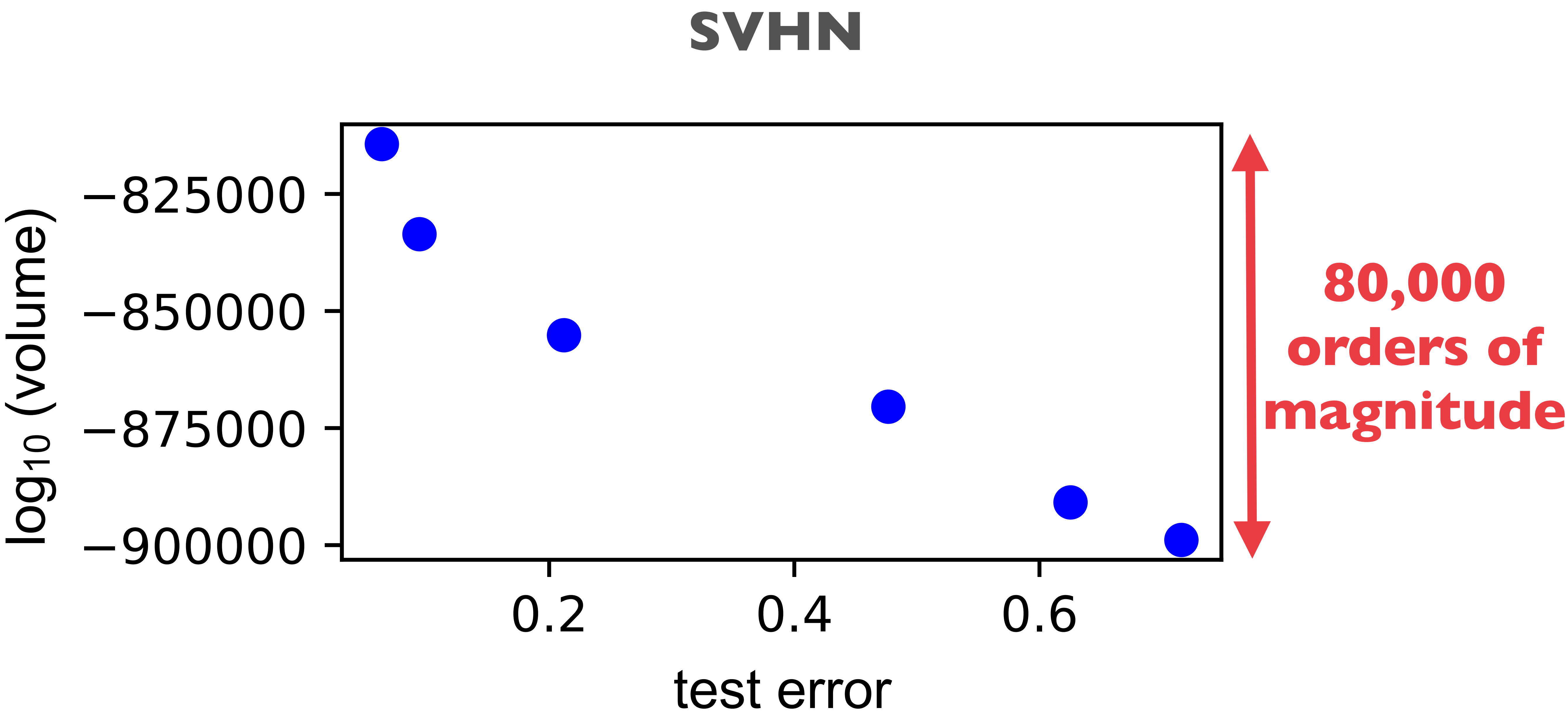
Understanding generalization through visualizations, Under Review

Generalization gap vs. volume



Understanding generalization through visualizations, Under Review

Generalization gap vs. volume



Understanding generalization through visualizations, Under Review

Let’s Summarize!

Why do neural networks generalize?

- Are all minima good? **No!**
- Nothing special about the **optimizer**
- Good minima are **easy to find!**

Let's Summarize!

Why do neural networks generalize?

- Are all minima good? **No!**
- Nothing special about the **optimizer**
- Good minima are **easy to find!**

Experiments

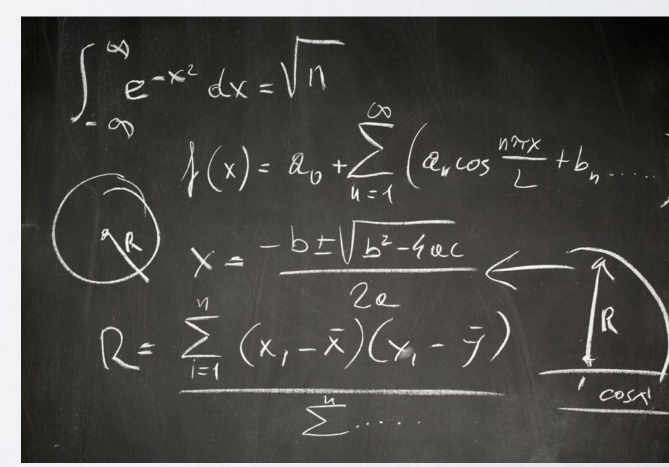
Let's Summarize!

Why do neural networks generalize?

- Are all minima good? **No!**
- Nothing special about the **optimizer**
- Good minima are **easy to find!**

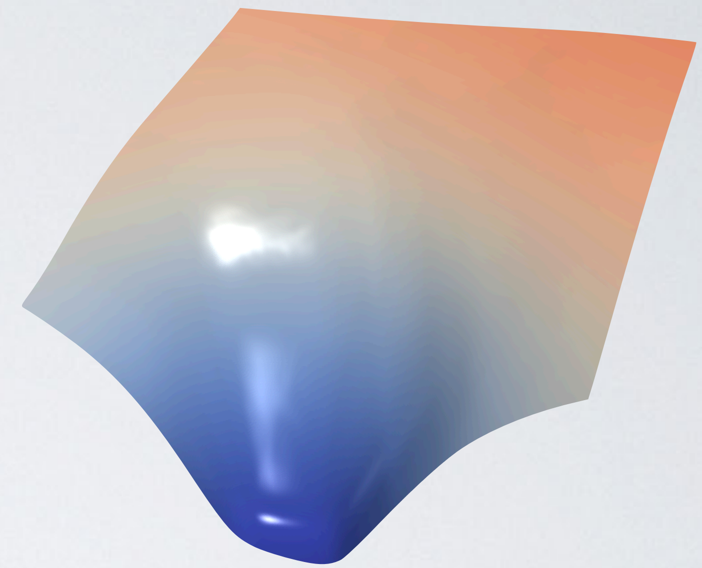
Experiments

Theory?



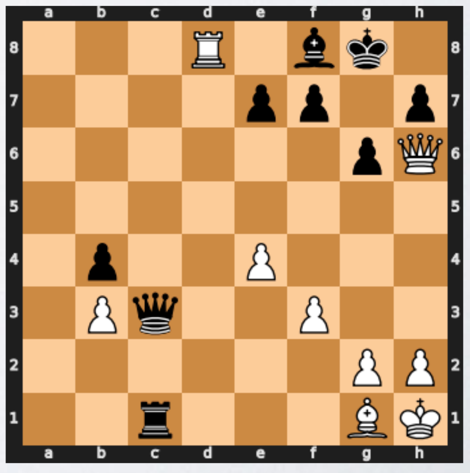
Why do neural networks work?

What are the properties of good minima and why do optimizers find them?



Theories that predict generalization

Observing generalization in reasoning problems



An Anatomy of PAC-Bayes Generalization Bounds

(McAllester 1998)

An Anatomy of PAC-Bayes Generalization Bounds

(McAllester 1998)

P - prior (over parameters)
Q - posterior

An Anatomy of PAC-Bayes Generalization Bounds

(McAllester 1998)

P - prior (over parameters)
Q - posterior

With probability at least $1 - \delta$,

An Anatomy of PAC-Bayes Generalization Bounds

(McAllester 1998)

P - prior (over parameters)
Q - posterior

With probability at least $1 - \delta$,

$$\mathop{\mathbb{E}}_{h \sim Q}\left[R(h)\right]$$

Risk (test error)

An Anatomy of PAC-Bayes Generalization Bounds

(McAllester 1998)

P - prior (over parameters)
Q - posterior

With probability at least $1 - \delta$,

$$\mathop{\mathbb{E}}_{h\sim Q}\left[R\left(h\right)\right] \leq \mathop{\mathbb{E}}_{h\sim Q}[\hat{R}\left(h\right)]$$

Risk (test error)

Empirical Risk
(train error)

An Anatomy of PAC-Bayes Generalization Bounds

(McAllester 1998)

P - prior (over parameters)
Q - posterior

With probability at least $1 - \delta$,

$$\mathop{\mathbb{E}}_{h \sim Q} \left[R(h) \right] \leq \mathop{\mathbb{E}}_{h \sim Q} [\hat{R}(h)] + \sqrt{\frac{\mathbb{KL}(Q \parallel P) + \log(n/\delta) + 2}{2n - 1}}$$

Risk (test error)

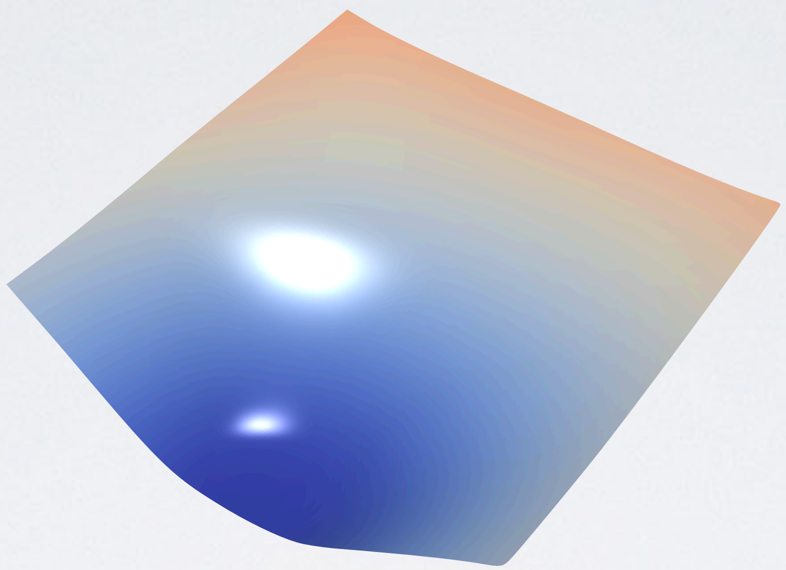
Empirical Risk (train error)

Complexity

PAC-Bayes prefers flat minima

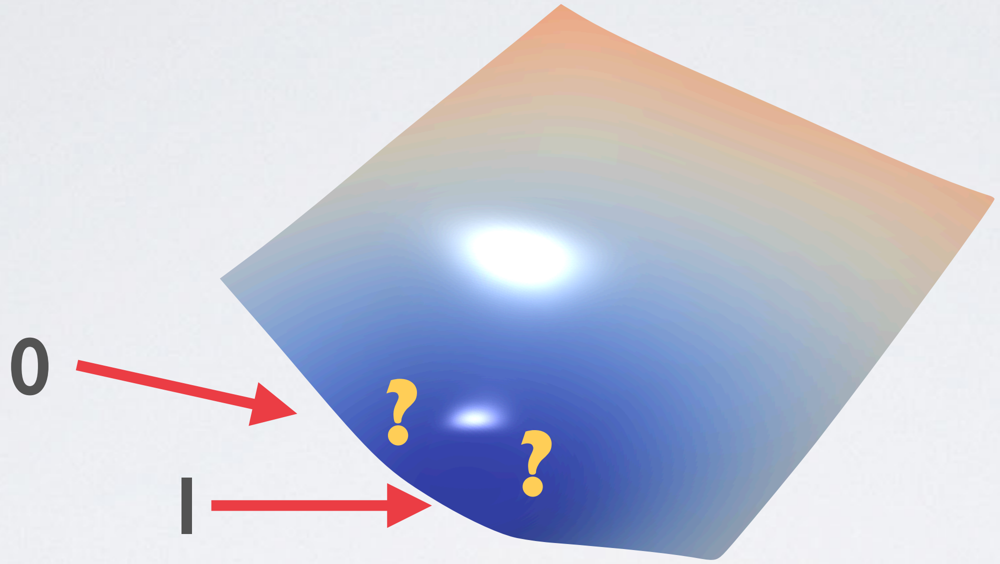
PAC-Bayes prefers flat minima

flat minima → compressible posteriors



PAC-Bayes prefers flat minima

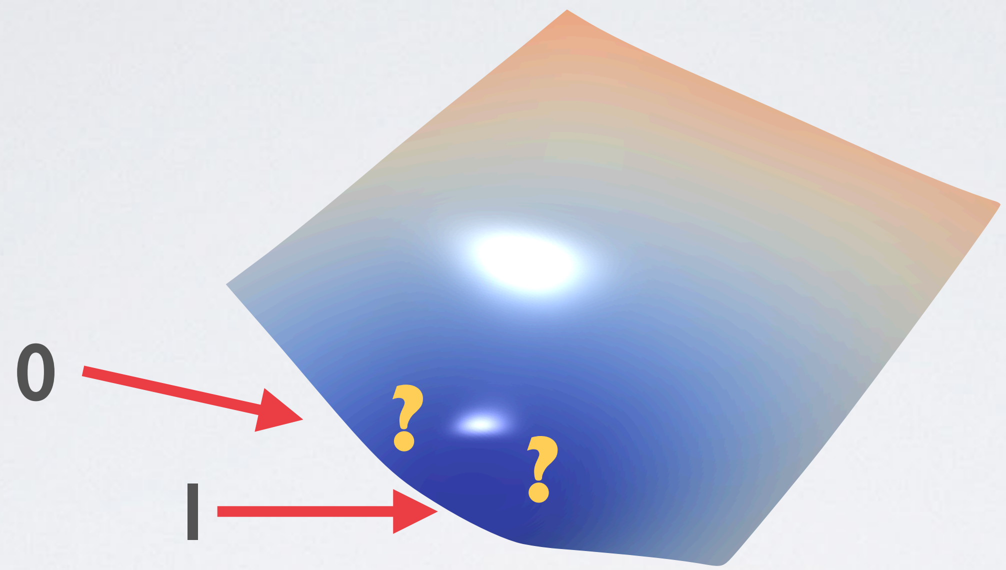
flat minima → compressible posteriors



By choosing parameters, we can encode more information!

PAC-Bayes prefers flat minima

flat minima → compressible posteriors



By choosing parameters, we can encode more information!

Diffuse posteriors achieve better bounds

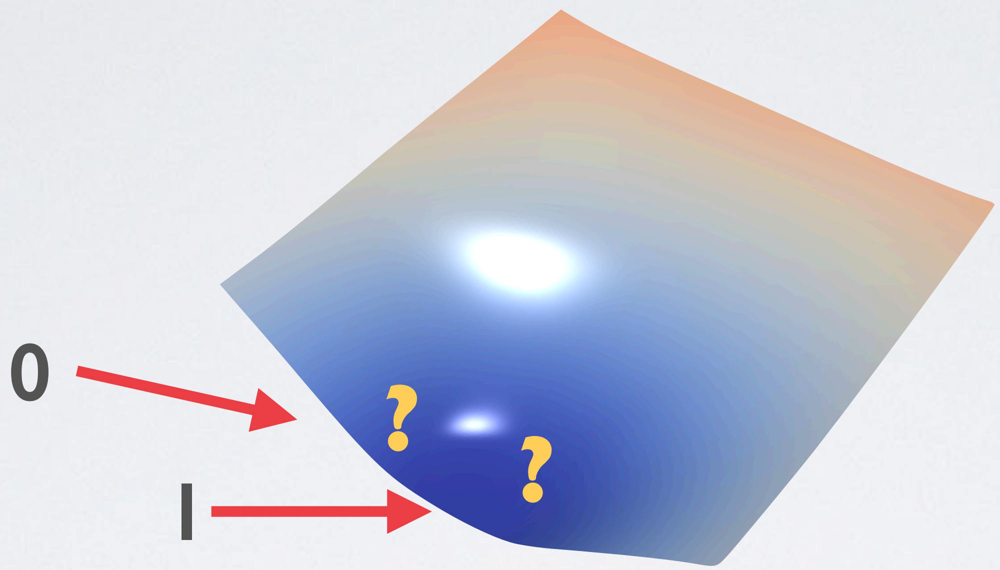
$$\mathbb{KL}(Q \parallel P) = H(Q, P) - H(Q)$$

Cross-Entropy

Shannon Entropy

PAC-Bayes prefers flat minima

flat minima → compressible posteriors



By choosing parameters, we can encode more information!

Diffuse posteriors achieve better bounds

$$\mathbb{KL}(Q \parallel P) = H(Q, P) - H(Q)$$

Cross-Entropy

Shannon Entropy

How to craft tight bounds

How to craft tight bounds

- Frame the problem in terms of compression

How to craft tight bounds

- Frame the problem in terms of compression
- Reduce the number of parameters

How to craft tight bounds

- Frame the problem in terms of compression
- Reduce the number of parameters
- Transfer learning

How to craft tight bounds

- Frame the problem in terms of compression
- Reduce the number of parameters
- Transfer learning
- Quantization

How to craft tight bounds

- Frame the problem in terms of compression
- Reduce the number of parameters
- Transfer learning
- Quantization
- Arithmetic coding

PAC-Bayes Compression Bounds So Tight… NeurIPS '22

Squeezing the juice out of PAC-Bayes

Theoretical bounds on test error, lower is better

	Err. Bound (%)	Previous SOTA (%)
MNIST	**11.6**	21.7
+ SVHN Transfer	**9.0**	16.1
FashionMNIST	**32.8**	46.5
+ CIFAR-10 Transfer	**28.2**	30.1
CIFAR-10	**58.2**	89.9
+ ImageNet Transfer	**35.1**	54.2
CIFAR-100	**94.6**	100
+ ImageNet Transfer	**81.3**	98.1
ImageNet	**93.5**	96.5

PAC-Bayes Compression Bounds So Tight… NeurIPS '22

Squeezing the juice out of PAC-Bayes

Theoretical bounds on test error, lower is better

	Err. Bound (%)	Previous SOTA (%)
MNIST	**11.6**	21.7
+ SVHN Transfer	**9.0**	16.1
FashionMNIST	**32.8**	46.5
+ CIFAR-10 Transfer	**28.2**	30.1
CIFAR-10	**58.2**	89.9
+ ImageNet Transfer	**35.1**	54.2
CIFAR-100	**94.6**	100
+ ImageNet Transfer	**81.3**	98.1
ImageNet	**93.5**	96.5

PAC-Bayes Compression Bounds So Tight… NeurIPS '22

Squeezing the juice out of PAC-Bayes

Theoretical bounds on test error, lower is better

	Err. Bound (%)	Previous SOTA (%)
MNIST	**11.6**	21.7
+ SVHN Transfer	**9.0**	16.1
FashionMNIST	**32.8**	46.5
+ CIFAR-10 Transfer	**28.2**	30.1
CIFAR-10	**58.2**	89.9
+ ImageNet Transfer	**35.1**	54.2
CIFAR-100	**94.6**	100
+ ImageNet Transfer	**81.3**	98.1
ImageNet	**93.5**	96.5

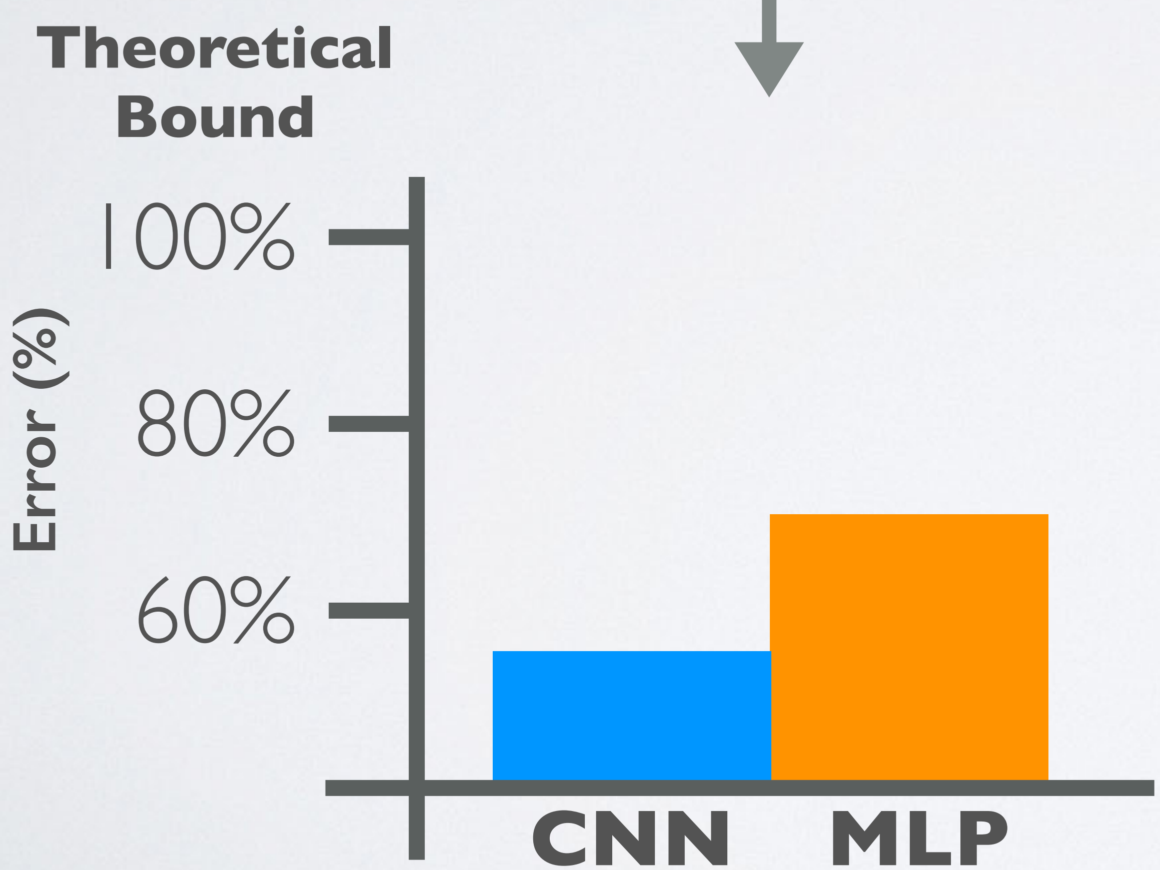
PAC-Bayes Compression Bounds So Tight… NeurIPS '22

Can our theory predict important phenomena in real architectures?

