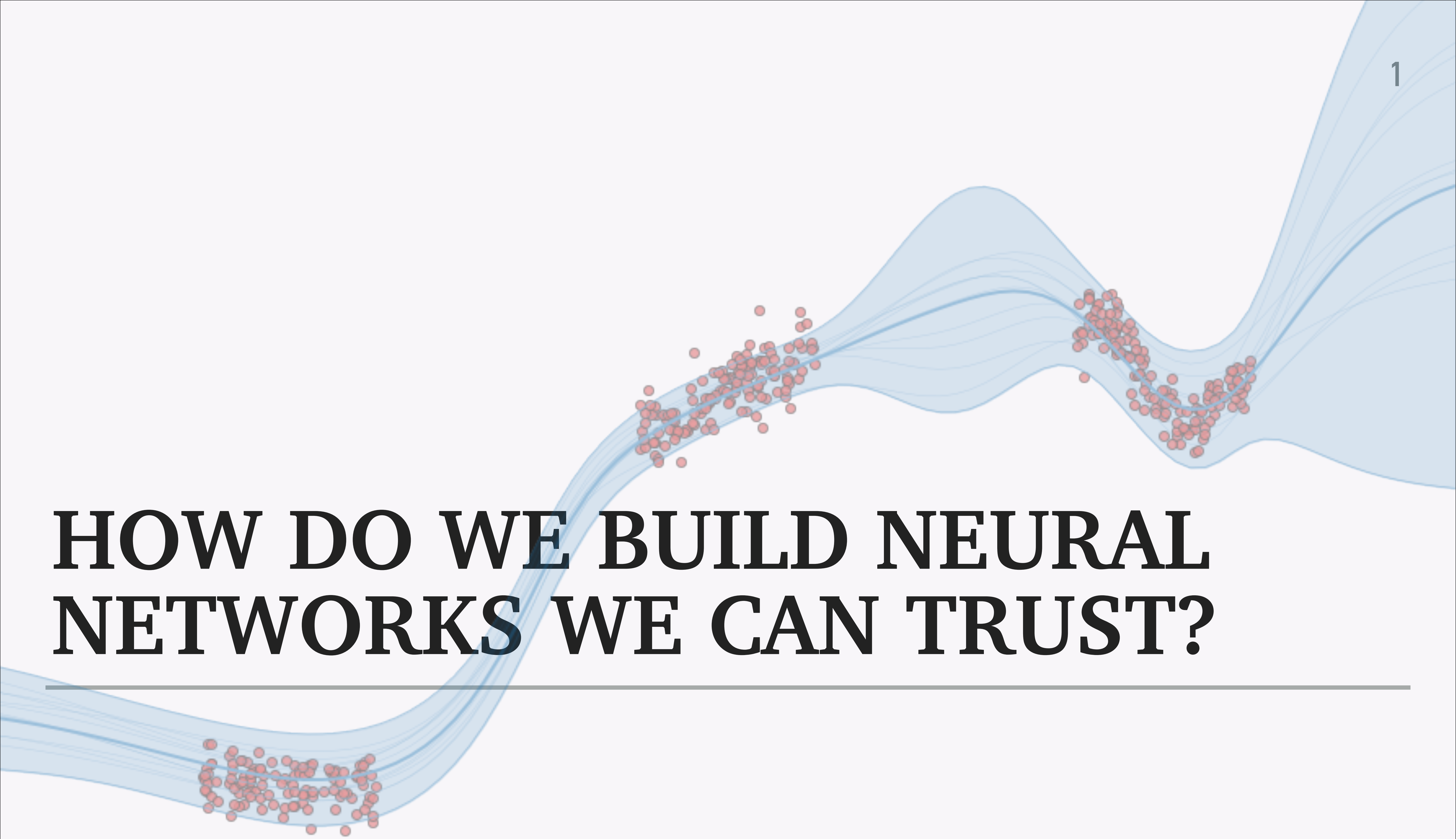
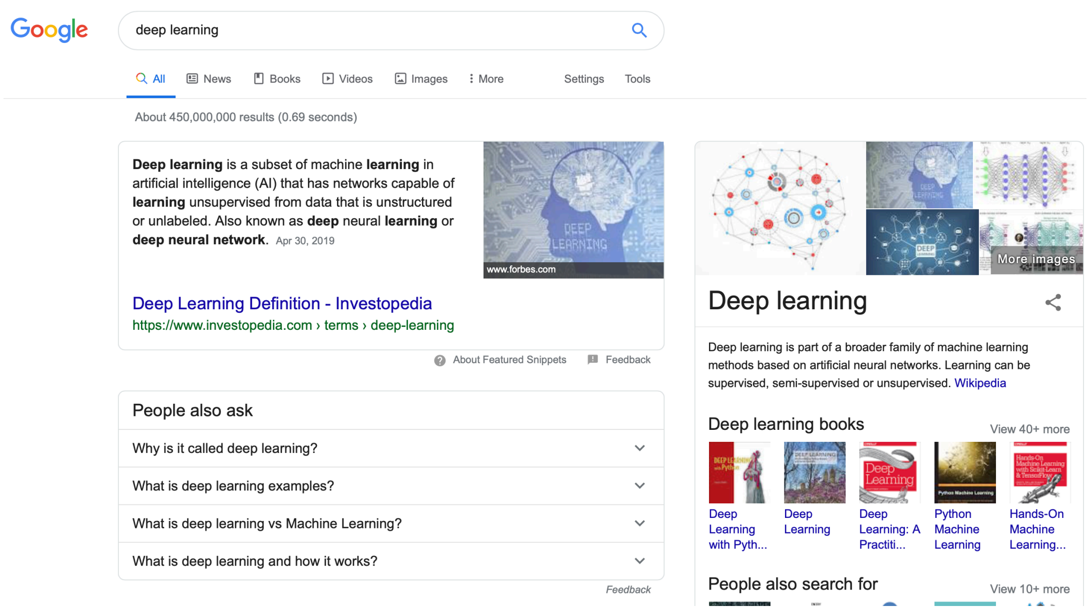


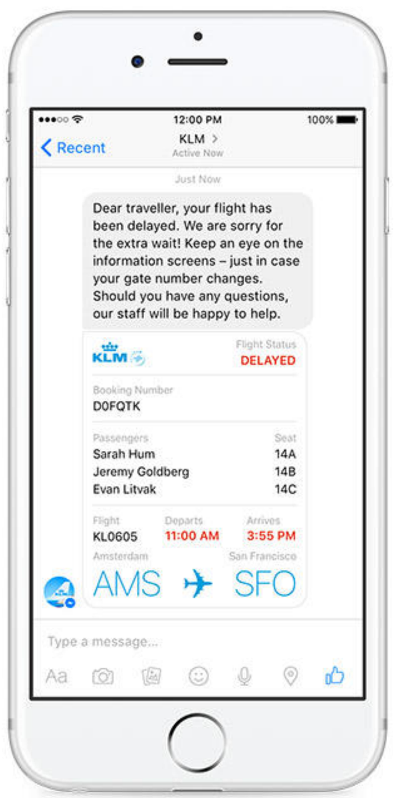
HOW DO WE BUILD NEURAL NETWORKS WE CAN TRUST?