CNNs prefer data with spatial structure

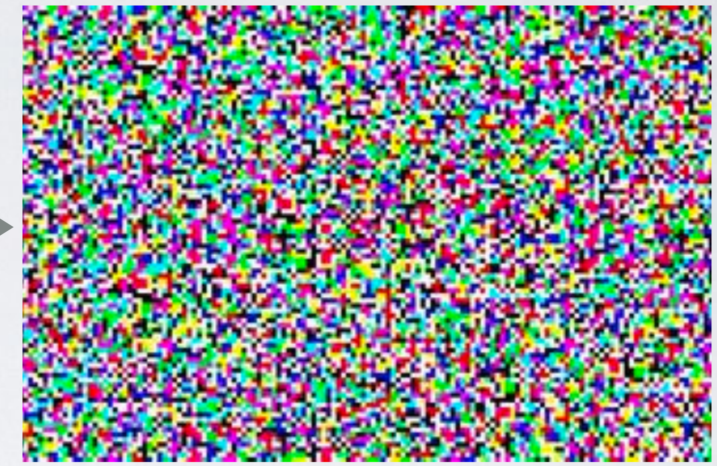
CNNs prefer data with spatial structure

CNNs prefer data with spatial structure



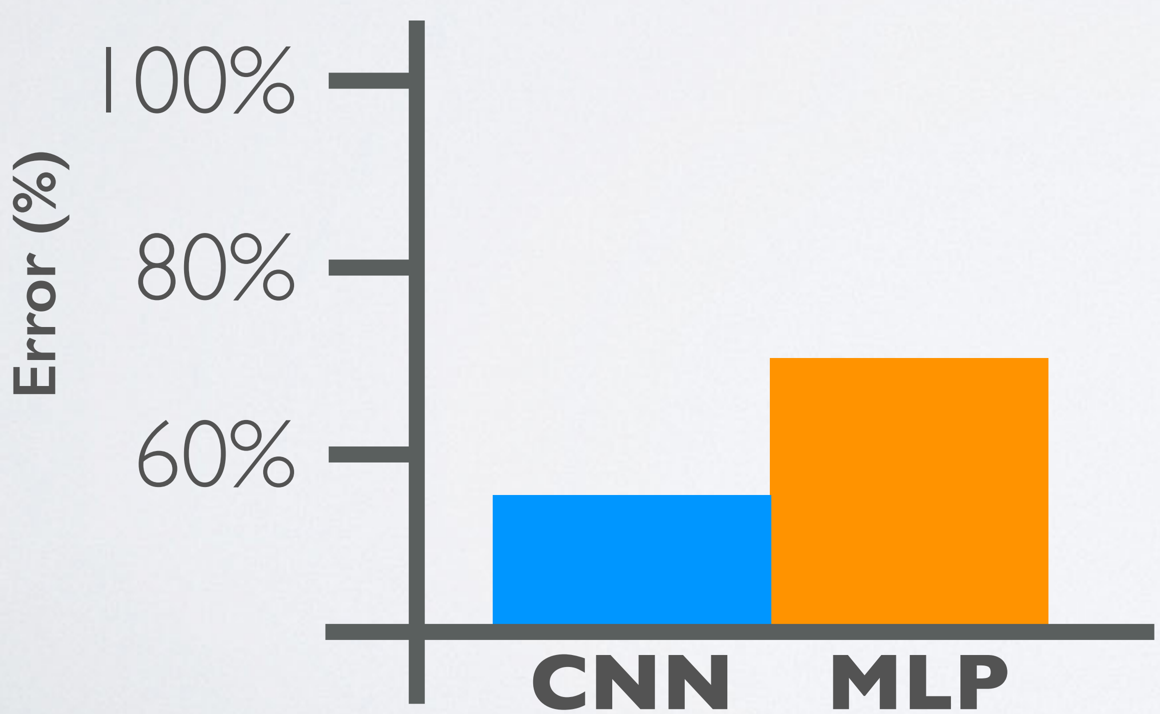
PAC-Bayes Compression Bounds So Tight… NeurIPS '22

CNNs prefer data with spatial structure

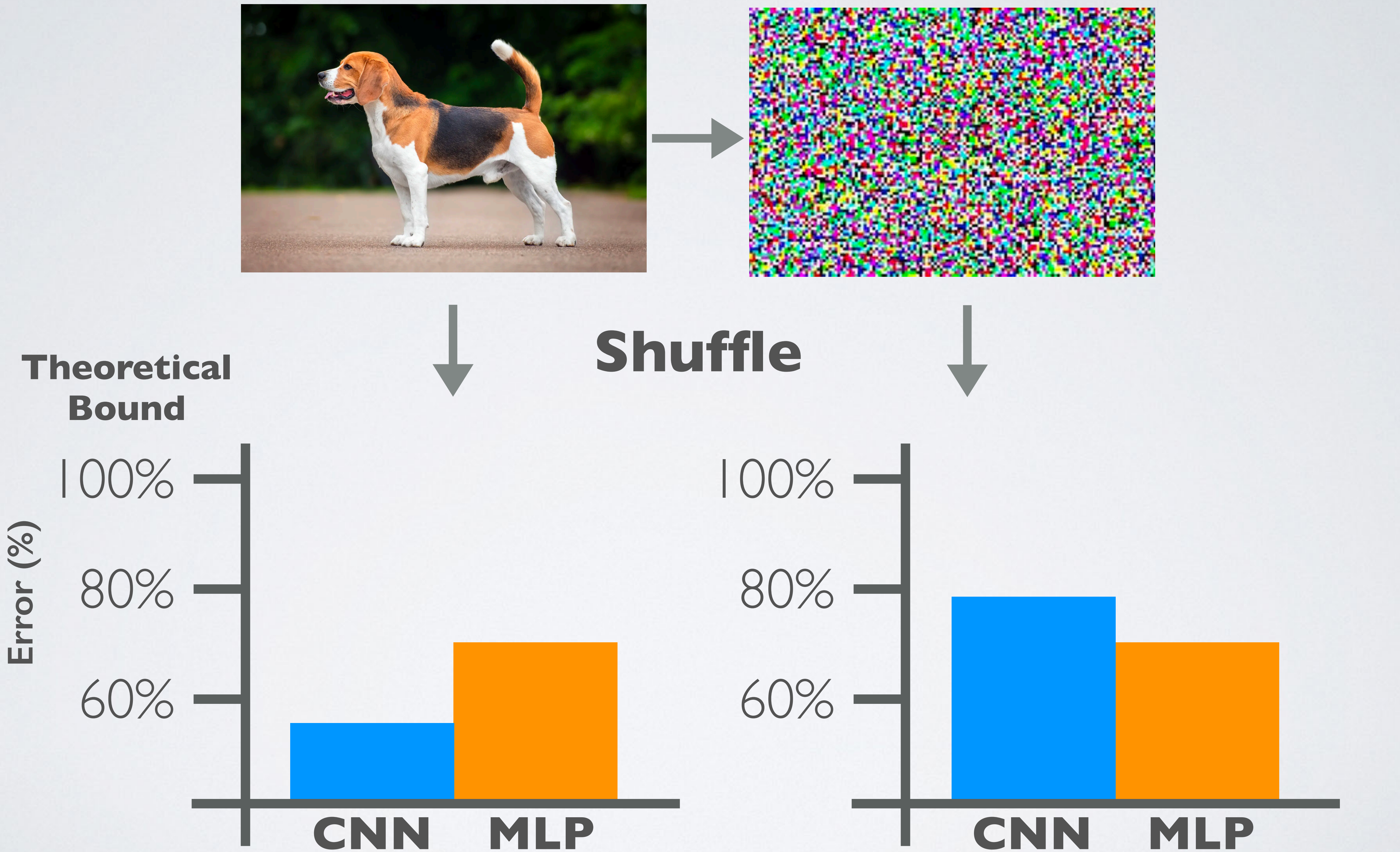


Shuffle

Theoretical Bound



CNNs prefer data with spatial structure



PAC-Bayes Compression Bounds So Tight… NeurIPS '22

The Marginal Likelihood and Generalization

Bayesian Model Selection, the Marginal Likelihood, and Generalization

Outstanding Paper Award - ICML 2022

The Marginal Likelihood

Marginal likelihood: $p(D|M) = \int p(D|M, w)p(w|M)dw$

Bayesian Model Selection, the Marginal Likelihood, and Generalization
Outstanding Paper Award - ICML 2022

The Marginal Likelihood

Marginal likelihood: $p(D|M) = \int p(D|M,w)p(w|M)dw$

Probability that a random draw from the prior generates the training data

Bayesian Model Selection, the Marginal Likelihood, and Generalization
Outstanding Paper Award - ICML 2022

The Marginal Likelihood

Marginal likelihood: $p(D|M) = \int p(D|M,w)p(w|M)dw$

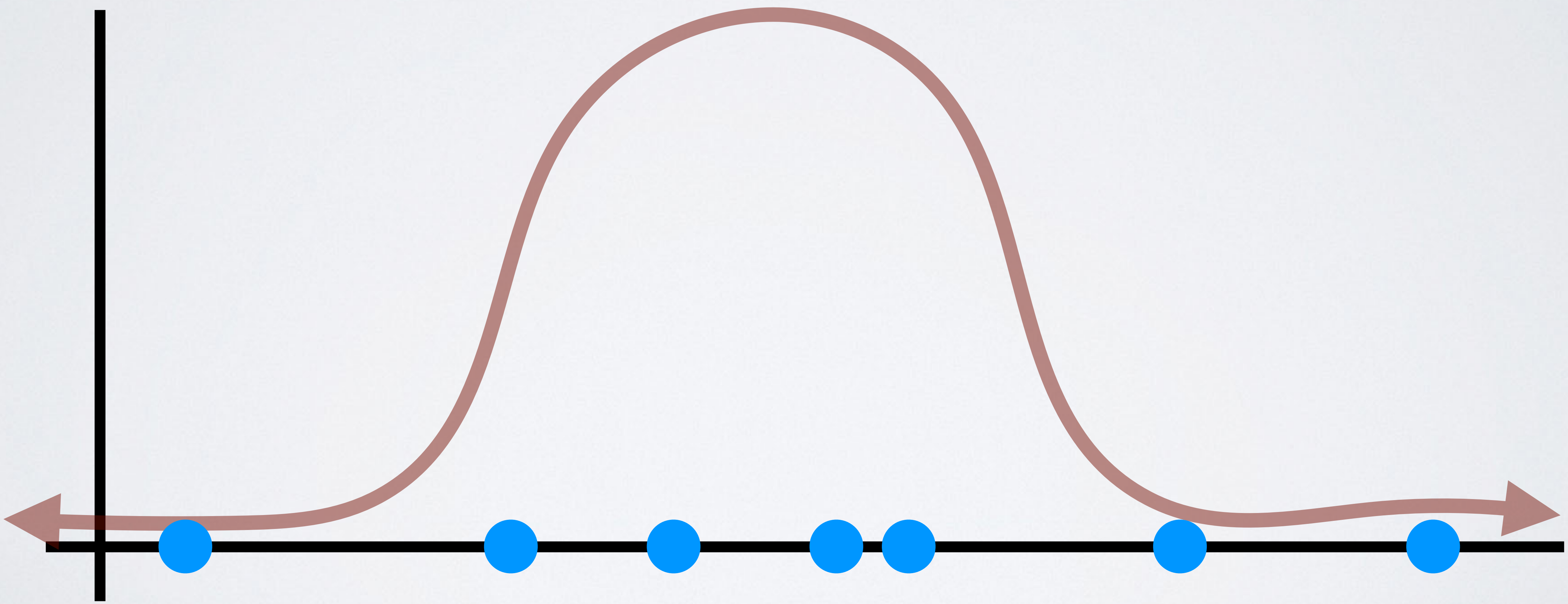
Probability that a random draw from the prior generates the training data

Bayesian Model Selection, the Marginal Likelihood, and Generalization
Outstanding Paper Award - ICML 2022

The Marginal Likelihood

Marginal likelihood: $p(D|M) = \int p(D|M,w)p(w|M)dw$

Probability that a random draw from the prior generates the training data



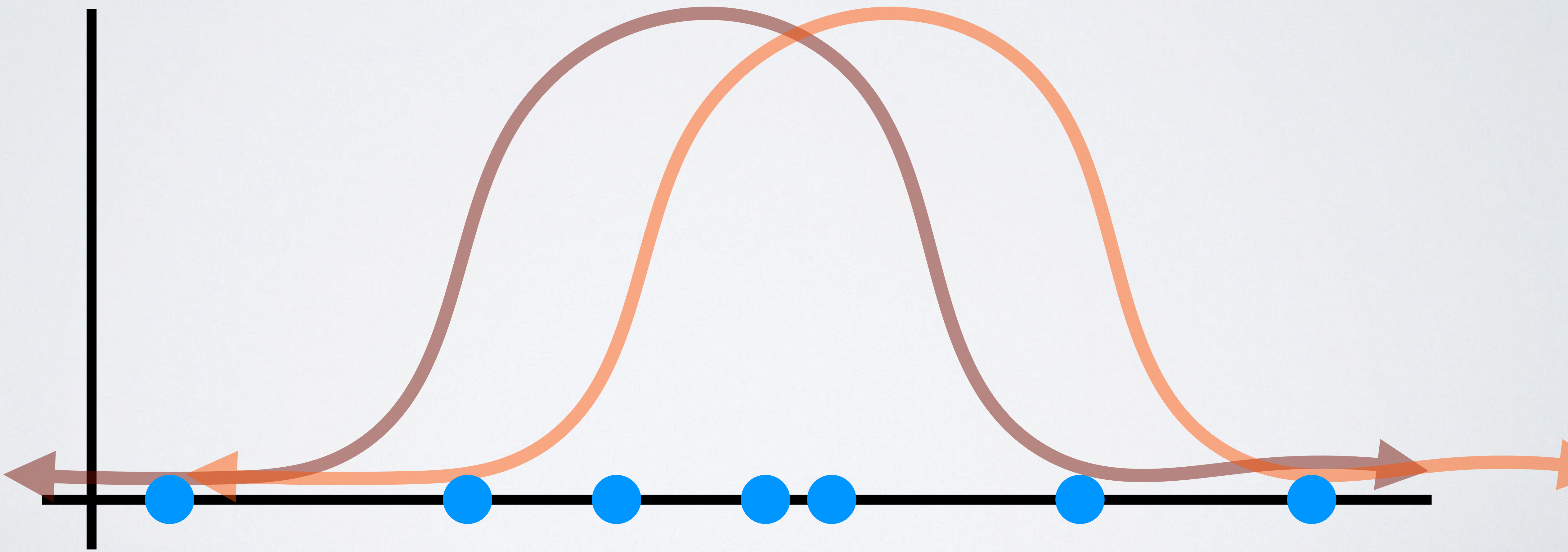
Bayesian Model Selection, the Marginal Likelihood, and Generalization

Outstanding Paper Award - ICML 2022

The Marginal Likelihood

Marginal likelihood: $p(D|M) = \int p(D|M,w)p(w|M)dw$

Probability that a random draw from the prior generates the training data

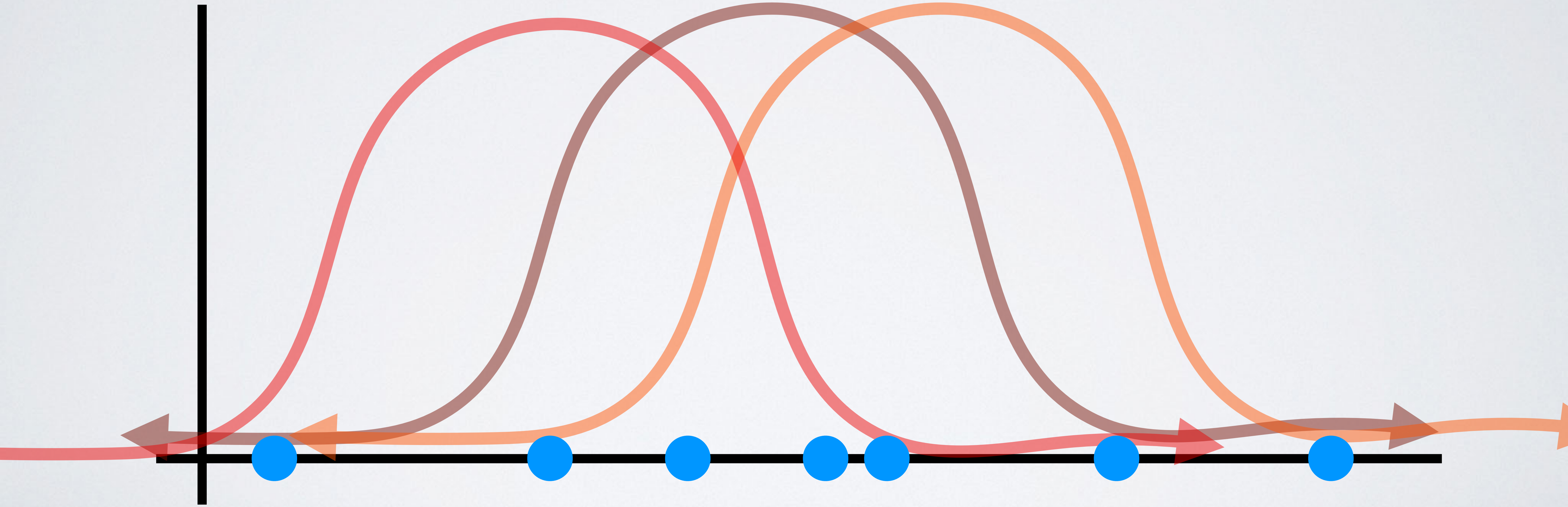


Bayesian Model Selection, the Marginal Likelihood, and Generalization
Outstanding Paper Award - ICML 2022

The Marginal Likelihood

Marginal likelihood: $p(D|M) = \int p(D|M,w)p(w|M)dw$

Probability that a random draw from the prior generates the training data



Bayesian Model Selection, the Marginal Likelihood, and Generalization
Outstanding Paper Award - ICML 2022

The Marginal Likelihood

Marginal likelihood: $p(D|M) = \int p(D|M,w)p(w|M)dw$

Probability that a random draw from the prior generates the training data

- Model selection

Bayesian Model Selection, the Marginal Likelihood, and Generalization
Outstanding Paper Award - ICML 2022

The Marginal Likelihood

Marginal likelihood: $p(D|M) = \int p(D|M,w)p(w|M)dw$

Probability that a random draw from the prior generates the training data

- Model selection
- Hyperparameter tuning

Bayesian Model Selection, the Marginal Likelihood, and Generalization
Outstanding Paper Award - ICML 2022

The Marginal Likelihood

Marginal likelihood: $p(D|M) = \int p(D|M,w)p(w|M)dw$

Probability that a random draw from the prior generates the training data

- Model selection
- Hyperparameter tuning
- Hypothesis testing

Bayesian Model Selection, the Marginal Likelihood, and Generalization
Outstanding Paper Award - ICML 2022

Trouble in Bayesian-Land

But…

can fail to predict generalization

Bayesian Model Selection, the Marginal Likelihood, and Generalization

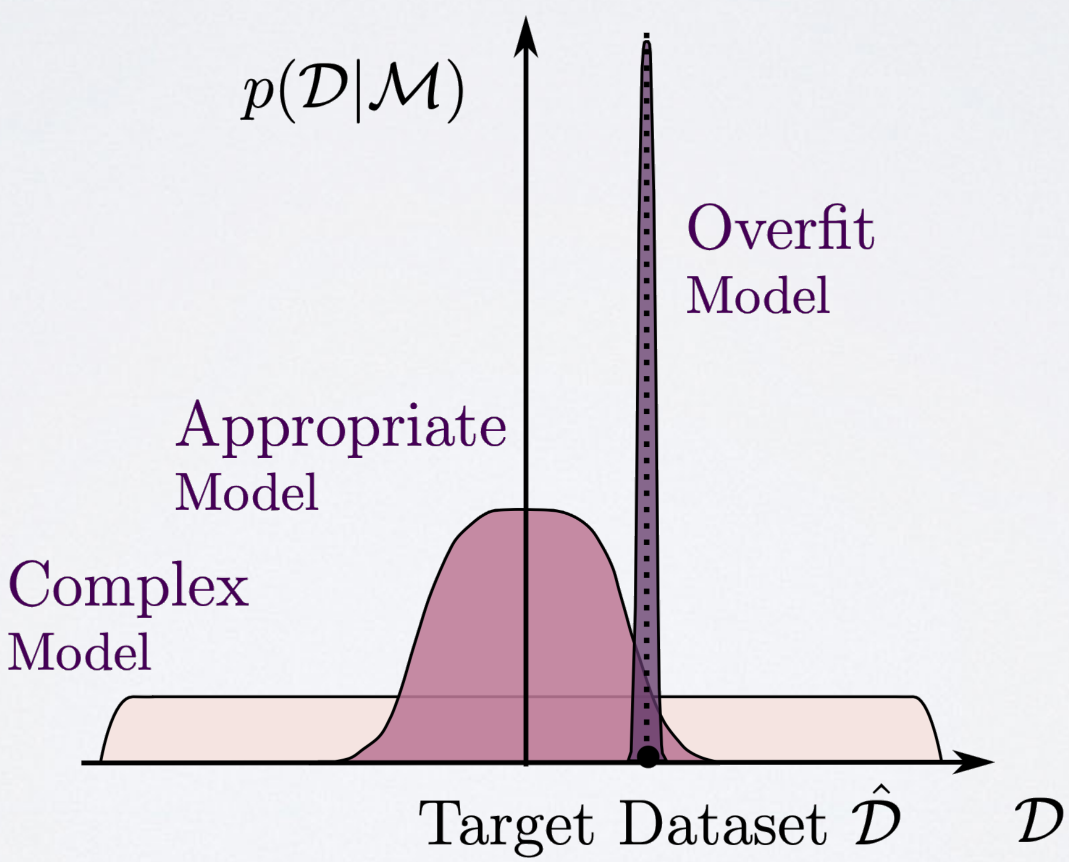
Outstanding Paper Award - ICML 2022

Trouble in Bayesian-Land

But…

can fail to predict generalization

Overfitting



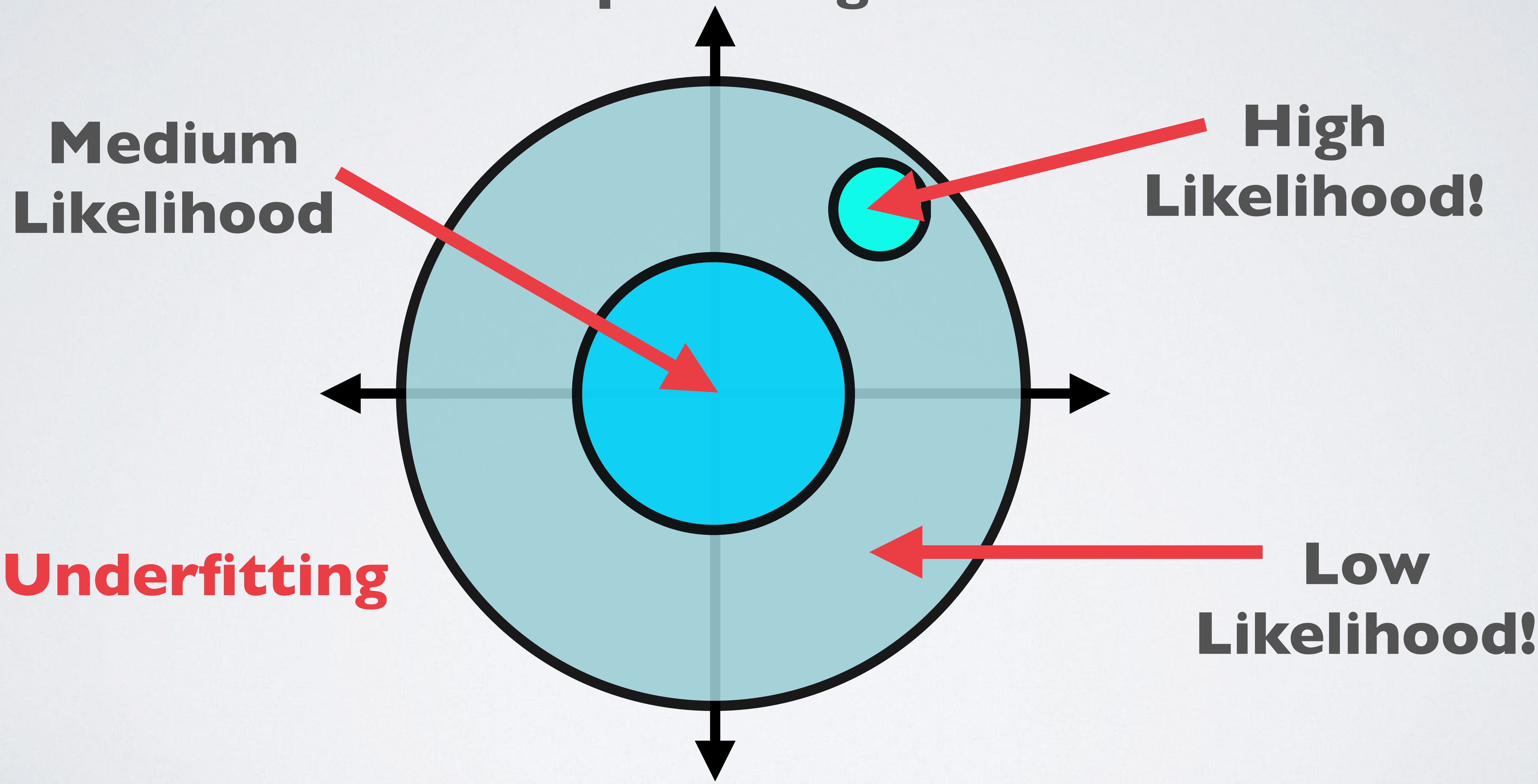
Bayesian Model Selection, the Marginal Likelihood, and Generalization

Outstanding Paper Award - ICML 2022

Trouble in Bayesian-Land

But…

can fail to predict generalization



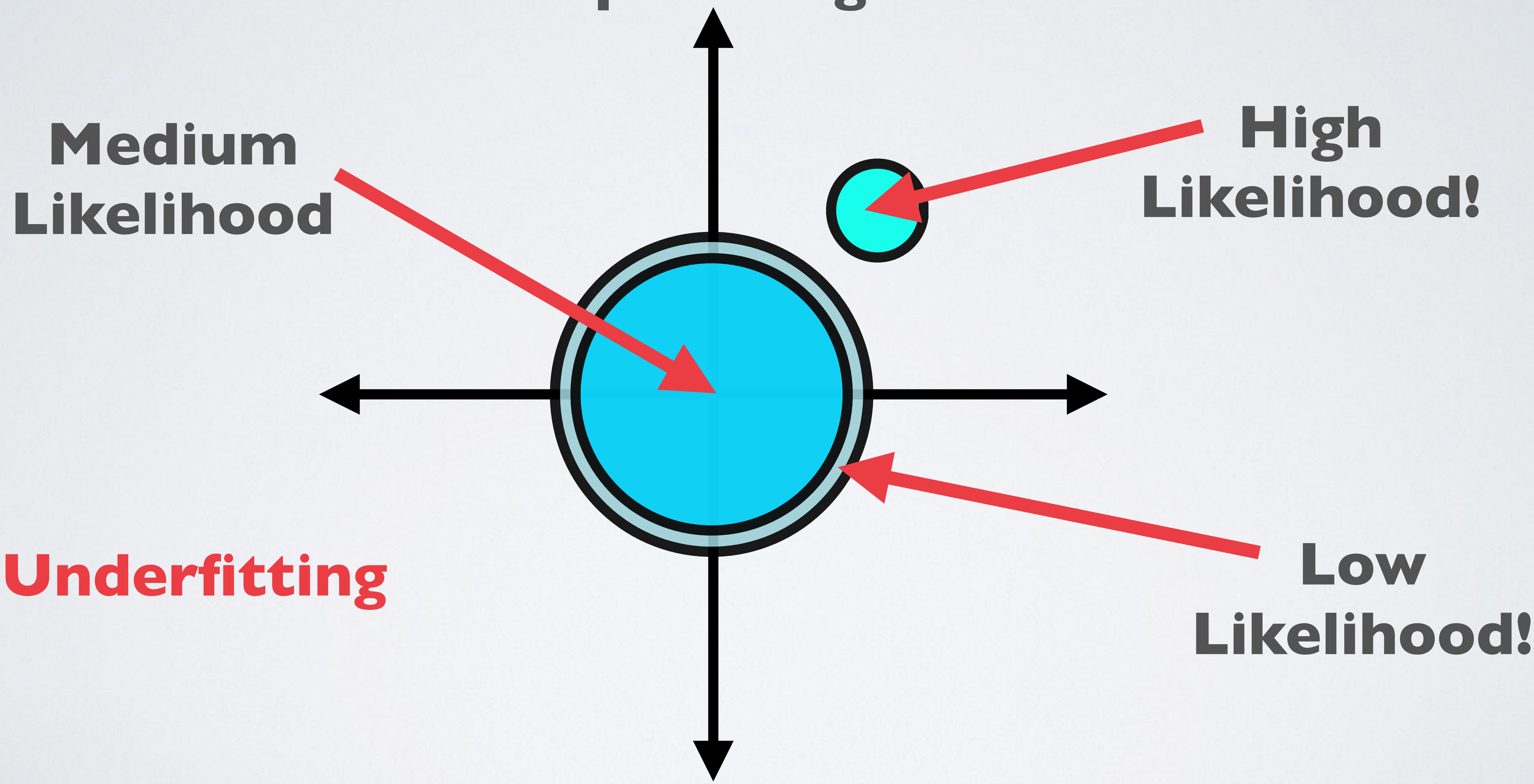
Bayesian Model Selection, the Marginal Likelihood, and Generalization

Outstanding Paper Award - ICML 2022

Trouble in Bayesian-Land

But…

can fail to predict generalization



Bayesian Model Selection, the Marginal Likelihood, and Generalization

Outstanding Paper Award - ICML 2022

Why does the marginal likelihood fail to predict generalization?

The Marginal Likelihood and PAC-Bayes

- Minimum description length (MacKay 2003)

- Marginal likelihood ⇔ PAC-Bayes bound (Germain et al. 2016)

What does PAC-Bayes say about tuning the prior?

Bayesian Model Selection, the Marginal Likelihood, and Generalization
Outstanding Paper Award - ICML 2022

What does PAC-Bayes say about tuning the prior?

Goal: choose between k models
Which ones generalize better?

What does PAC-Bayes say about tuning the prior?

Goal: choose between k models
Which ones generalize better?

Construct a bound for each model

Bayesian Model Selection, the Marginal Likelihood, and Generalization
Outstanding Paper Award - ICML 2022

What does PAC-Bayes say about tuning the prior?

Goal: choose between k models
Which ones generalize better?

Construct a bound for each model

Probability of bounds holding:
$$1 - \delta \longrightarrow 1 - k\delta$$

What does PAC-Bayes say about tuning the prior?

Construct a bound for each model

Probability of bounds holding:
$$1 - \delta \longrightarrow 1 - k\delta$$

Keep high probability bound, but looser
$$\log(n/\delta) \longrightarrow \log(kn/\delta)$$

$$\mathbb{E}_{h\sim Q}\left[R(h)\right] \leq \mathbb{E}_{h\sim Q}[\hat{R}(h)] + \sqrt{\frac{\mathbb{KL}(Q \parallel P) + \log(kn/\delta) + 2}{2n - 1}}$$

Bayesian Model Selection, the Marginal Likelihood, and Generalization
Outstanding Paper Award - ICML 2022

What does PAC-Bayes say about tuning the prior?

Probability of bounds holding:

$$1 - \delta \longrightarrow 1 - k\delta$$

Keep high probability bound, but looser

$$\log(n/\delta) \longrightarrow \log(kn/\delta)$$

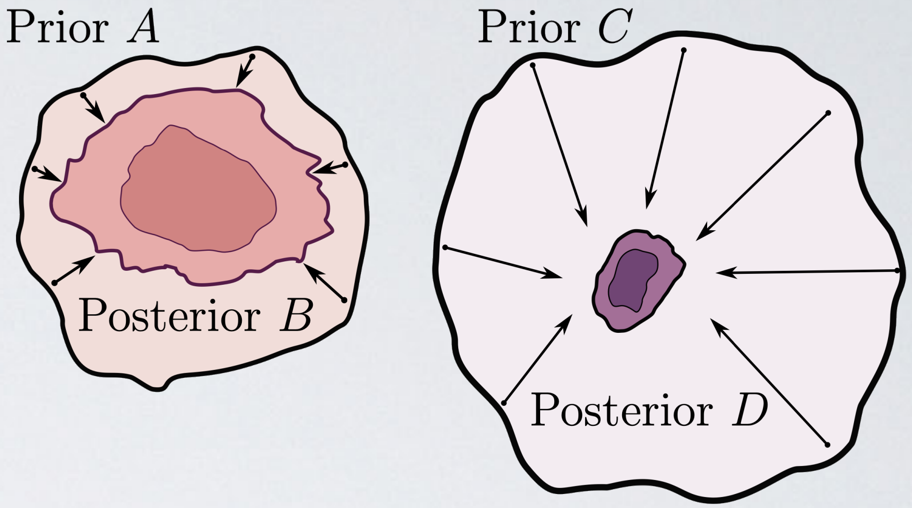
Cost: $\log(k)$

$$\mathbb{E}_{h\sim Q}\left[R(h)\right] \leq \mathbb{E}_{h\sim Q}[\hat{R}(h)] + \sqrt{\frac{\mathbb{KL}(Q \parallel P) + \log(k) + \log(n/\delta) + 2}{2n-1}}$$

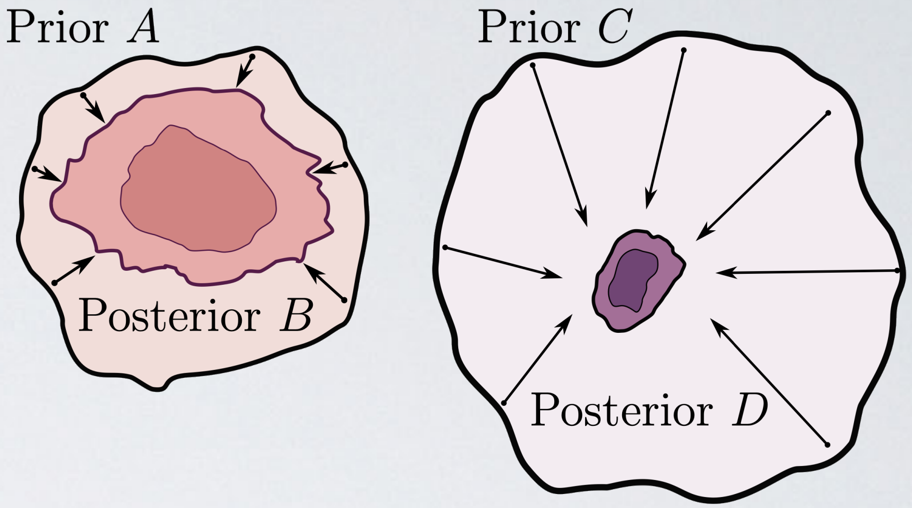
Bayesian Model Selection, the Marginal Likelihood, and Generalization

Outstanding Paper Award - ICML 2022

Underfitting

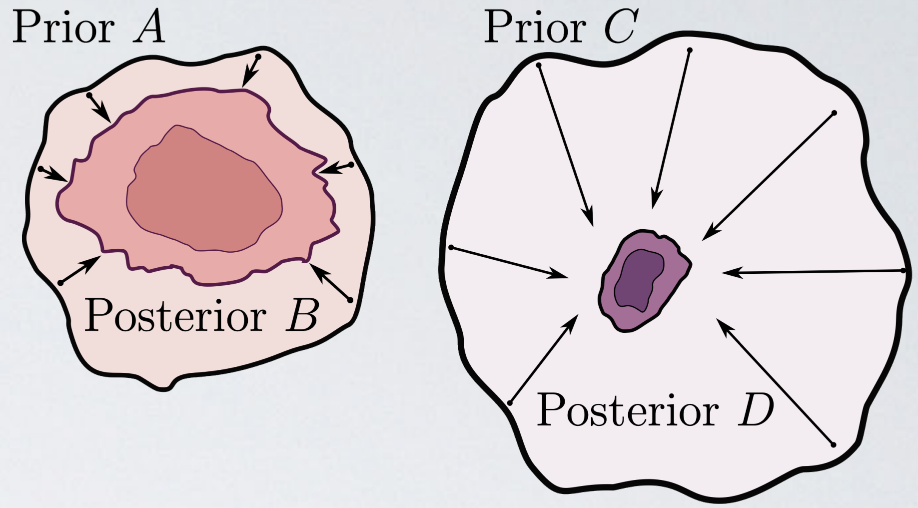


Underfitting



Marginal likelihood hates diffuse priors ⟶ Let prior contract before you measure the likelihood

Underfitting

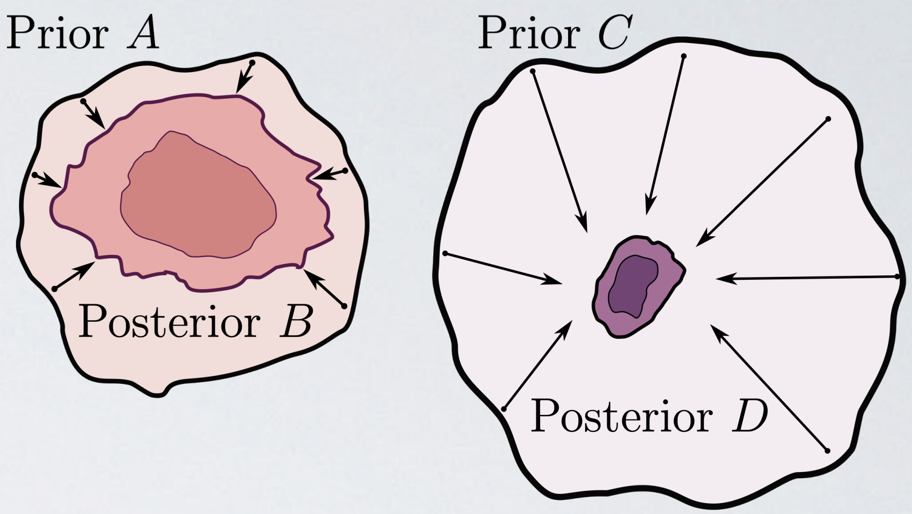


Marginal likelihood hates diffuse priors ⟶ Let prior contract before you measure the likelihood

Conditional marginal likelihood: $p(\mathcal{D}_{\geq m} | \mathcal{D}_{<m})$
Better aligned with generalization

Bayesian Model Selection, the Marginal Likelihood, and Generalization
Outstanding Paper Award - ICML 2022

Underfitting



Marginal likelihood hates diffuse priors $\longrightarrow$ Let prior contract before you measure the likelihood

Conditional marginal likelihood: $p(\mathcal{D}_{\geq m} | \mathcal{D}_{<m})$
Better aligned with generalization

↕

Sharper PAC-Bayes bounds via data-dependent priors (Dziugaite et al. 2020)

Bayesian Model Selection, the Marginal Likelihood, and Generalization
Outstanding Paper Award - ICML 2022

Wrap Up

- Neural networks admit simple solutions, despite having so many parameters.

Wrap Up

- Neural networks admit simple solutions, despite having so many parameters.

- Generalization bounds can predict generalization phenomena or problems with marginal likelihood.

Wrap Up

- Neural networks admit simple solutions, despite having so many parameters.

- Generalization bounds can predict generalization phenomena or problems with marginal likelihood.

- Can generalization theory inform deep learning in practice?

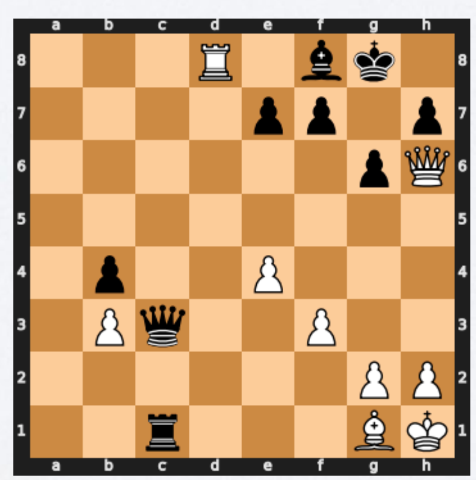
Wrap Up

- Neural networks admit simple solutions, despite having so many parameters.

- Generalization bounds can predict generalization phenomena or problems with marginal likelihood.