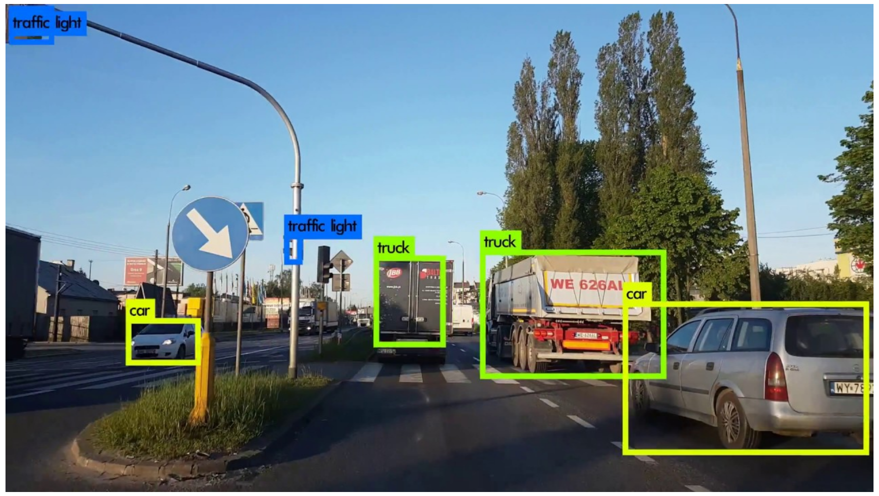
POLINA KIRICHENKO
PAVEL IZMAILOV

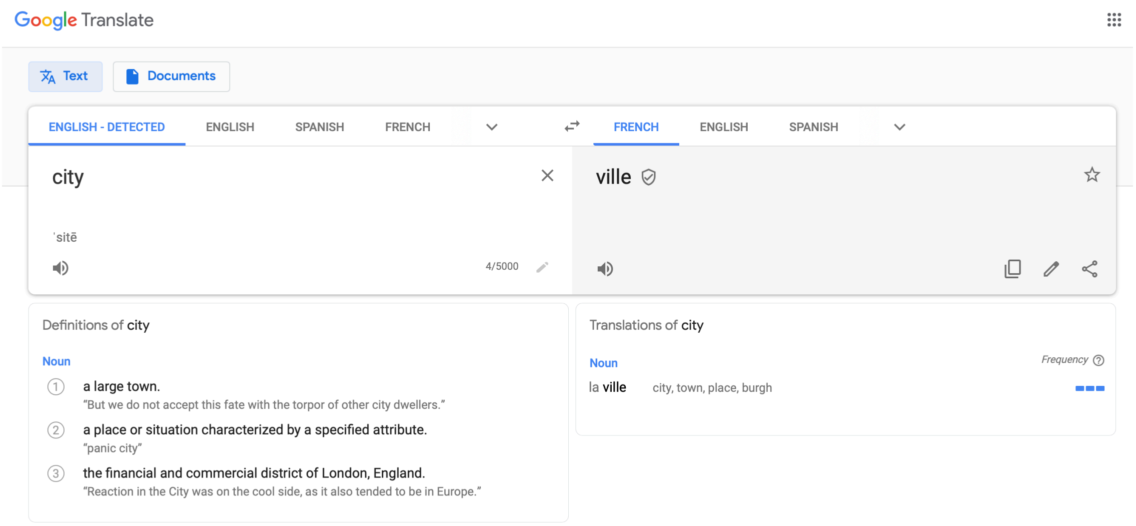
NYU

DEEP LEARNING SUCCESS



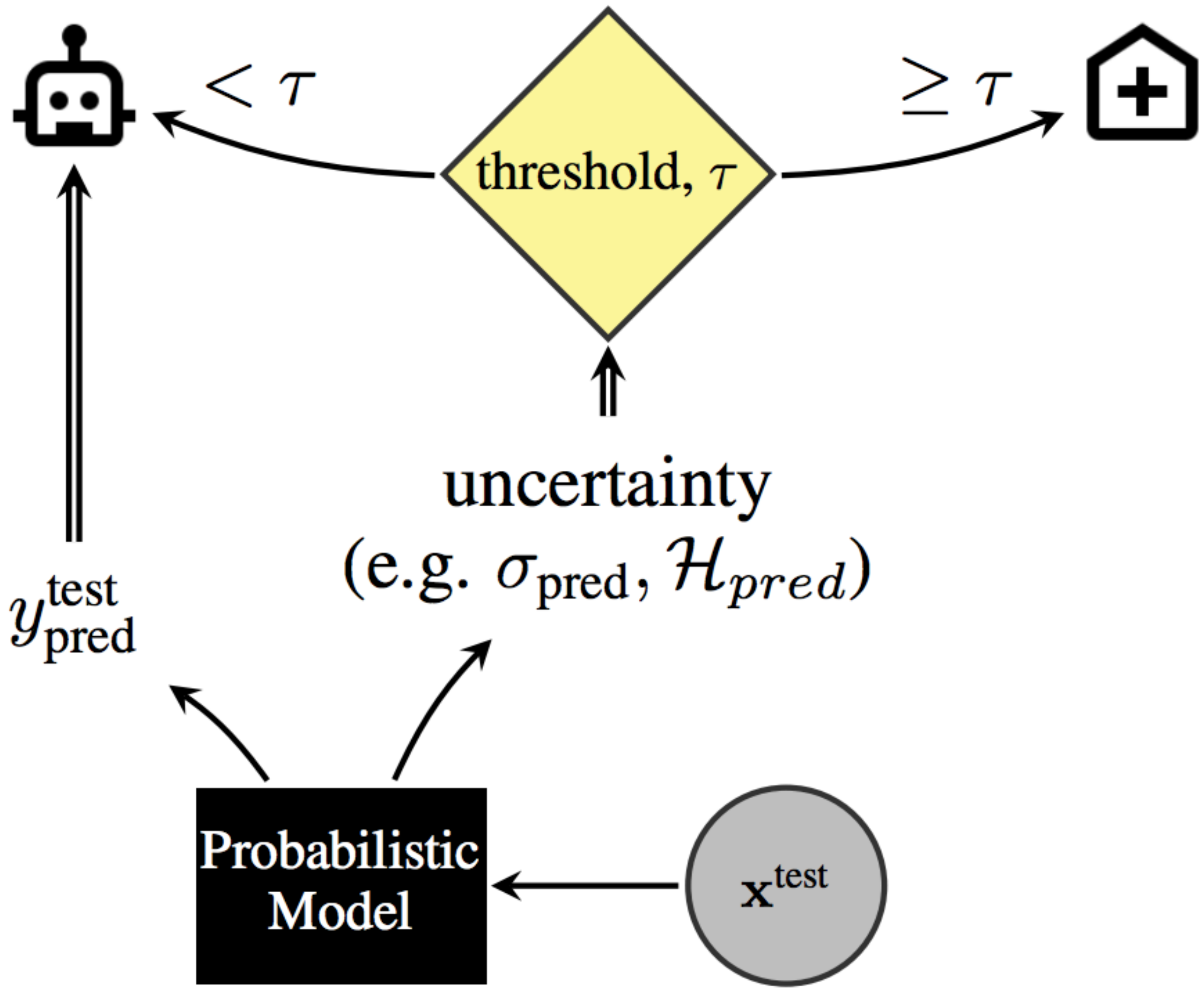






UNCERTAINTY IN DEEP LEARNING

Automated diagnosis: human-in-the-loop

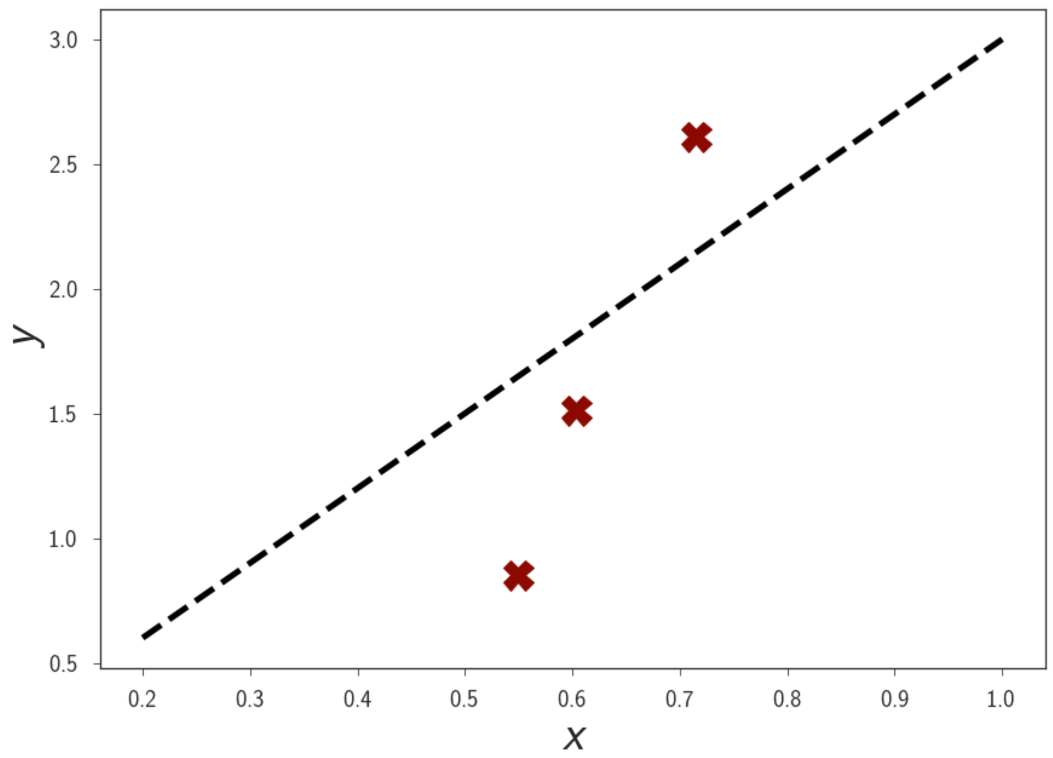