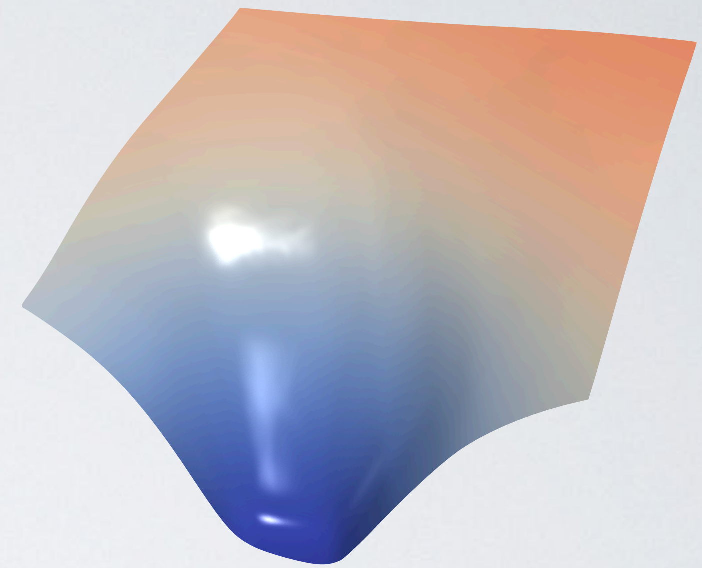
- Can generalization theory inform deep learning in practice?

How far can we push generalization?



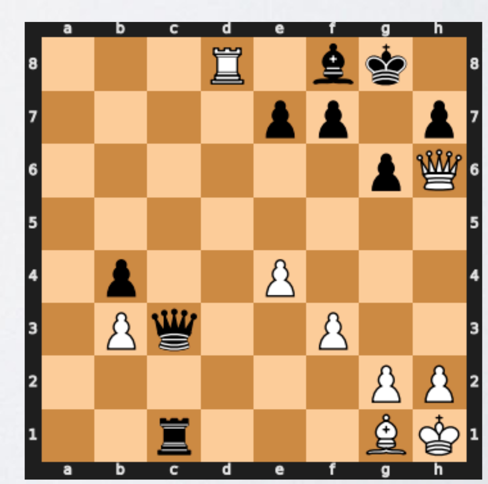
Why do neural networks work?

What are the properties of good minima and why do optimizers find them?



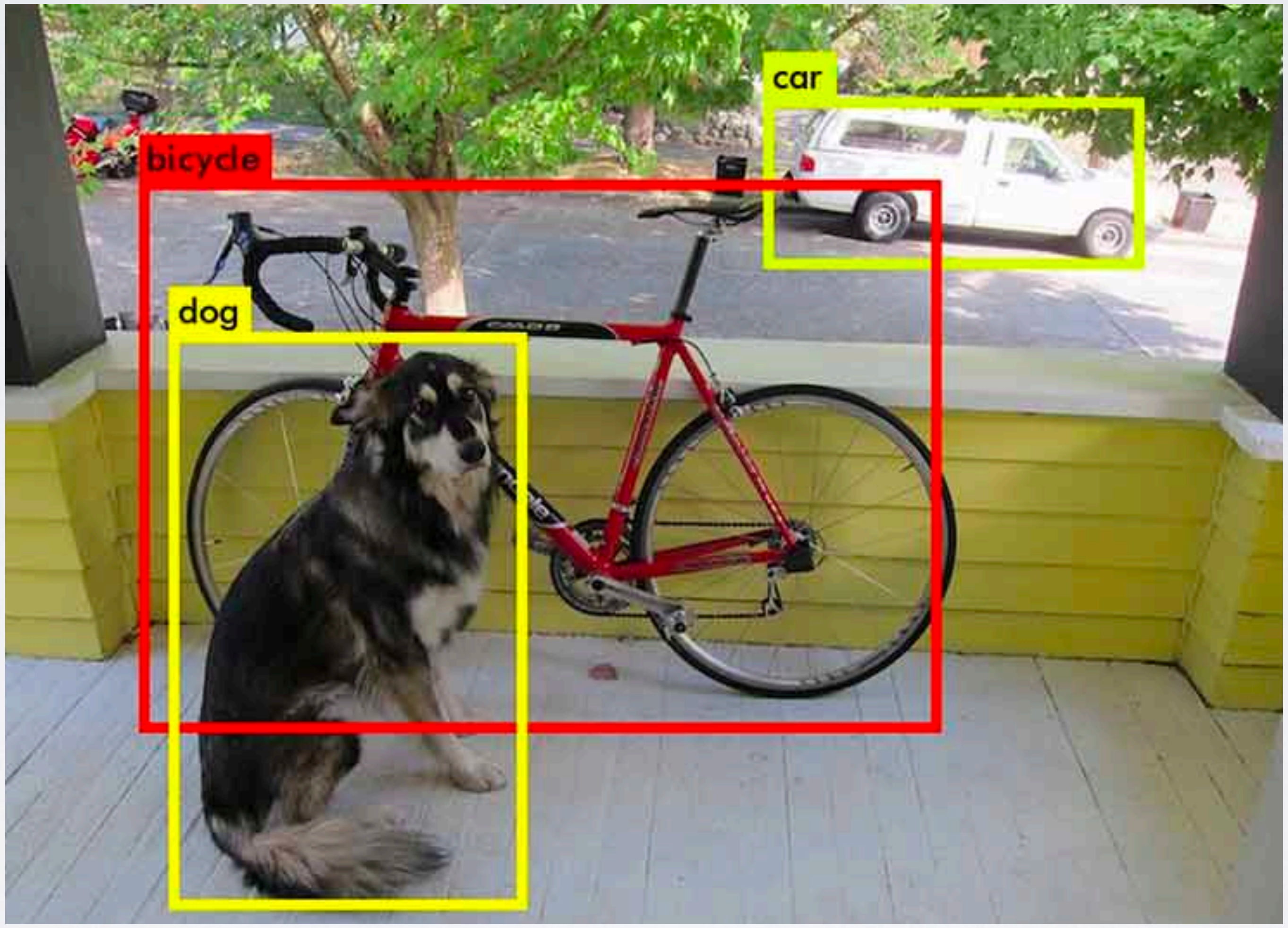
Theories that predict generalization

Observing generalization in reasoning problems



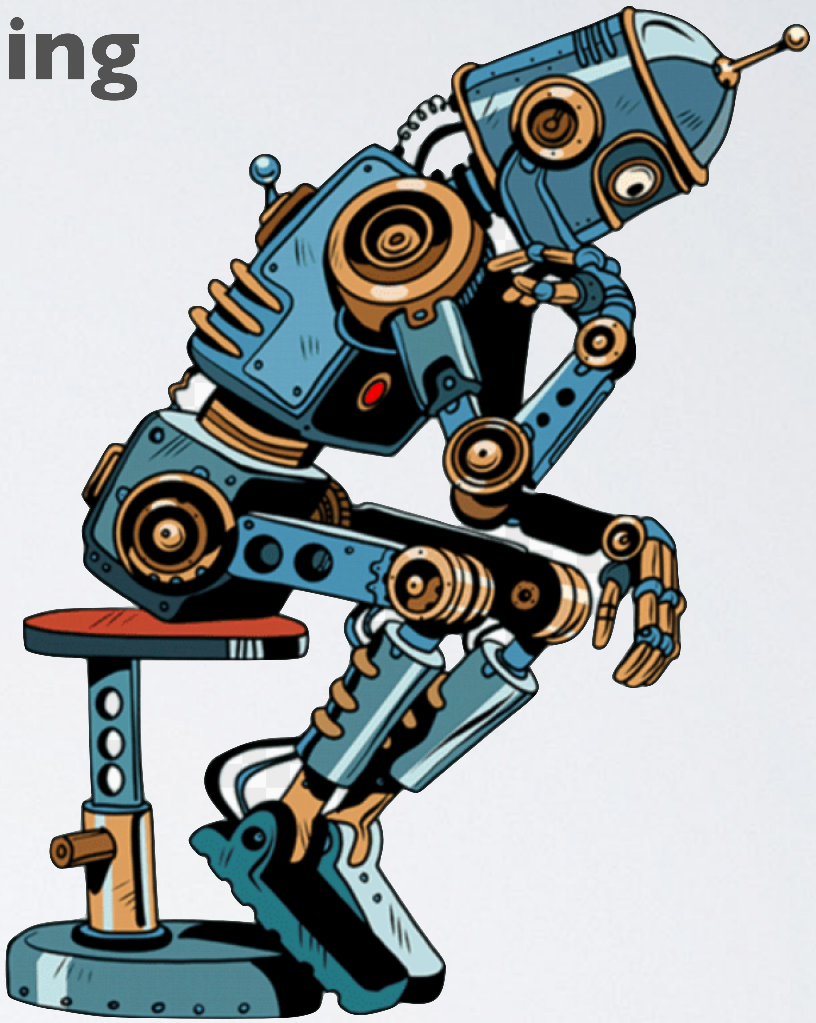
Machines are better than humans at…

Pattern matching



Humans are better than machines at...

Logical reasoning



Proof writing

Causality determination

Domain shift

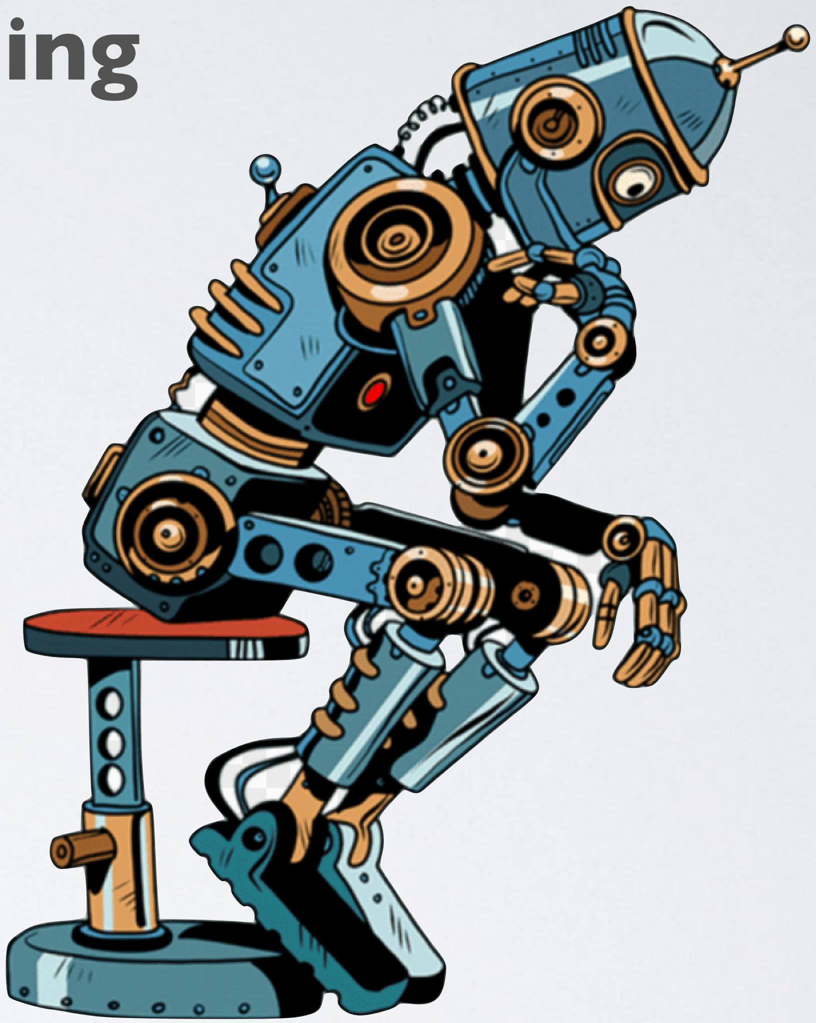
Humans are better than machines at…

Logical reasoning

Proof writing

Causality determination

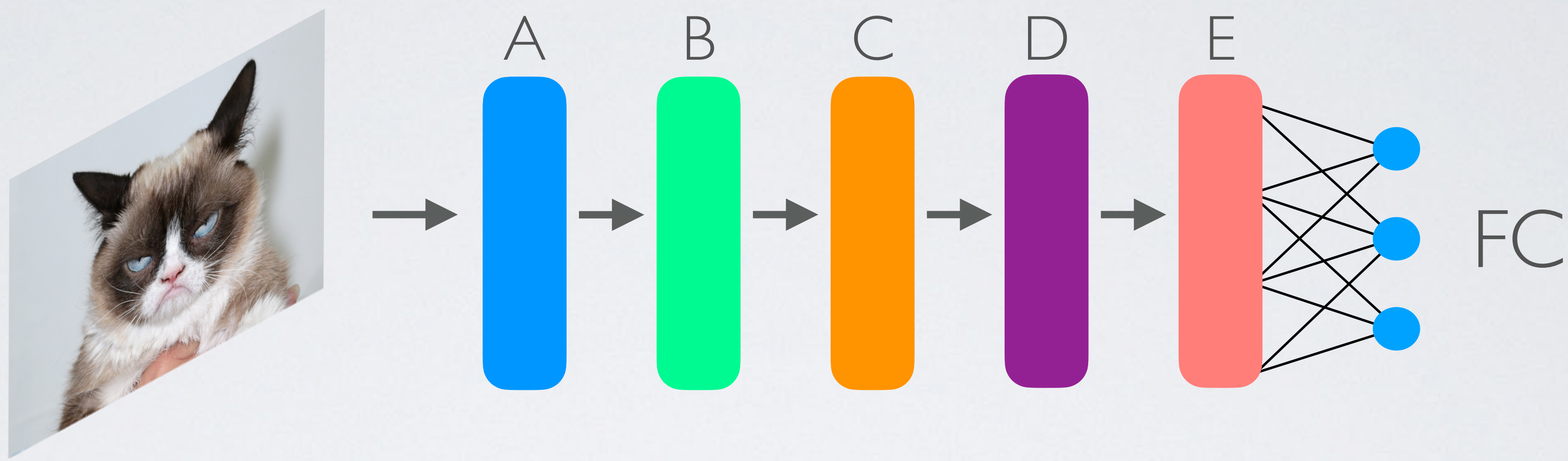
Domain shift



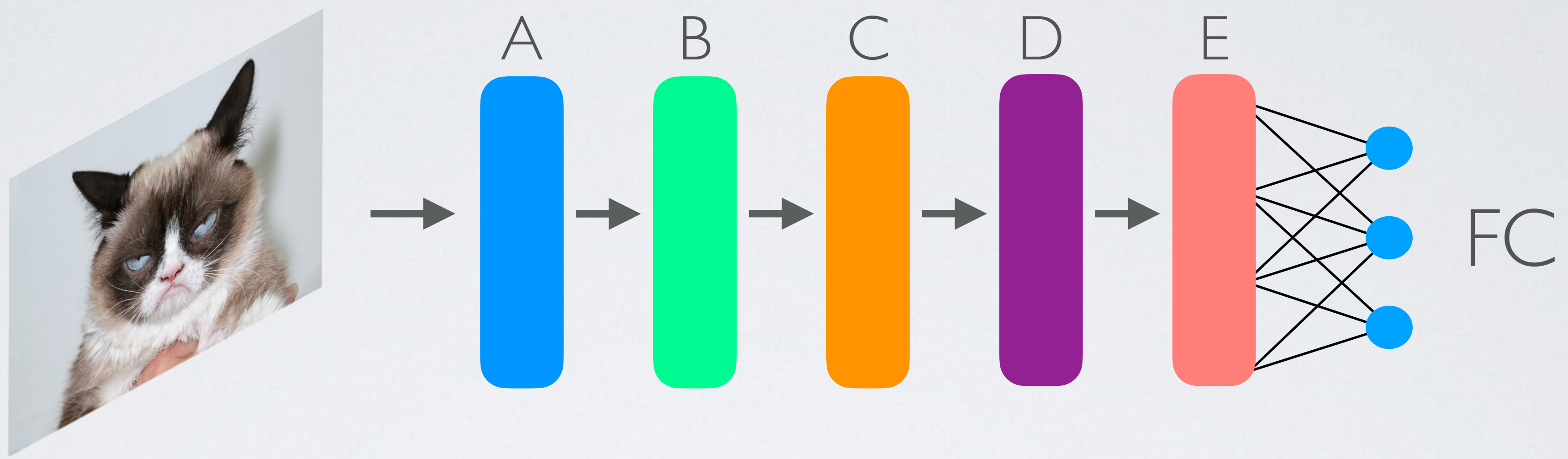
Solve problems of higher complexity by “thinking for longer”

Getting started: replace feed-forward computation with recurrence

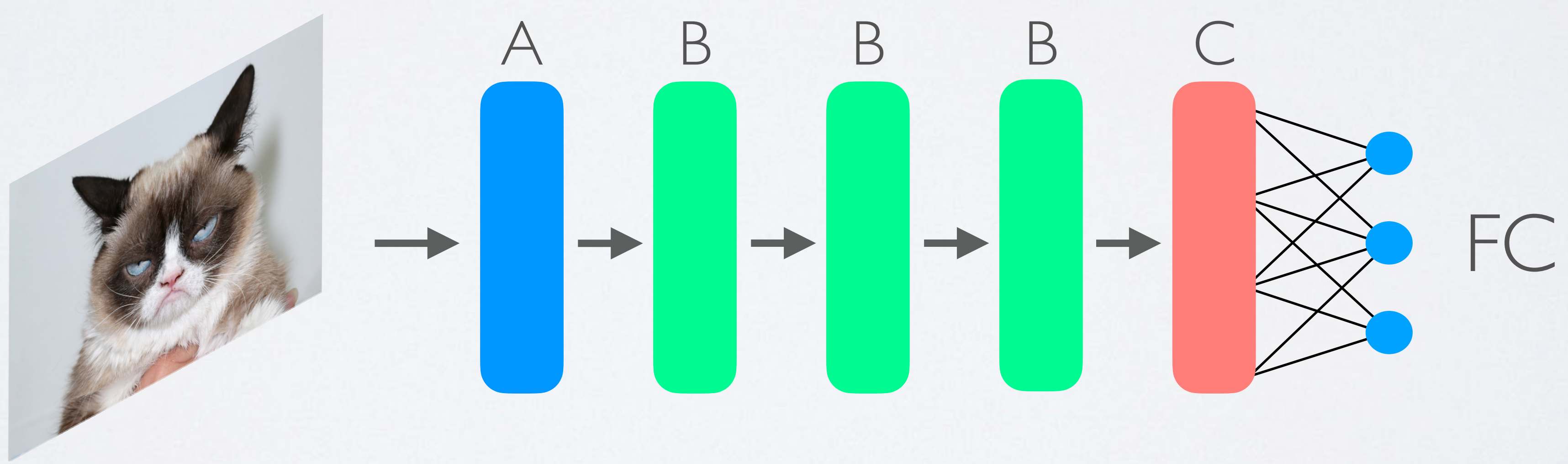
Feed-forward model



Feed-forward model



Recurrent model



Can recurrent nets extrapolate knowledge by “thinking”?

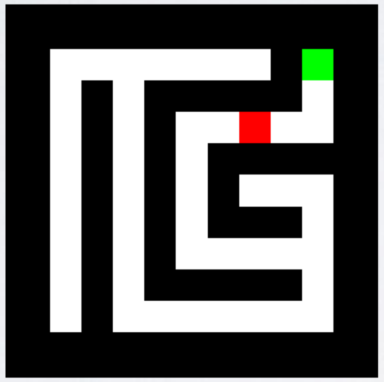
Procedurally generated mazes

Train on this.

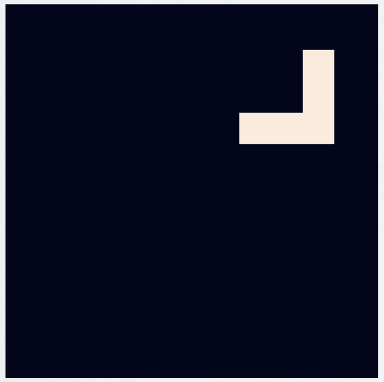
↓

9x9

inputs



labels



Procedurally generated mazes

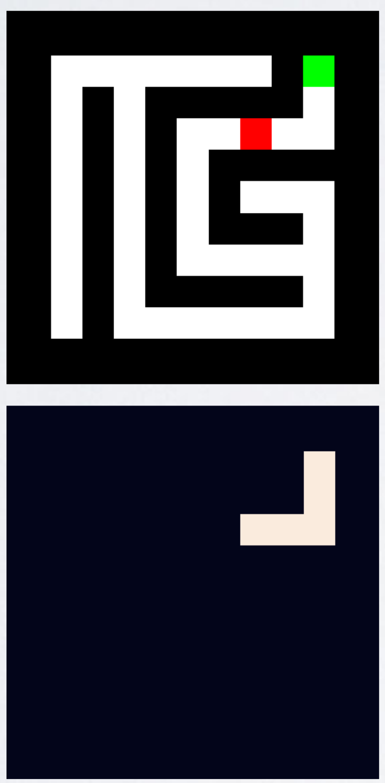
Train on this.