"Benchmarking Bayesian Deep Learning with Diabetic Retinopathy Diagnosis" by Angelos Filos et al.

BAYESIAN MACHINE LEARNING

Consider a simple linear regression problem:

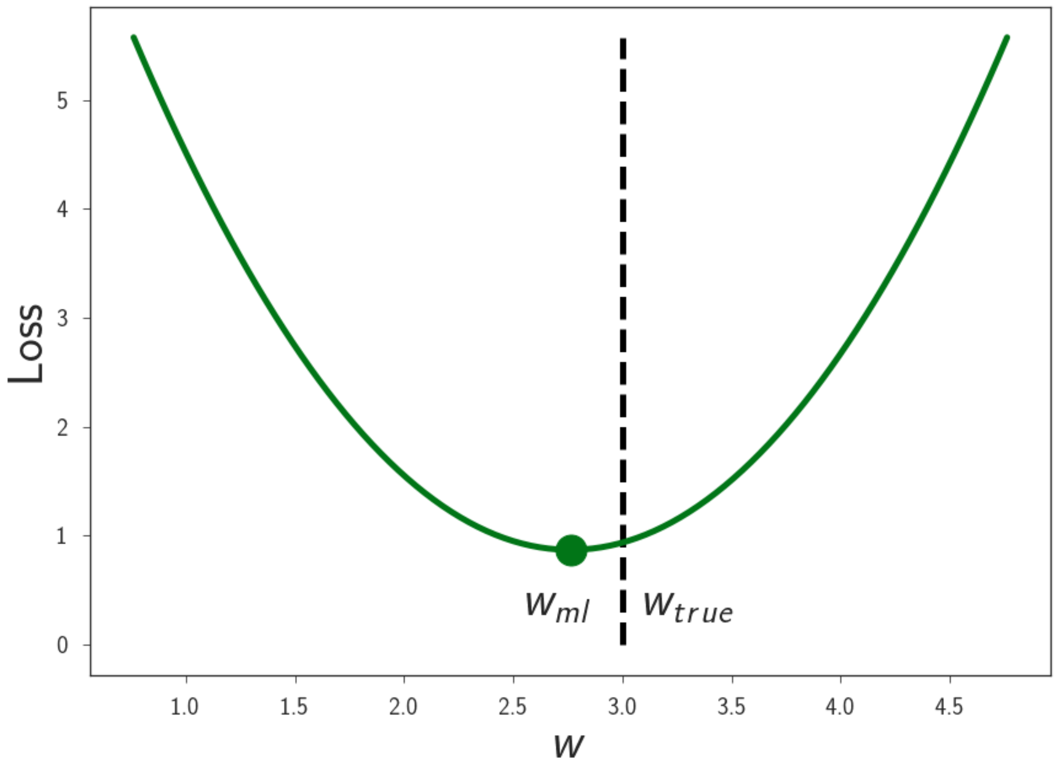
$$y = wx + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

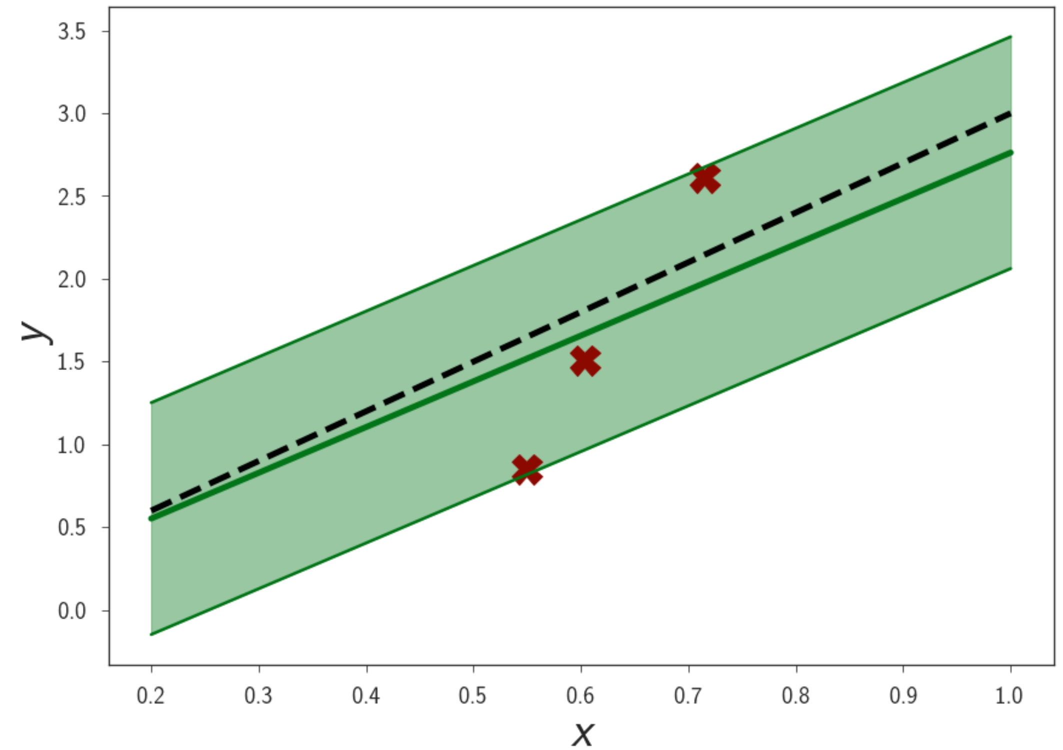


BAYESIAN MACHINE LEARNING

Standard linear regression:

$$\max_w \sum_{i=1}^{N} \log \mathcal{N}(y_i | w x_i, \sigma^2) \iff \min_w \frac{1}{N} \sum_{i=1}^{N} (y_i - w x_i)^2$$