↓

9x9

inputs

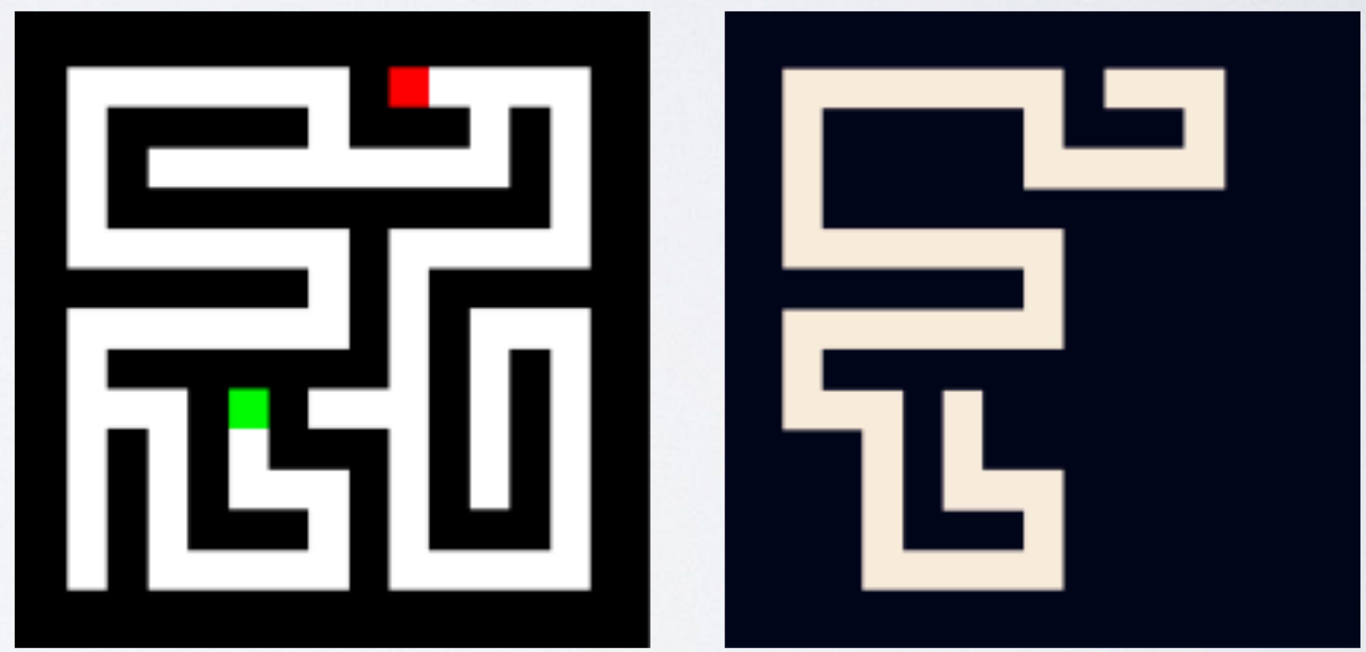
labels



Test on this.

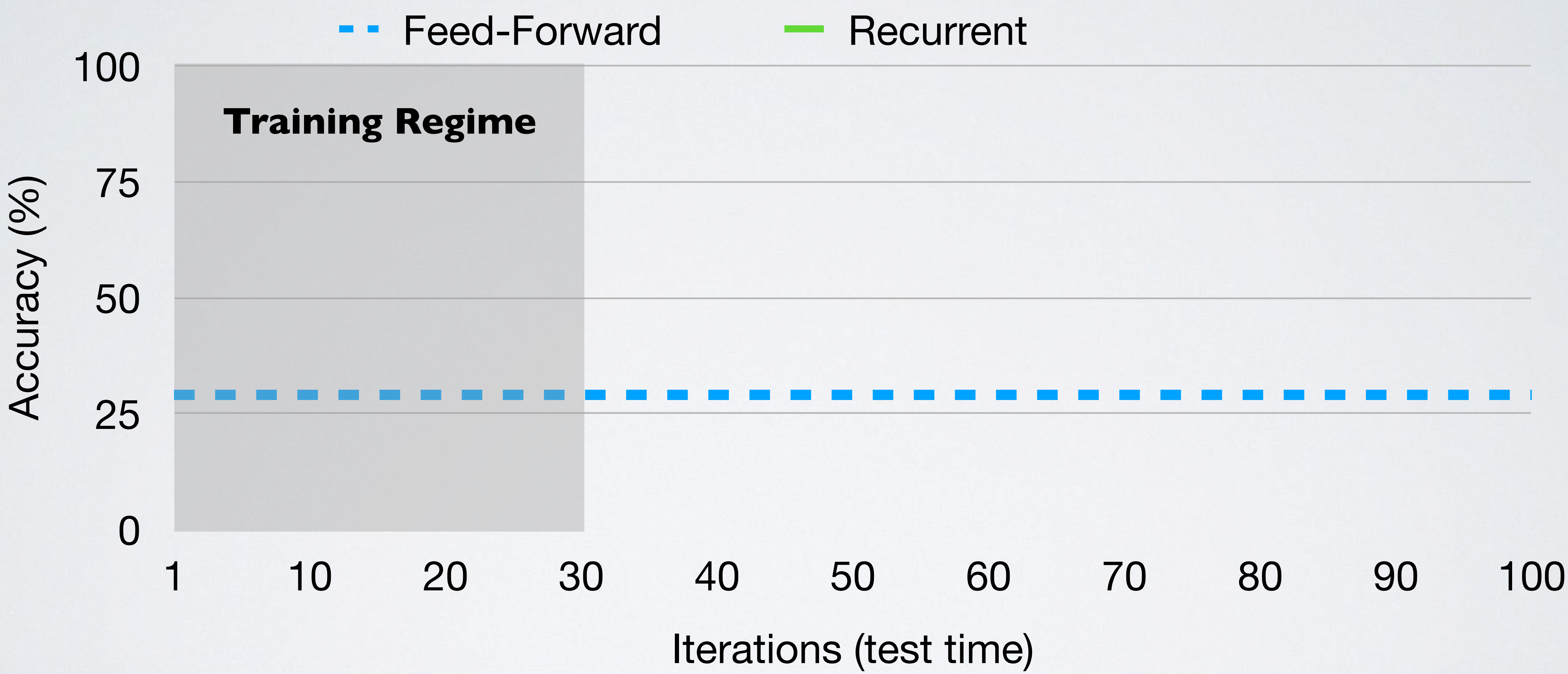
↓

13x13

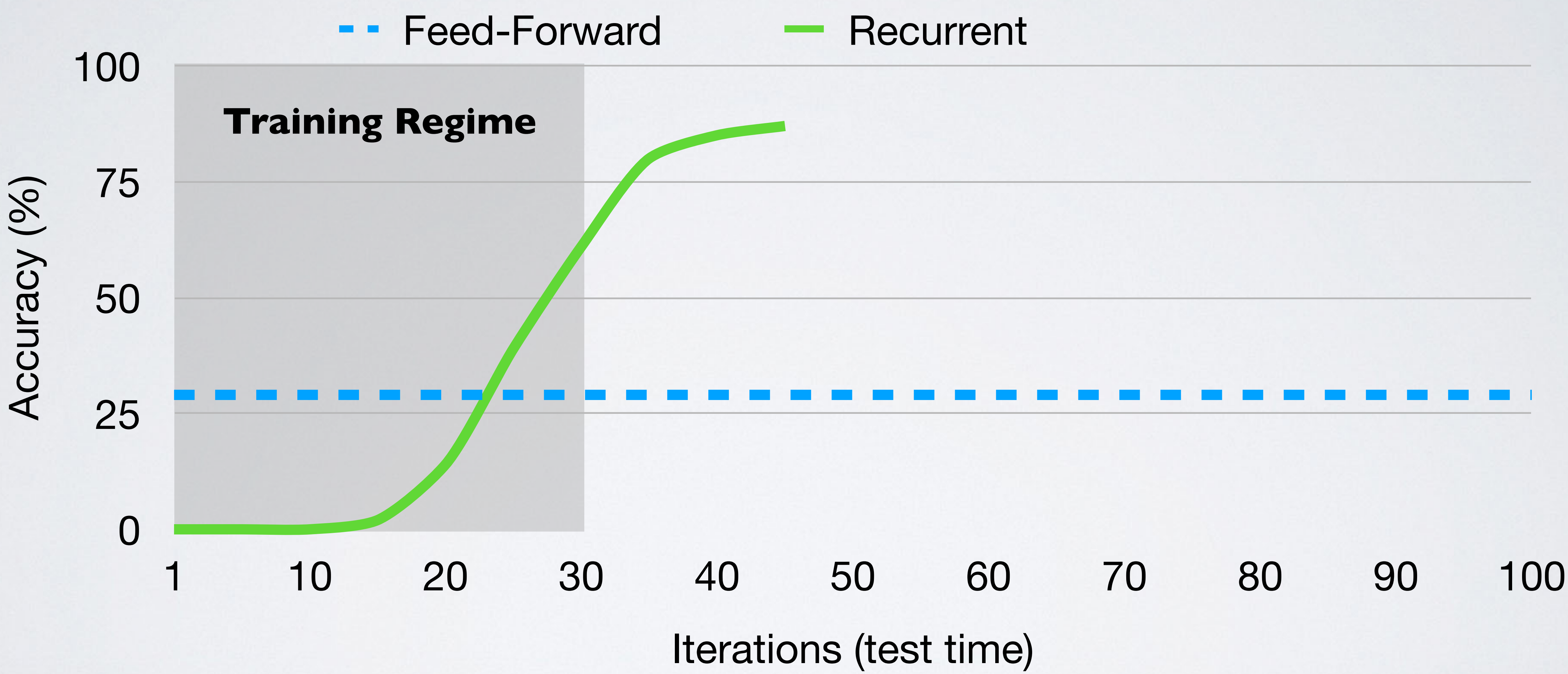


Can You Learn an Algorithm? NeurIPS ‘21

Train on 9x9 → Test on 13x13

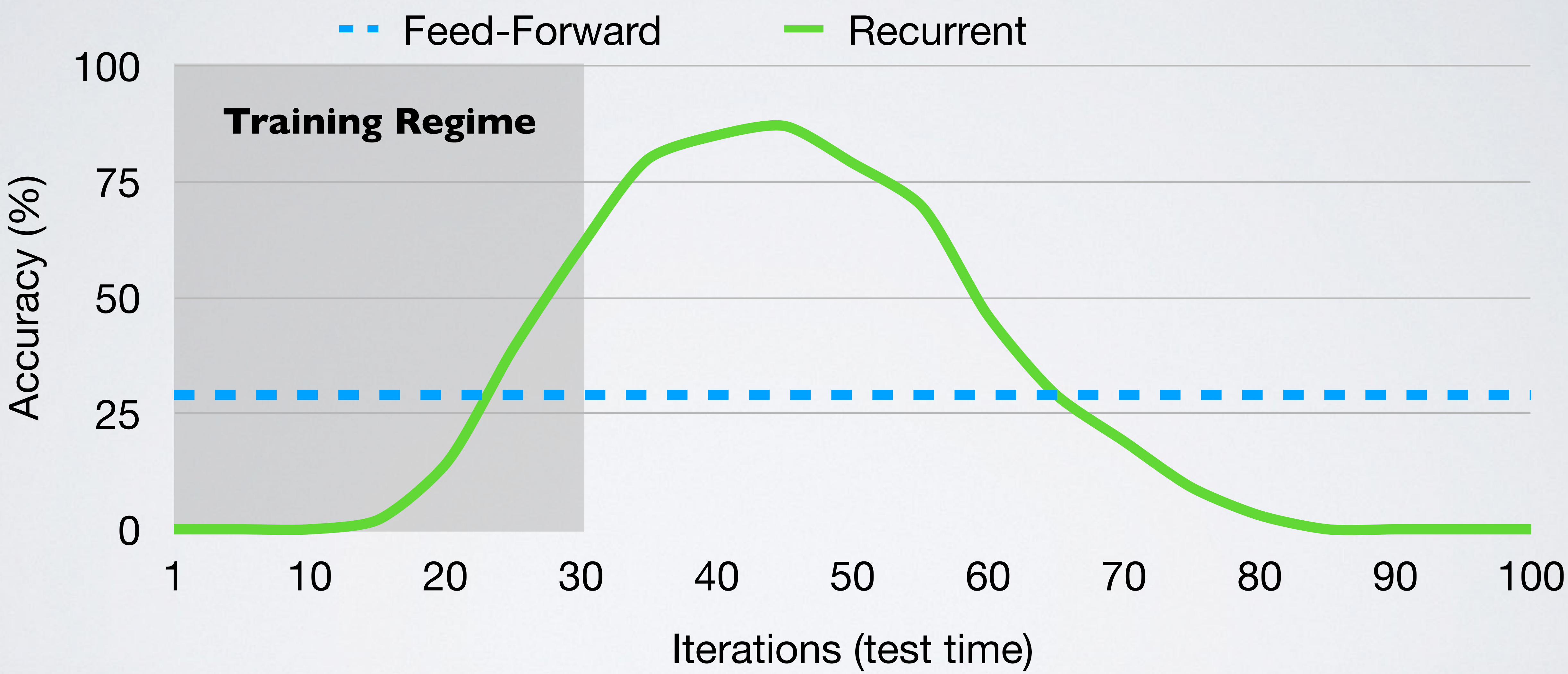


Train on 9x9 → Test on 13x13



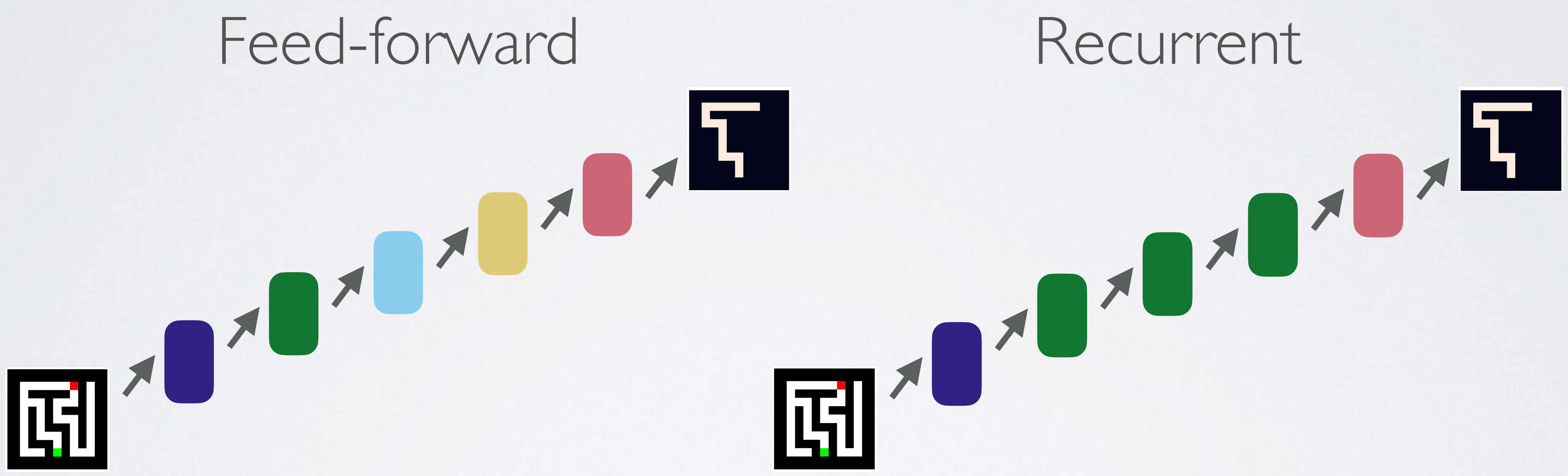
Can You Learn an Algorithm? NeurIPS '21

Train on 9x9 → Test on 13x13



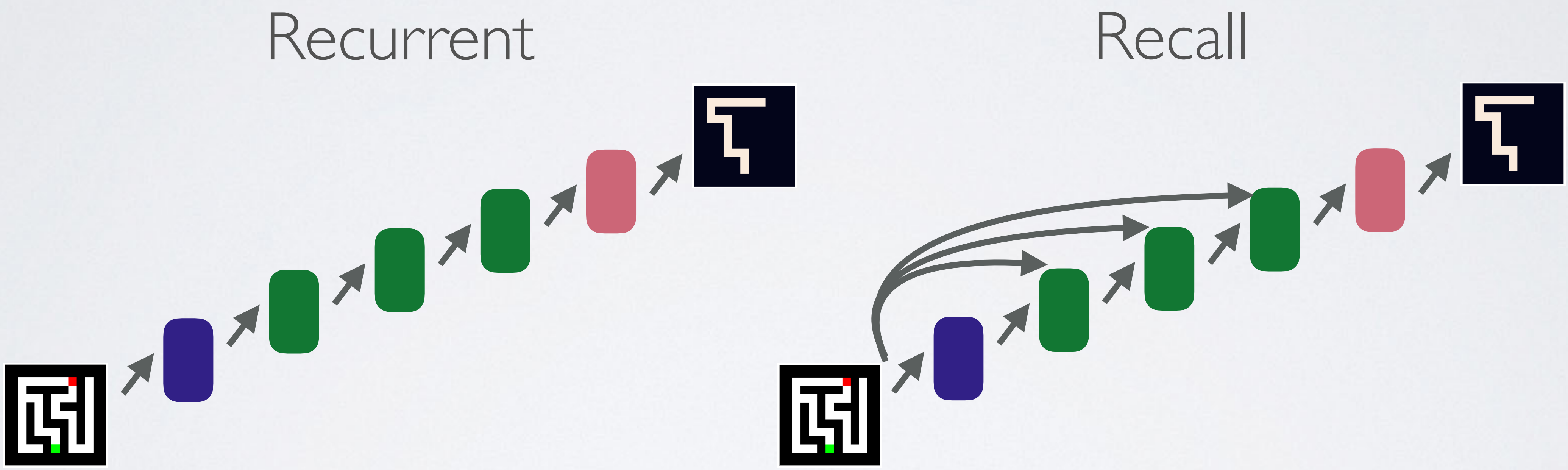
Can You Learn an Algorithm? NeurIPS '21

Architecture Improvement



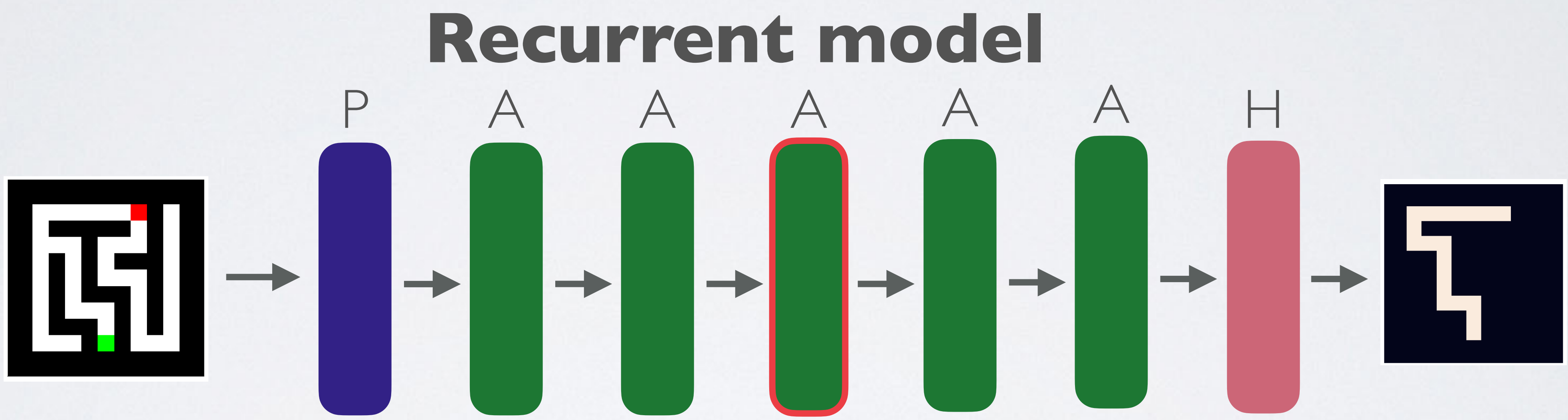
End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22

Architecture Improvement



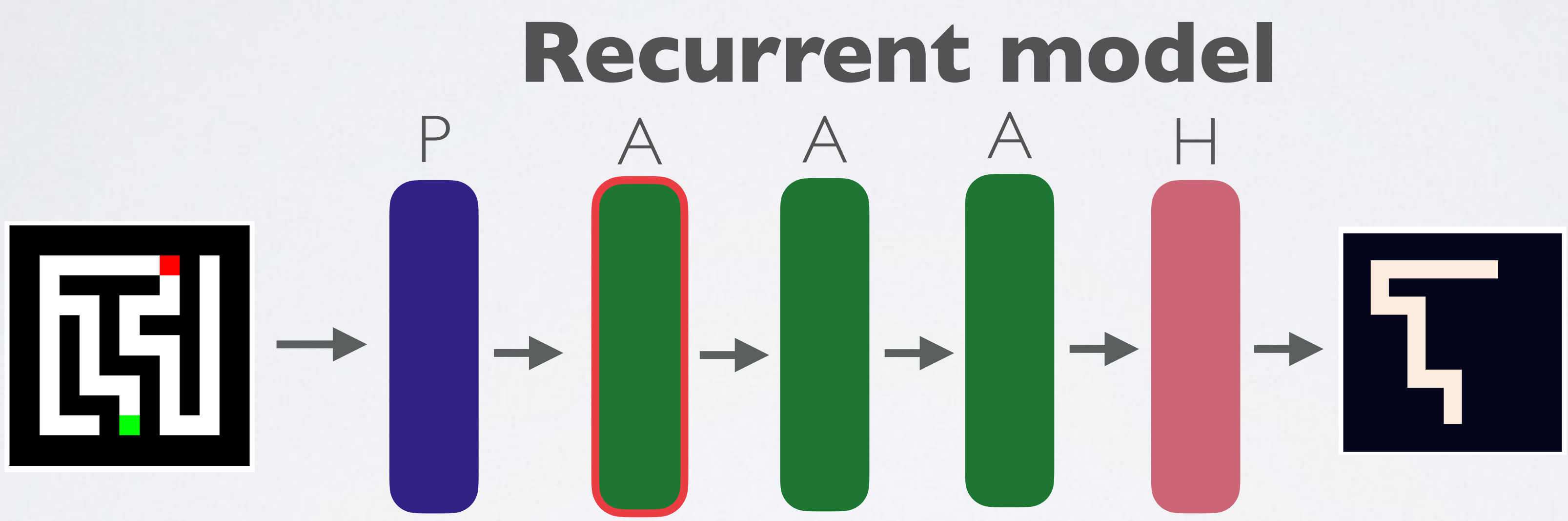
End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22