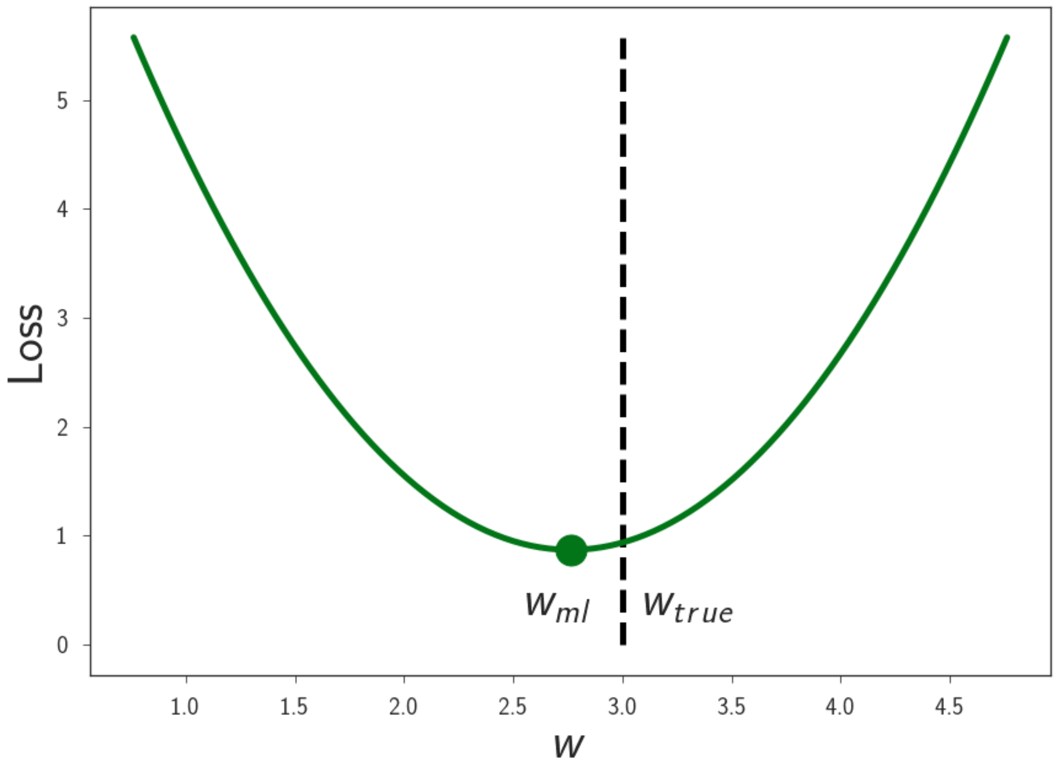


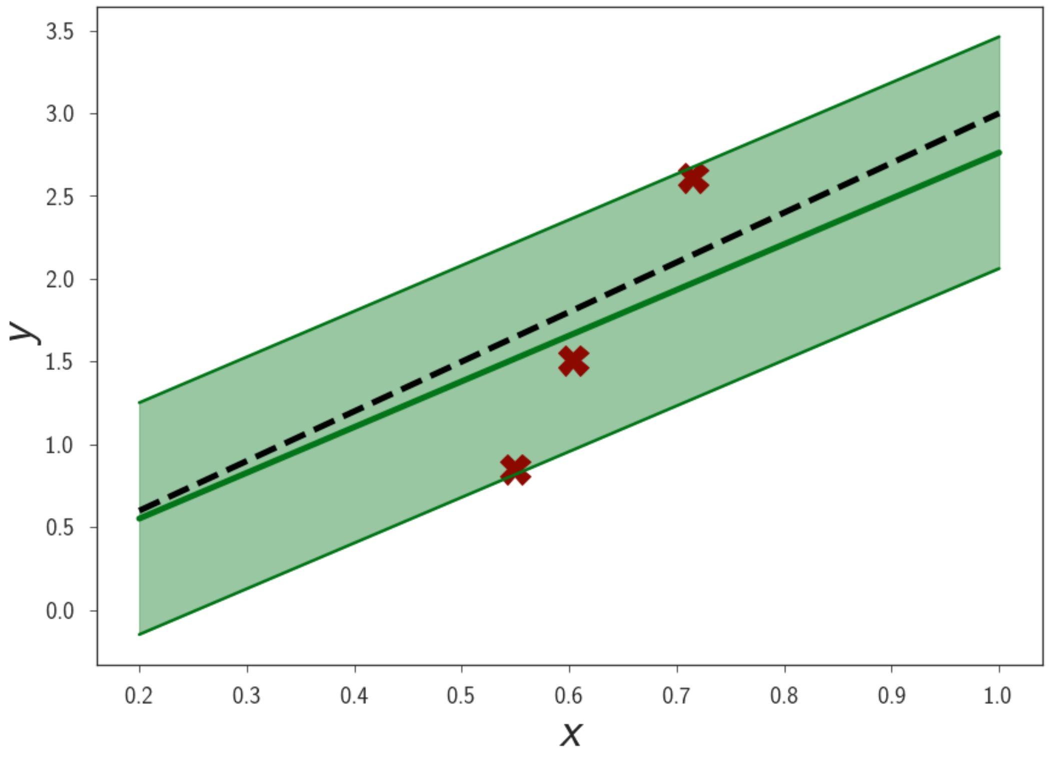
BAYESIAN MACHINE LEARNING

Standard linear regression:

$$\max_w \sum_{i=1}^{N} \log \mathcal{N}(y_i | w x_i, \sigma^2) \quad \Longleftrightarrow \quad \min_w \frac{1}{N} \sum_{i=1}^{N} (y_i - w x_i)^2$$

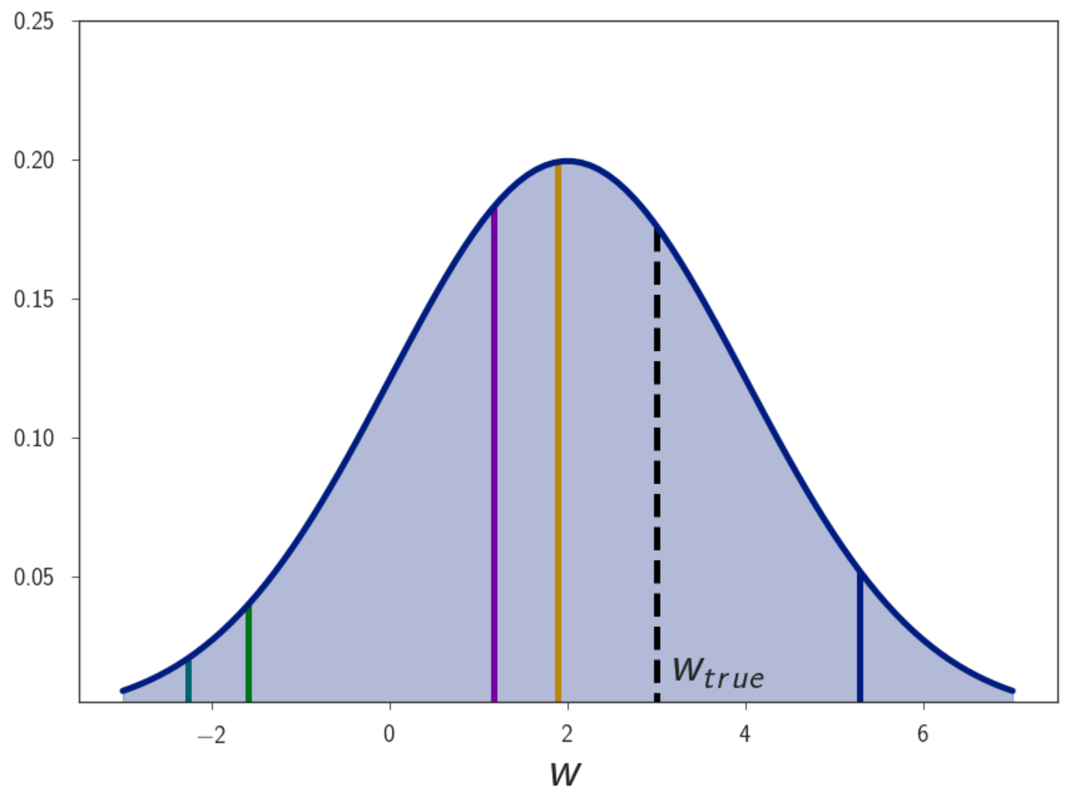
We want to model uncertainty over parameters of the model

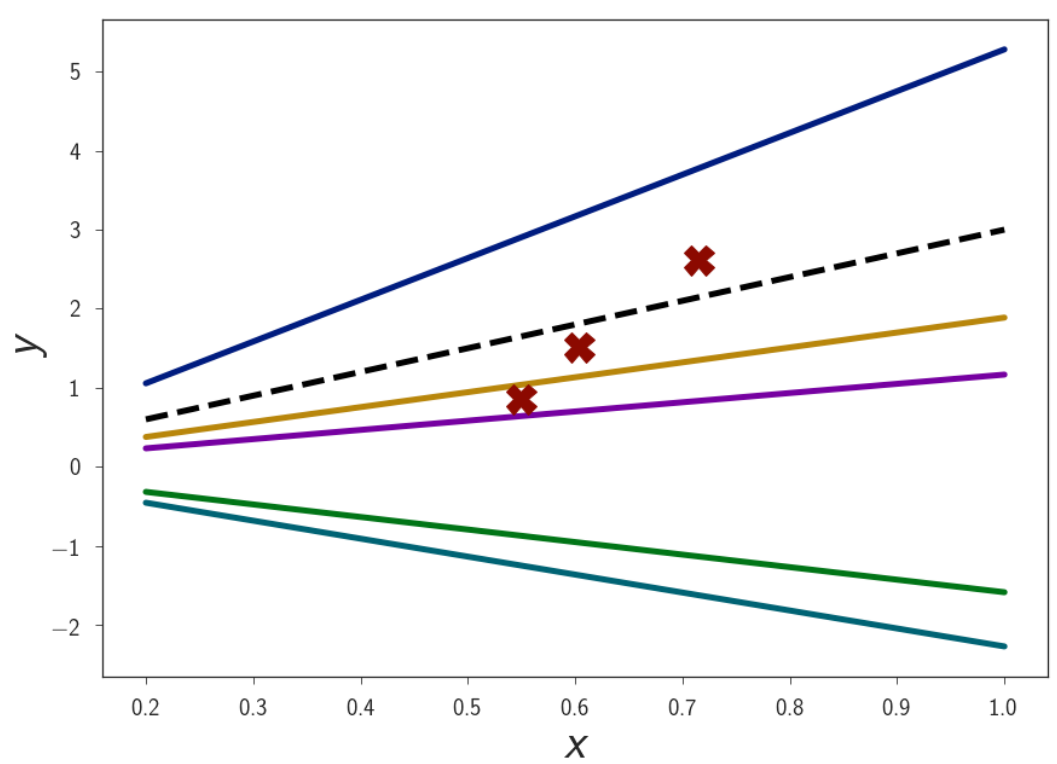




BAYESIAN MACHINE LEARNING

Step 1: introduce a prior distribution $p(w)$ over parameters

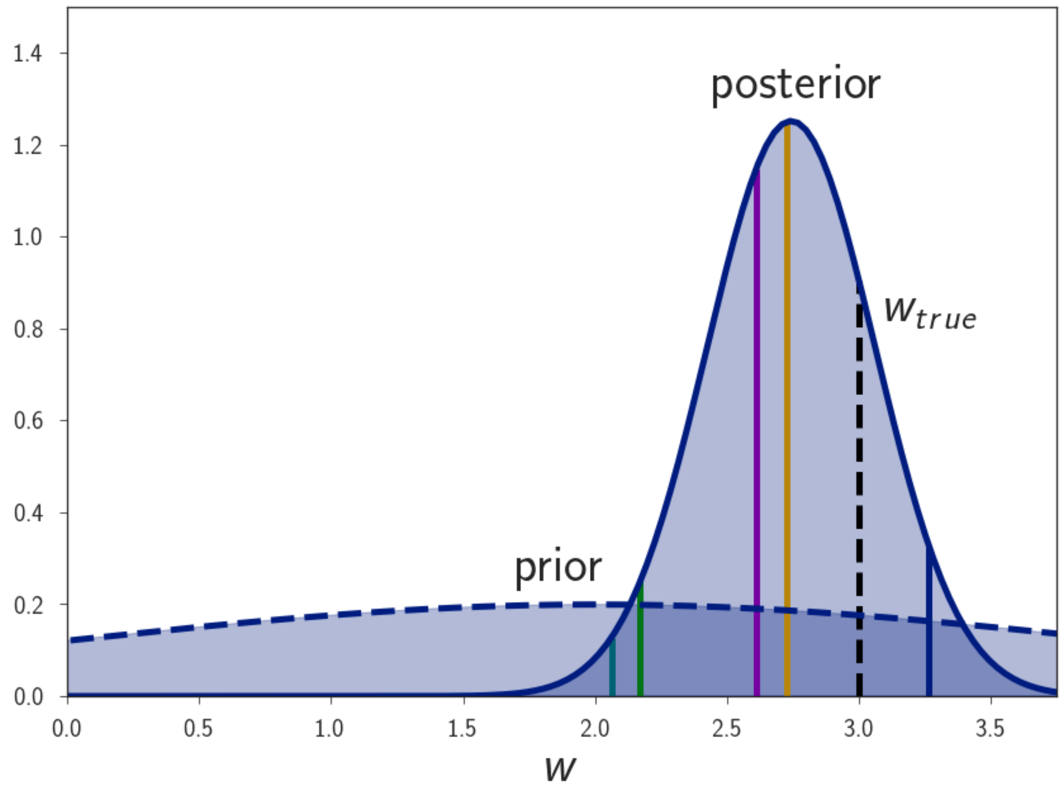