Incremental Training



End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22

Incremental Training

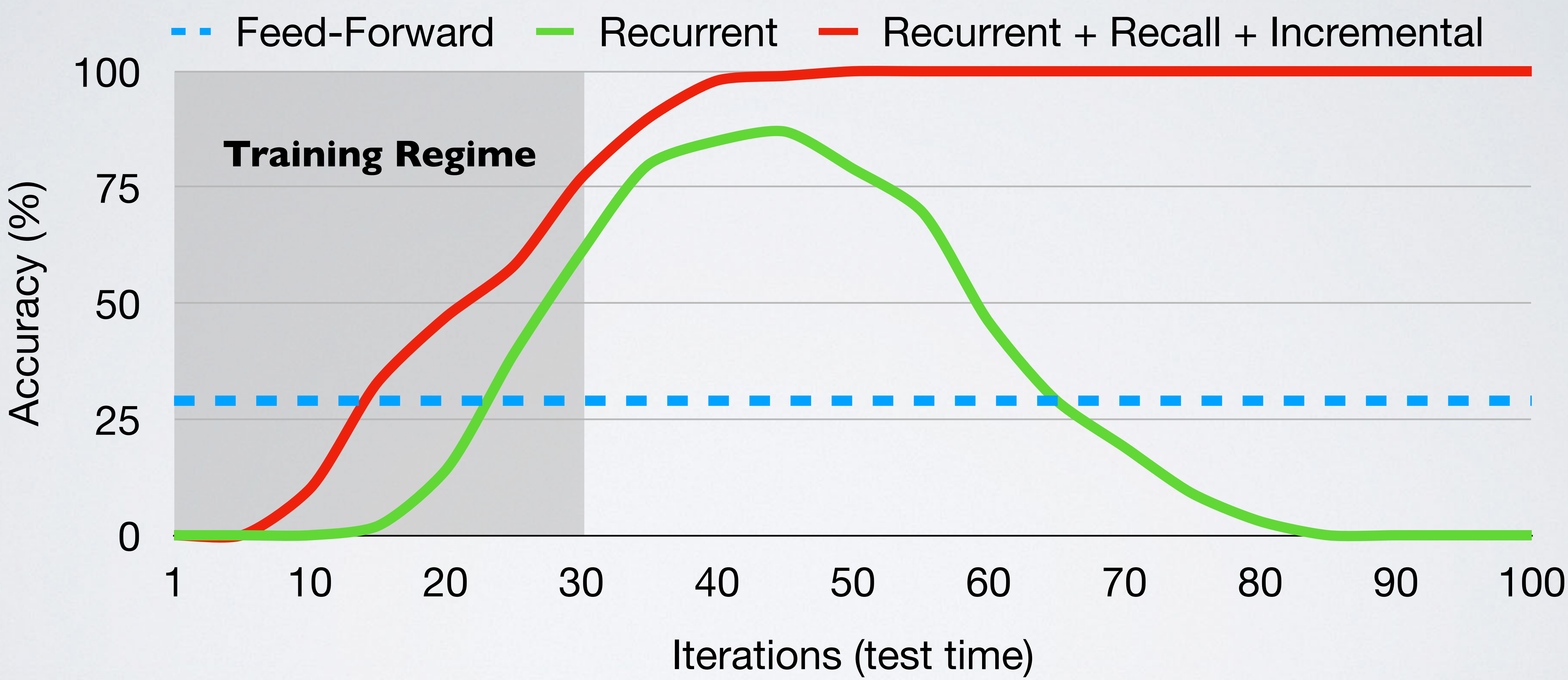


End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22

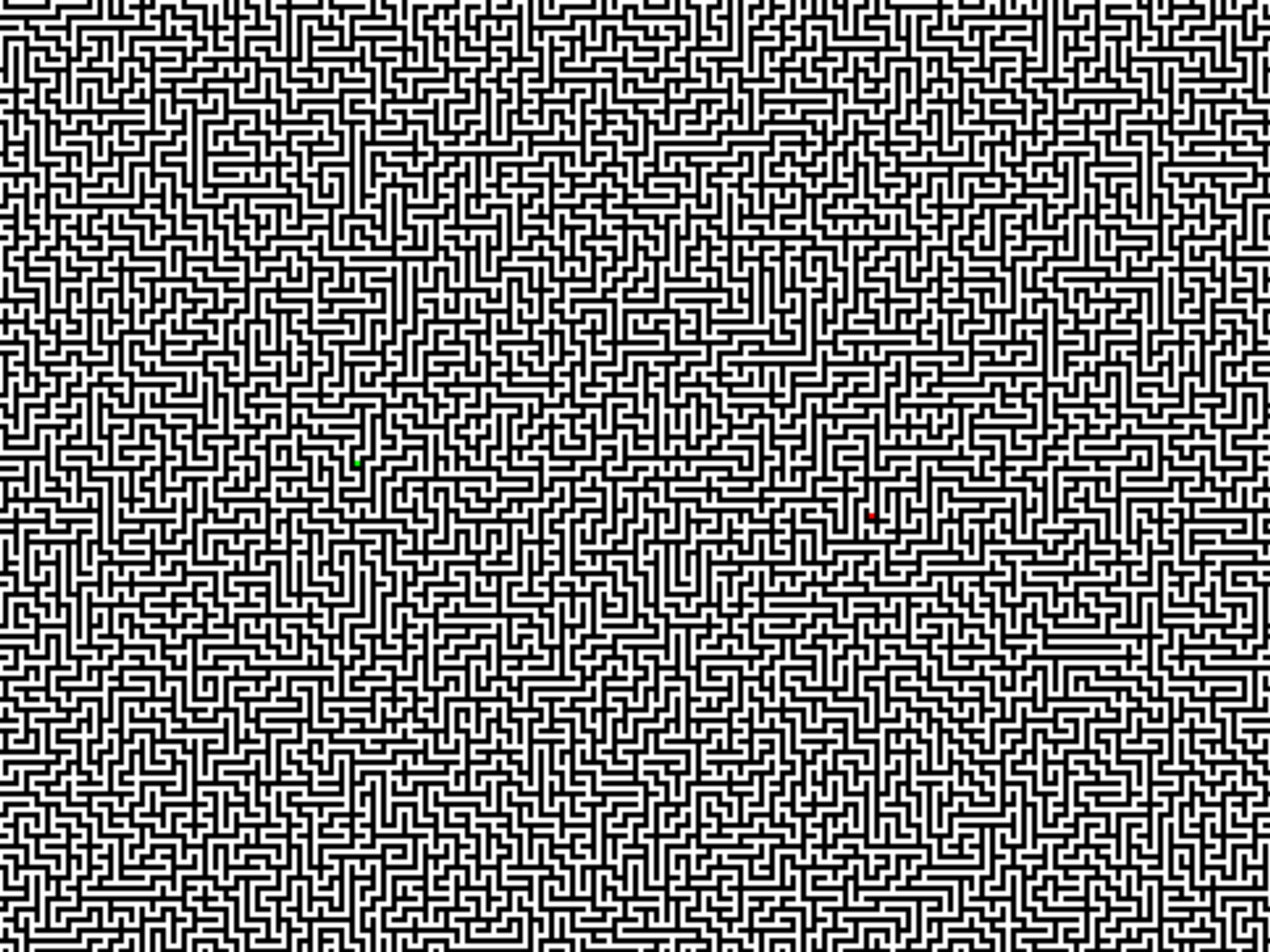
Train on 9x9 → Test on 13x13

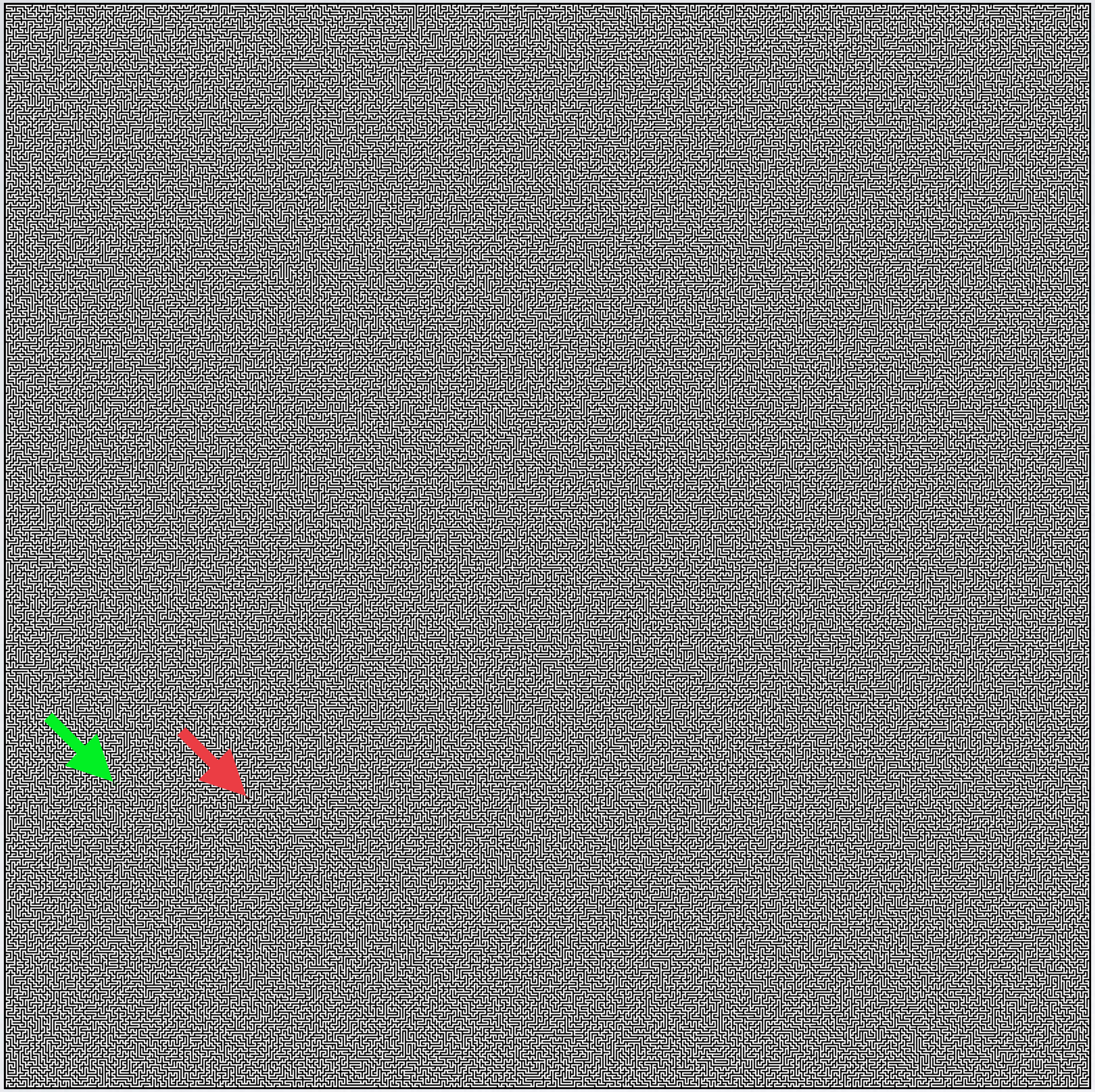
End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22

Train on 9x9 → Test on 13x13



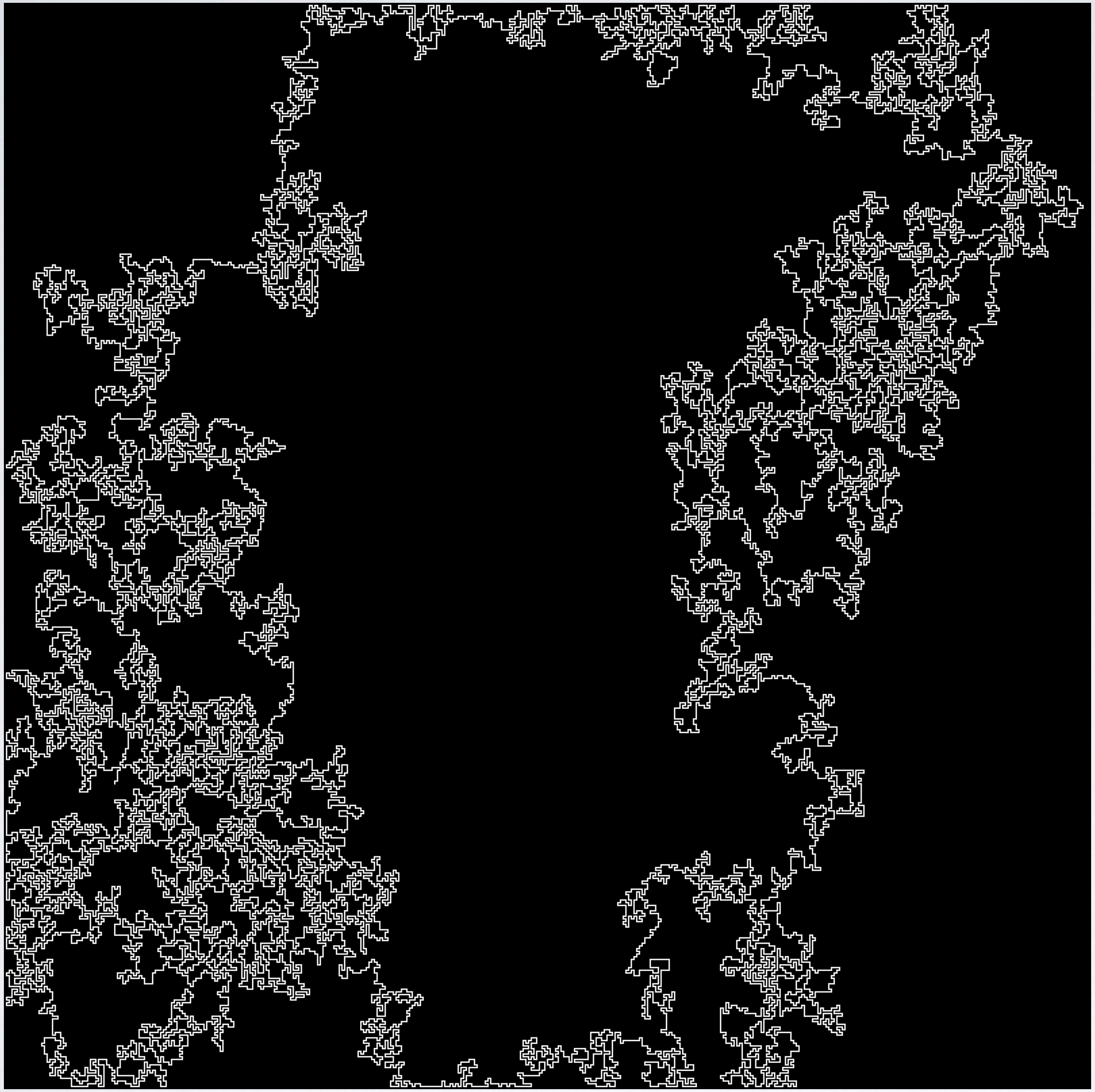
End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22





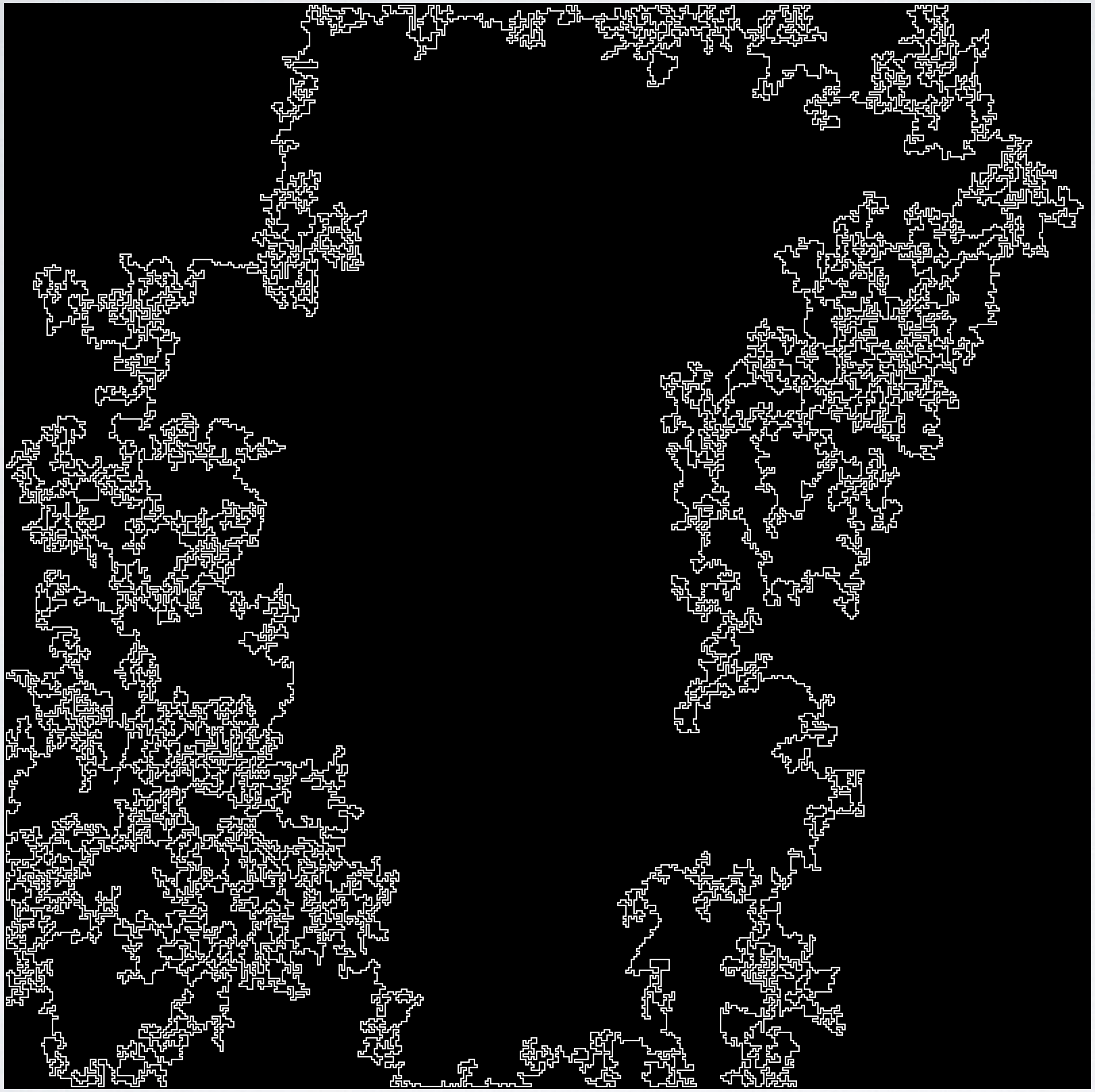
801x801

End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22



801x801

End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22



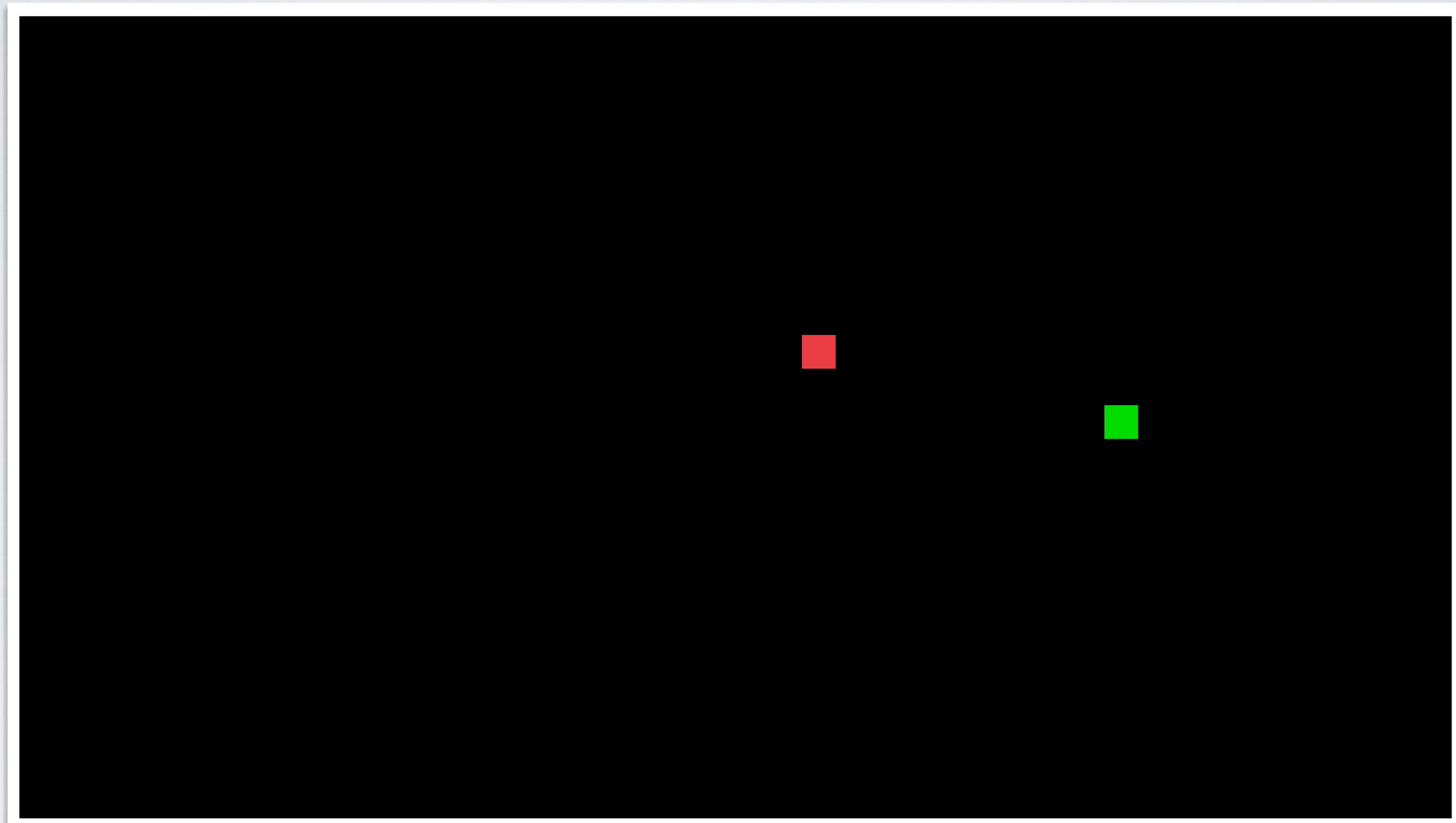
801x801

20,000 "thoughts"