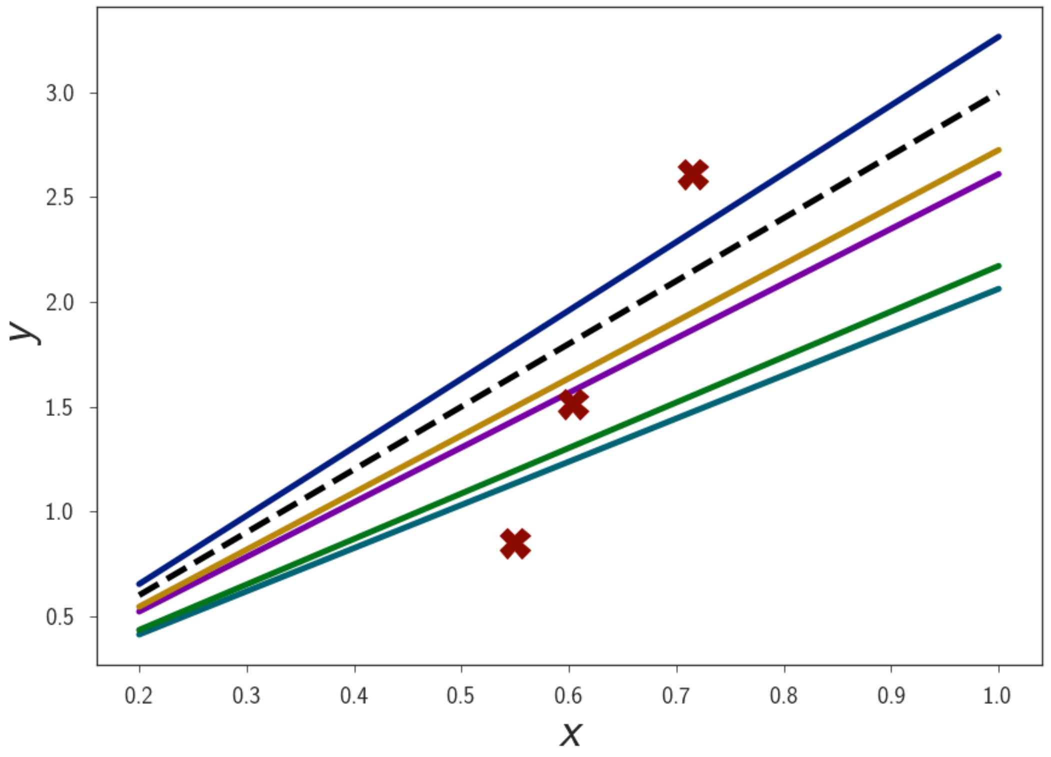


BAYESIAN MACHINE LEARNING

Step 2: Compute posterior $p(w|D)$ using Bayes rule

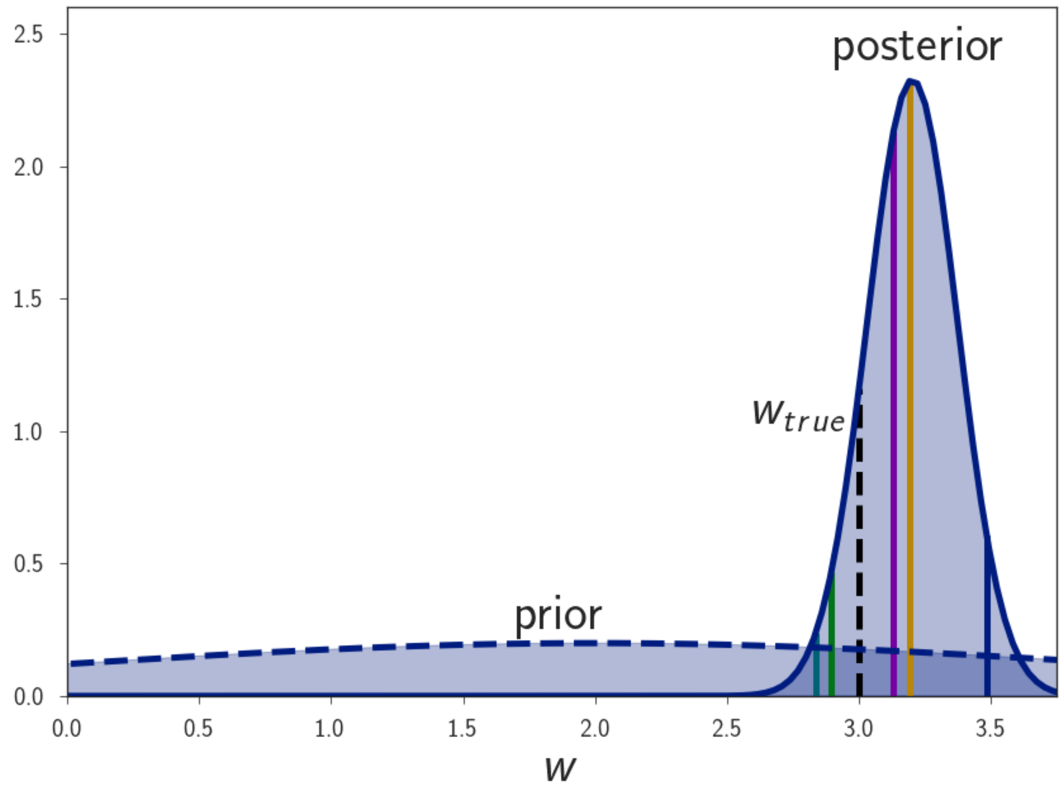
$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