100,004 layers

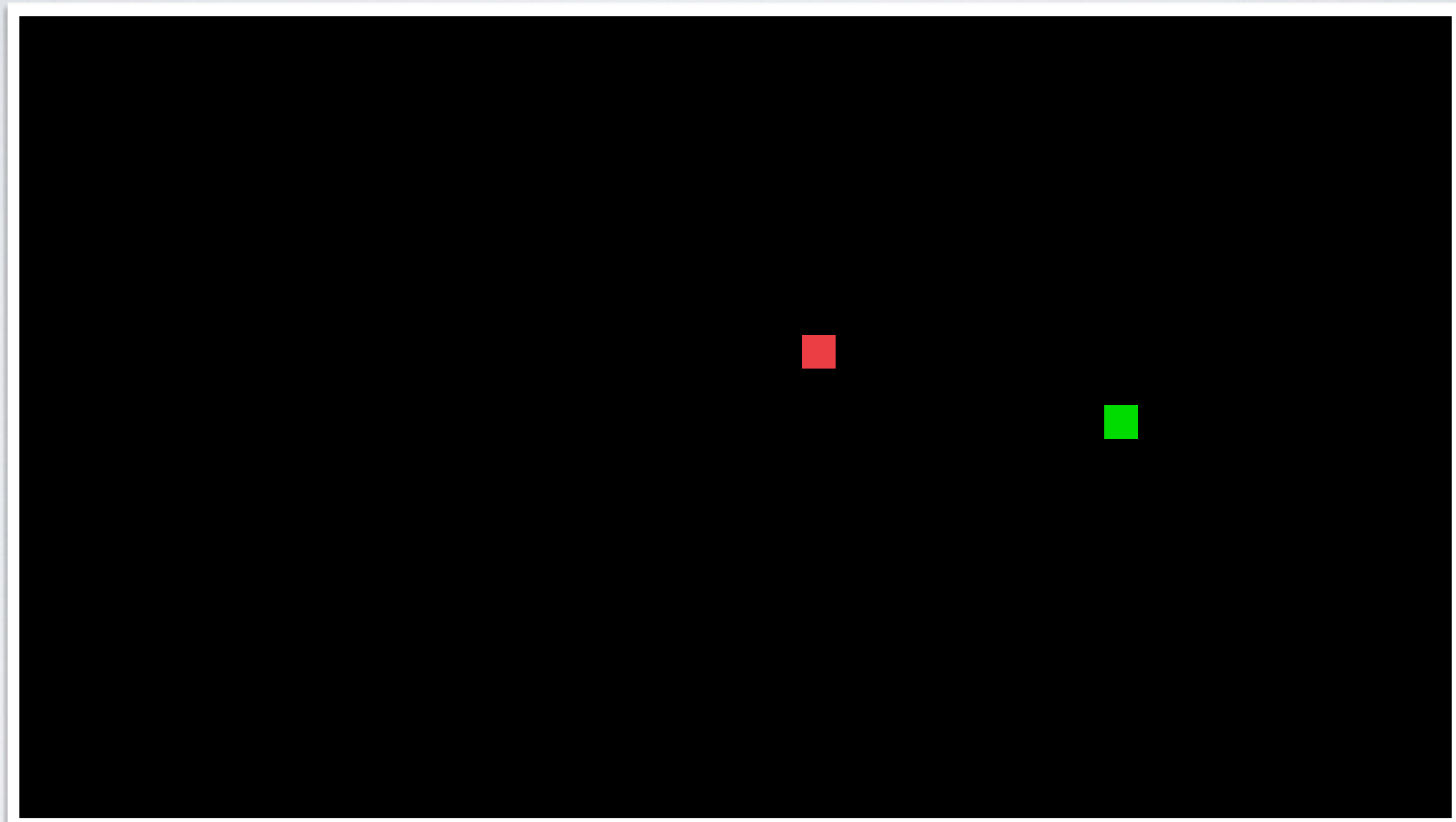
End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22

Solving a maze: start to finish



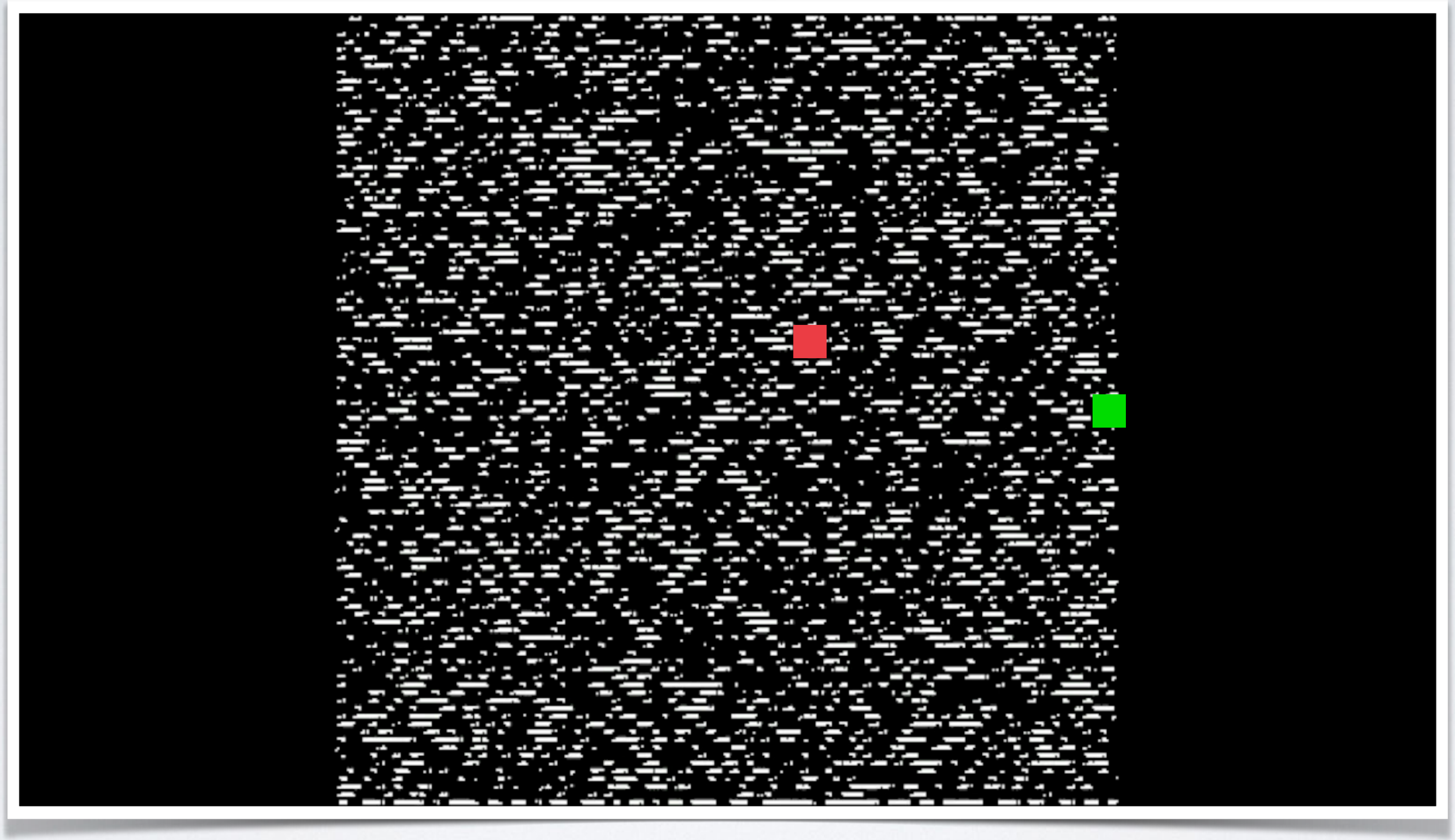
End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22

Solving a maze: start to finish



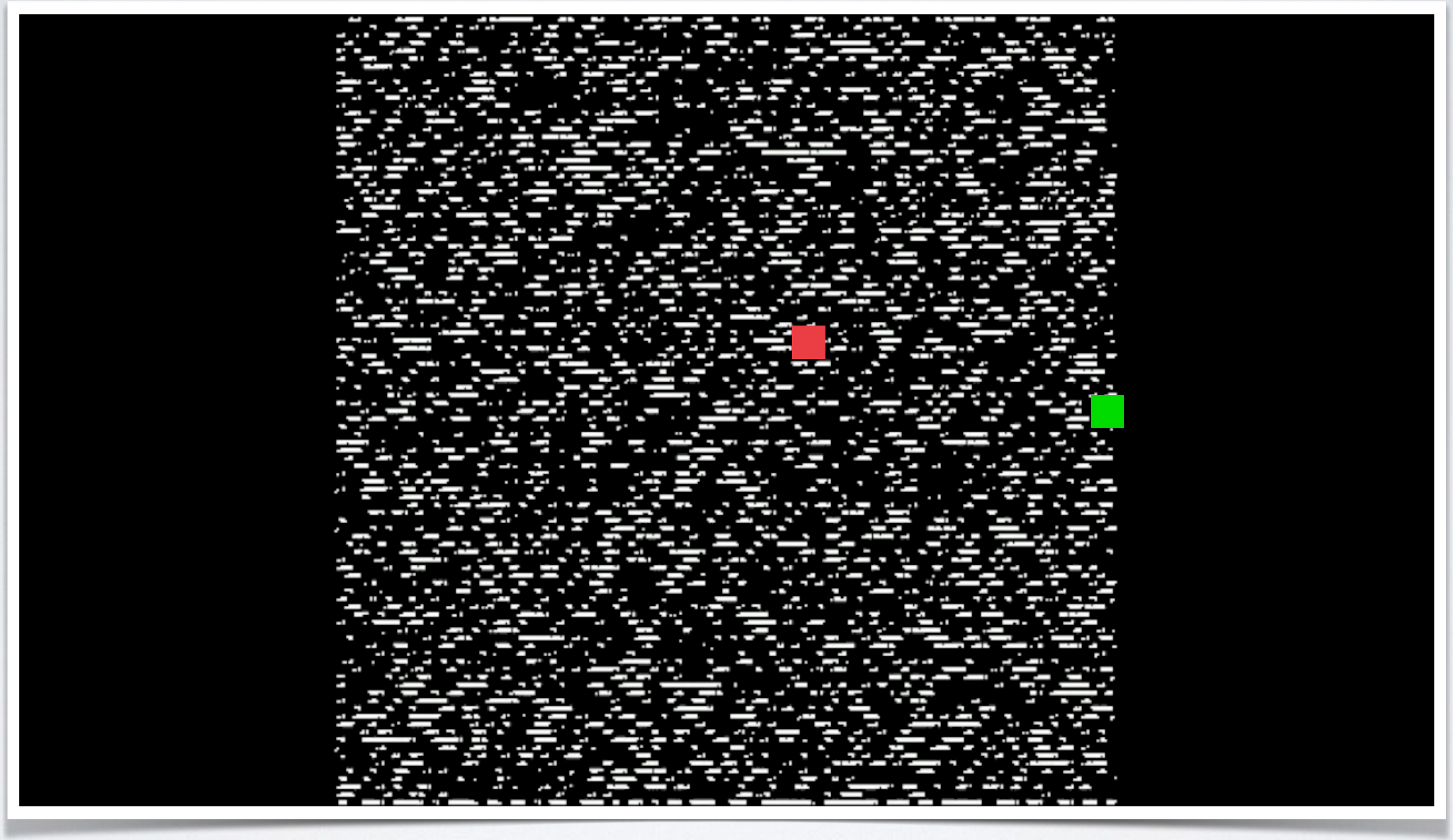
End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22

Corrupt memory with Gaussian noise



End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22

Corrupt memory with Gaussian noise



End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22

CHALLENGE PROBLEM

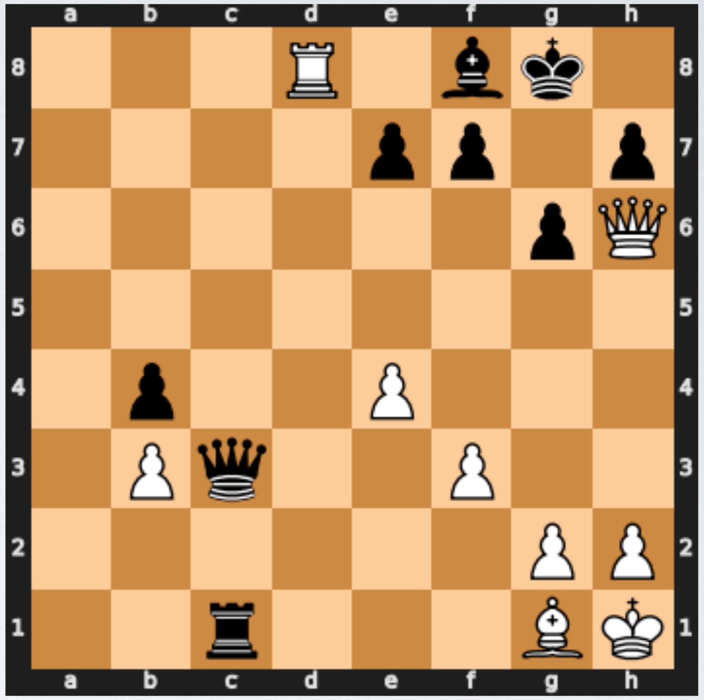
Chess

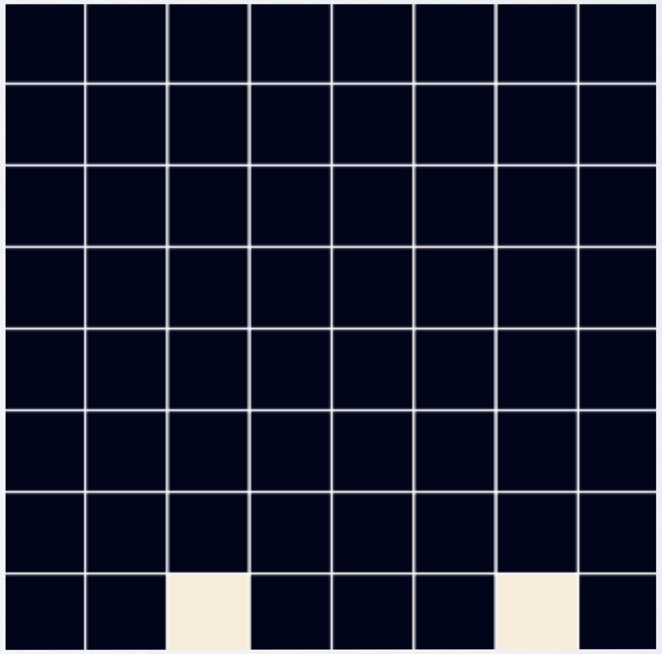
End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS ‘22 lichess.org

"Chess puzzles"

Game scenarios that have clear "best move"

Each puzzle has an Elo rating from human play.





Chess Data

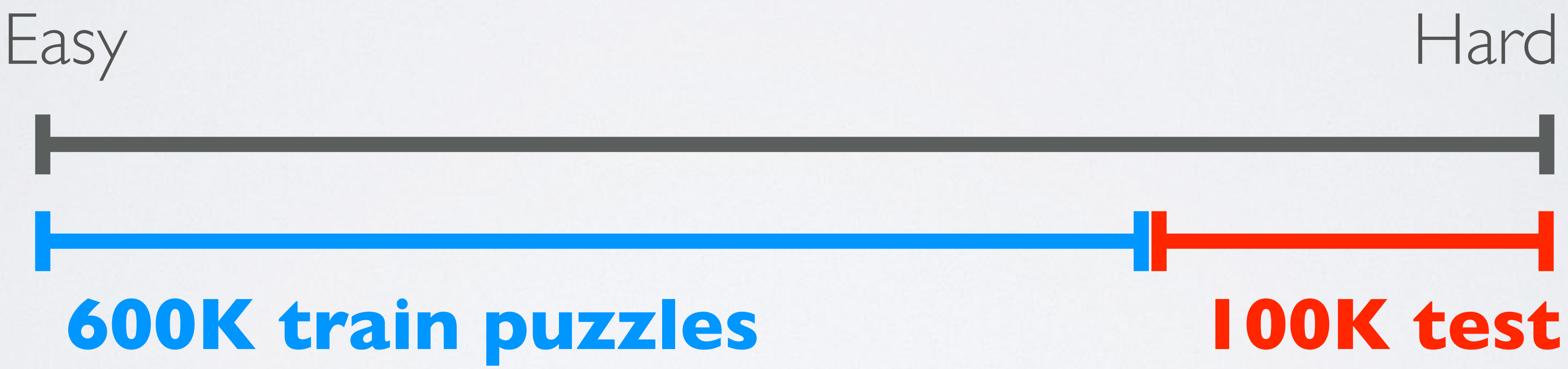
700K puzzles

Easy Hard

End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22

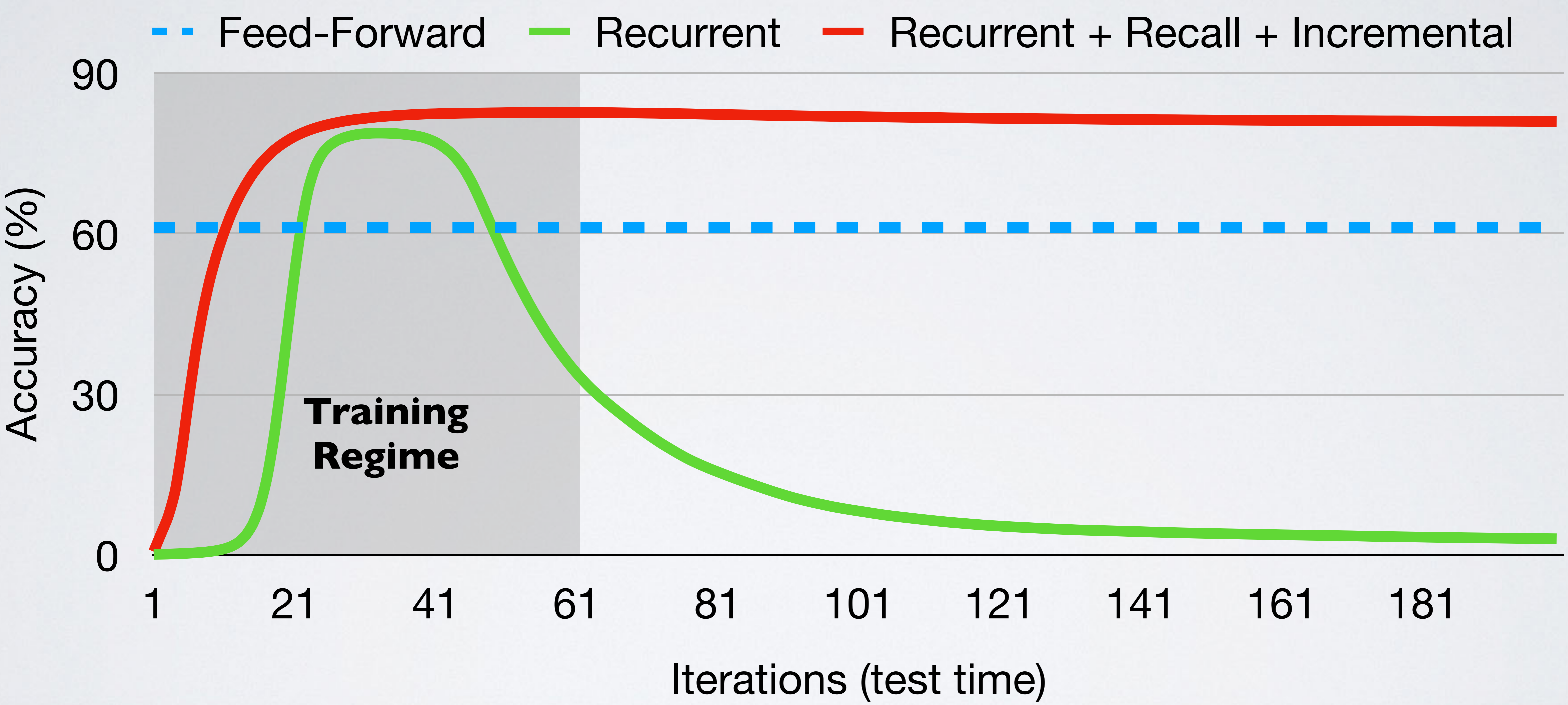
Chess Data

700K puzzles



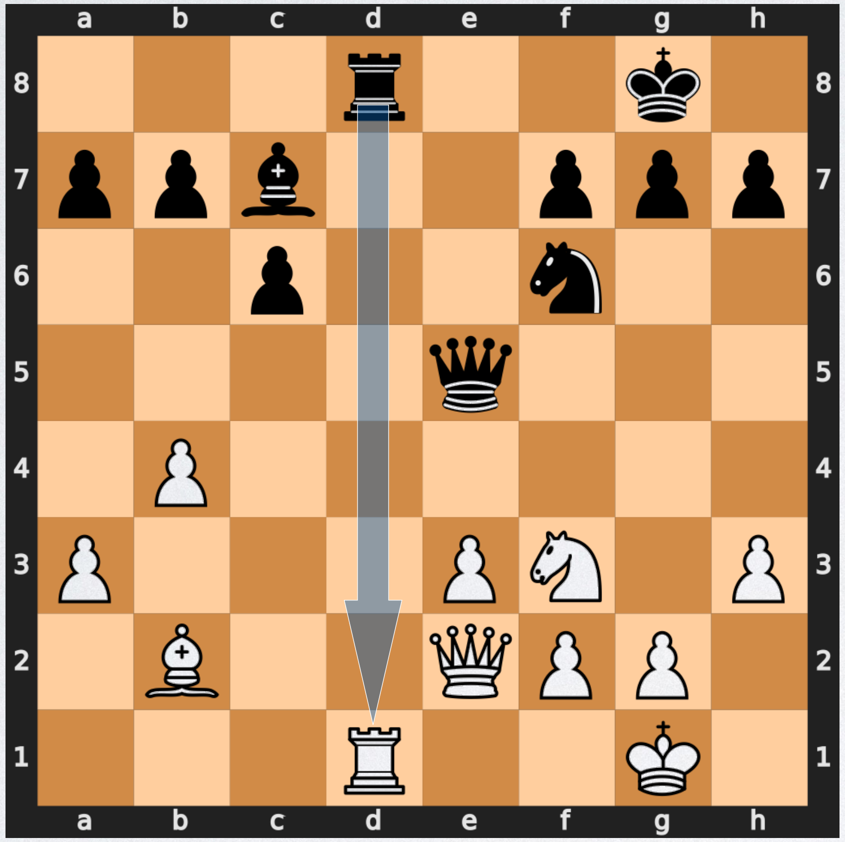
End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22

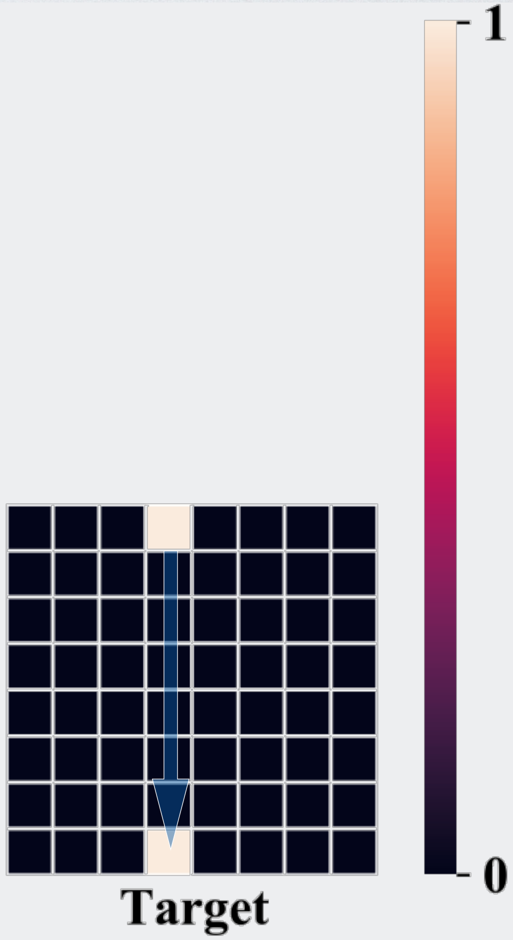
Chess



End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22

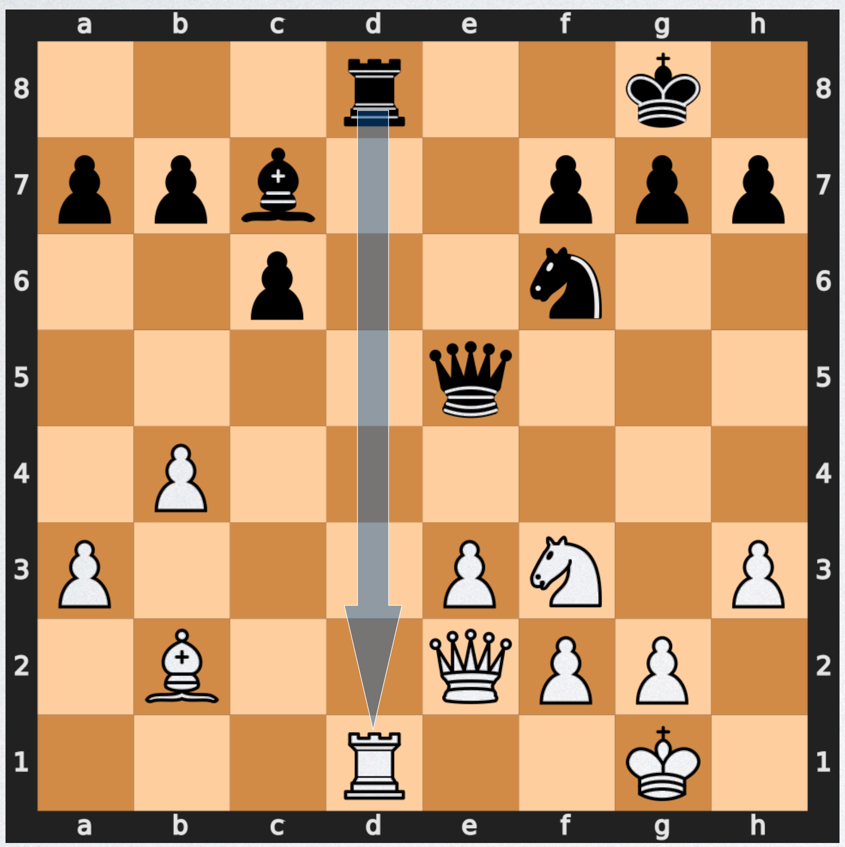
Chess Puzzles

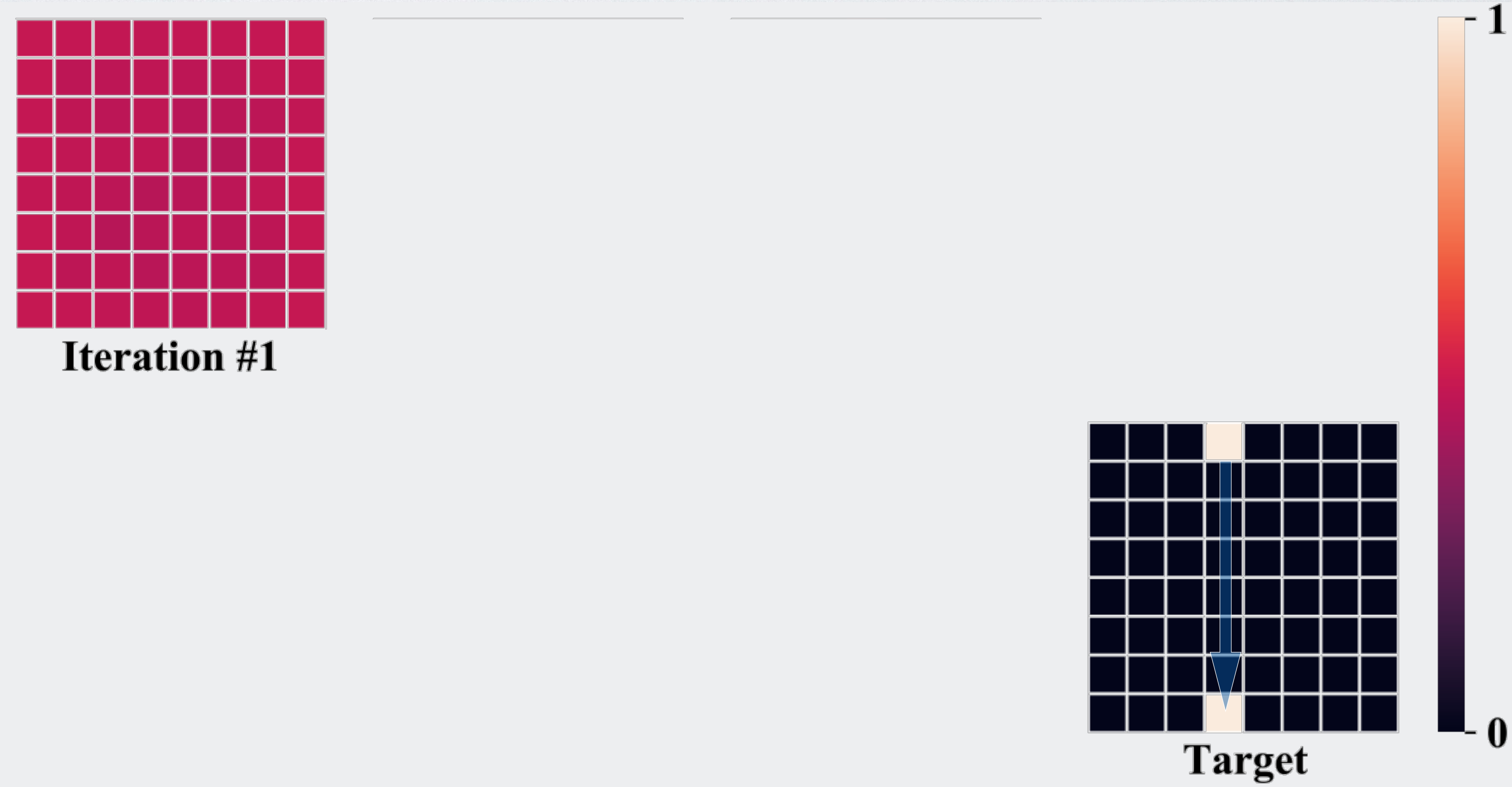




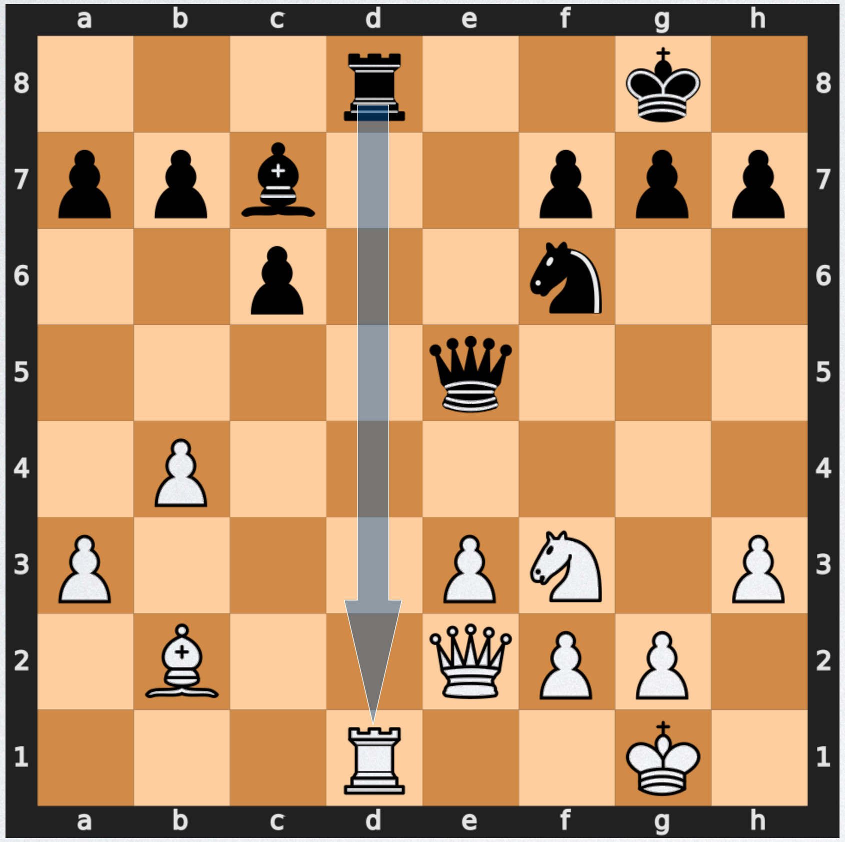
End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22

Chess Puzzles



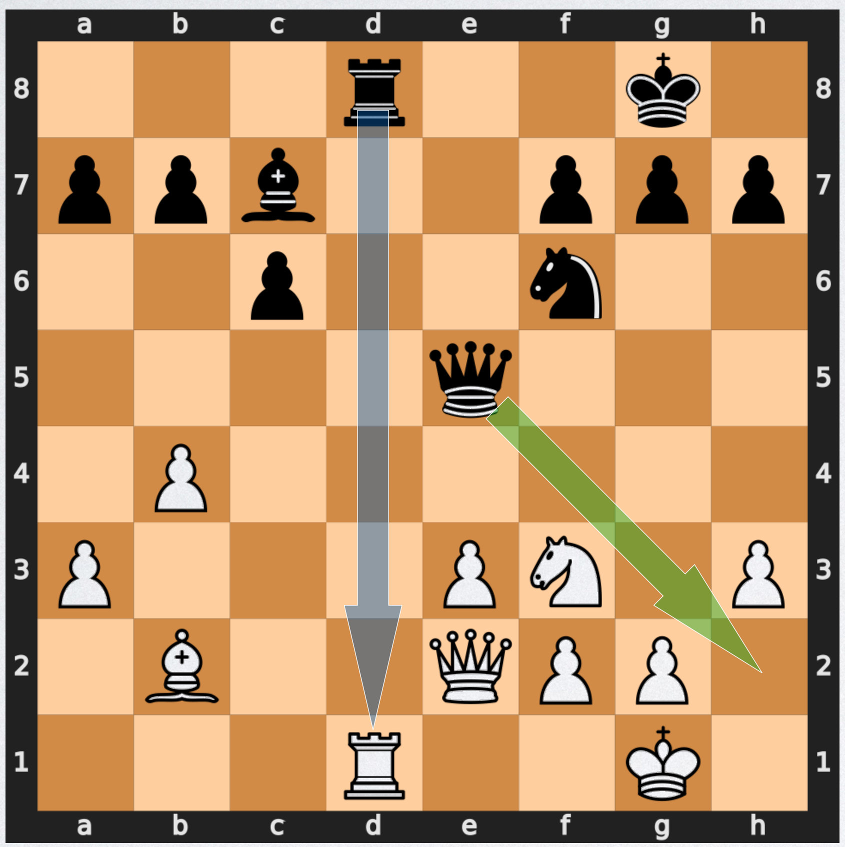


Chess Puzzles



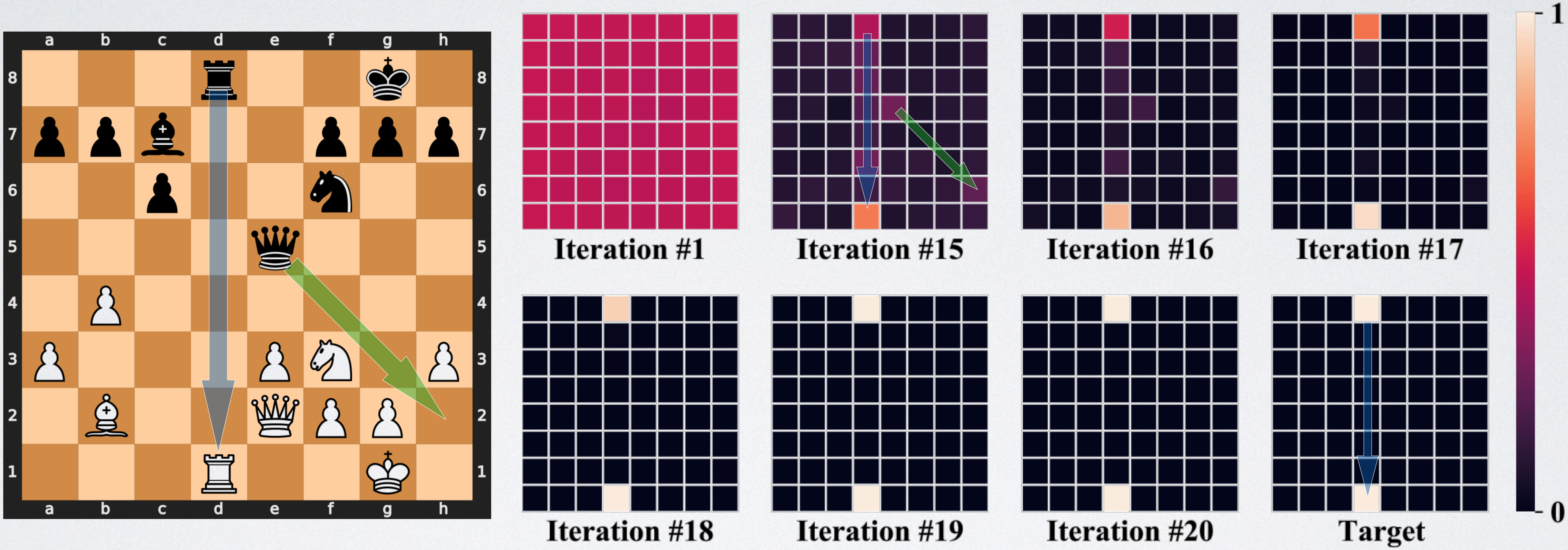
End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22

Chess Puzzles



End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22

Chess Puzzles



End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22

Some thoughts about thinking…

End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS ‘22

Some thoughts about thinking…

Generalize to "hard" problems that lie outside the training distribution.

End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22

Some thoughts about thinking…

Generalize to "hard" problems that lie outside the training distribution.

See only the *problem* and *solution*, and organically learn algorithms end-to-end.

Some thoughts about thinking…

Generalize to "hard" problems that lie outside the training distribution.

See only the *problem* and *solution*, and organically learn algorithms end-to-end.

Can we replace hand-crafted algorithms?

End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22

Some thoughts about thinking…

Generalize to "hard" problems that lie outside the training distribution.

See only the *problem* and *solution*, and organically learn algorithms end-to-end.

Can we replace hand-crafted algorithms?

What can humans do that neural networks can't?

End-to-end Algorithm Synthesis with Recurrent Networks, NeurIPS '22

Thanks!

Andrew Wilson

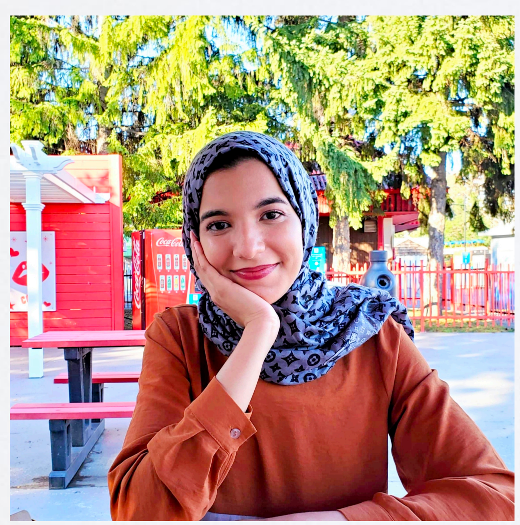
Tom Goldstein

Avi Schwarzschild

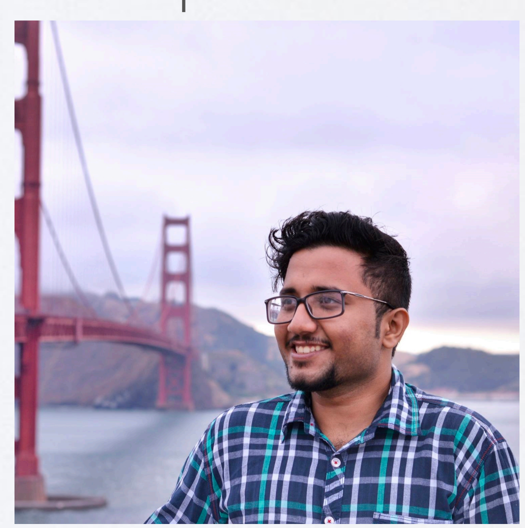
Ping Chiang

Jonas Geiping

Sanae Lotfi



Arpit Bansal



Ronny Huang