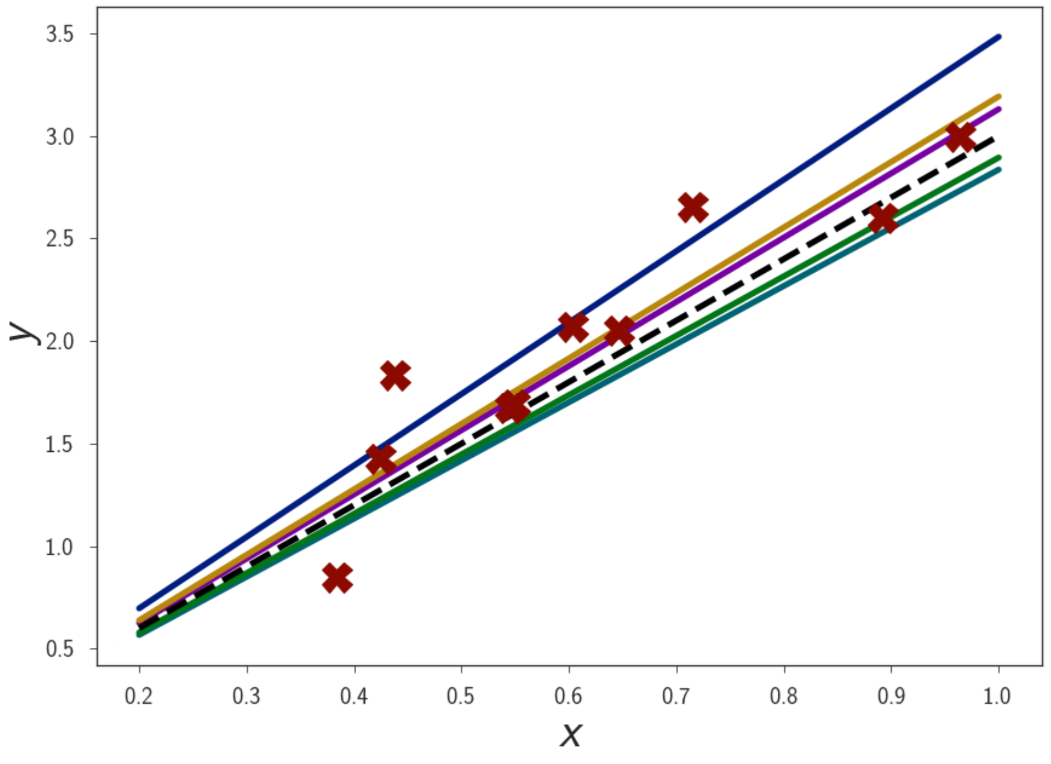




BAYESIAN MACHINE LEARNING: POSTERIOR CONTRACTION

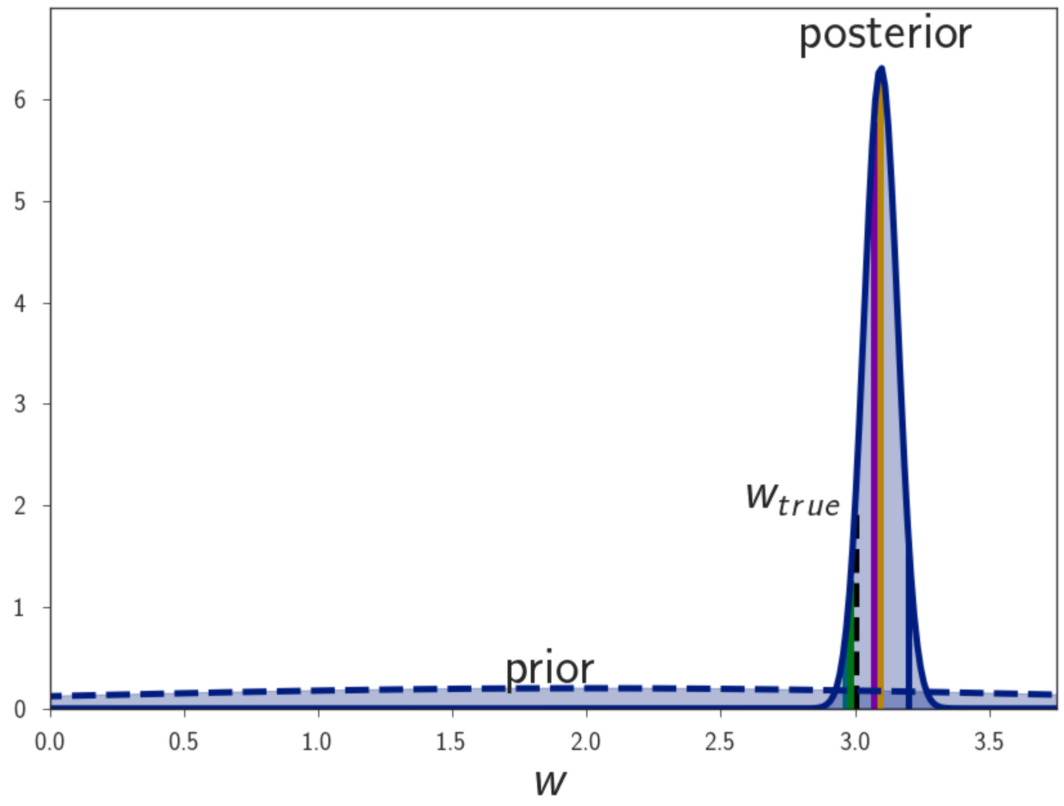
$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

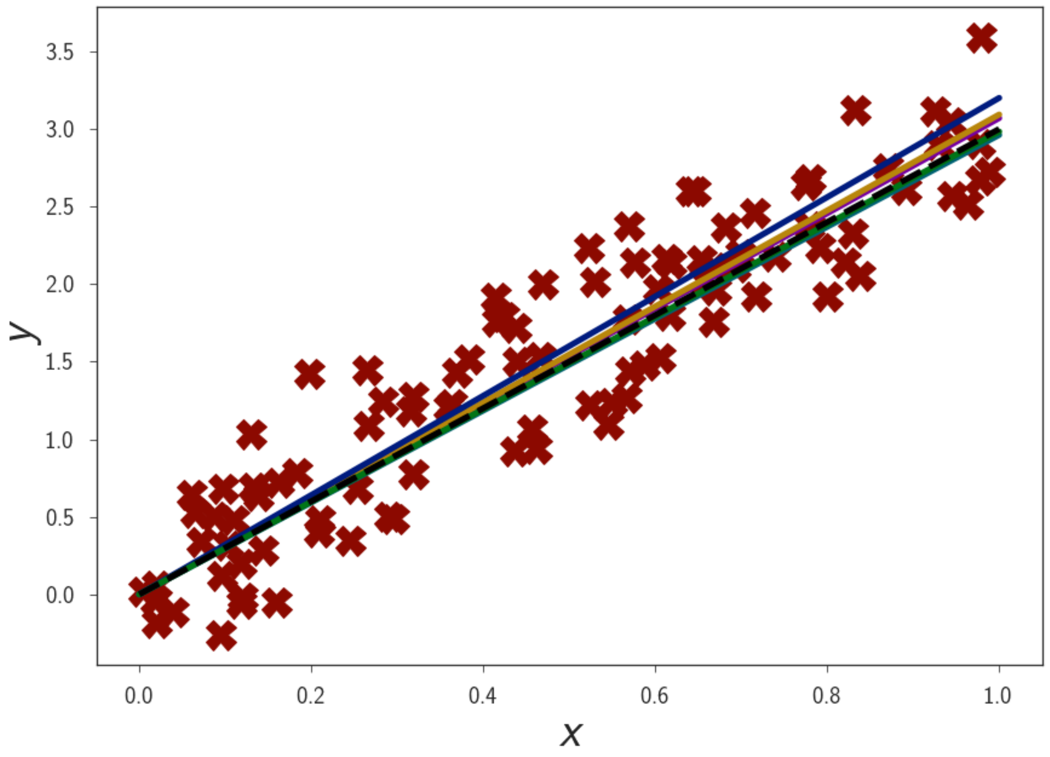




BAYESIAN MACHINE LEARNING: POSTERIOR CONTRACTION

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

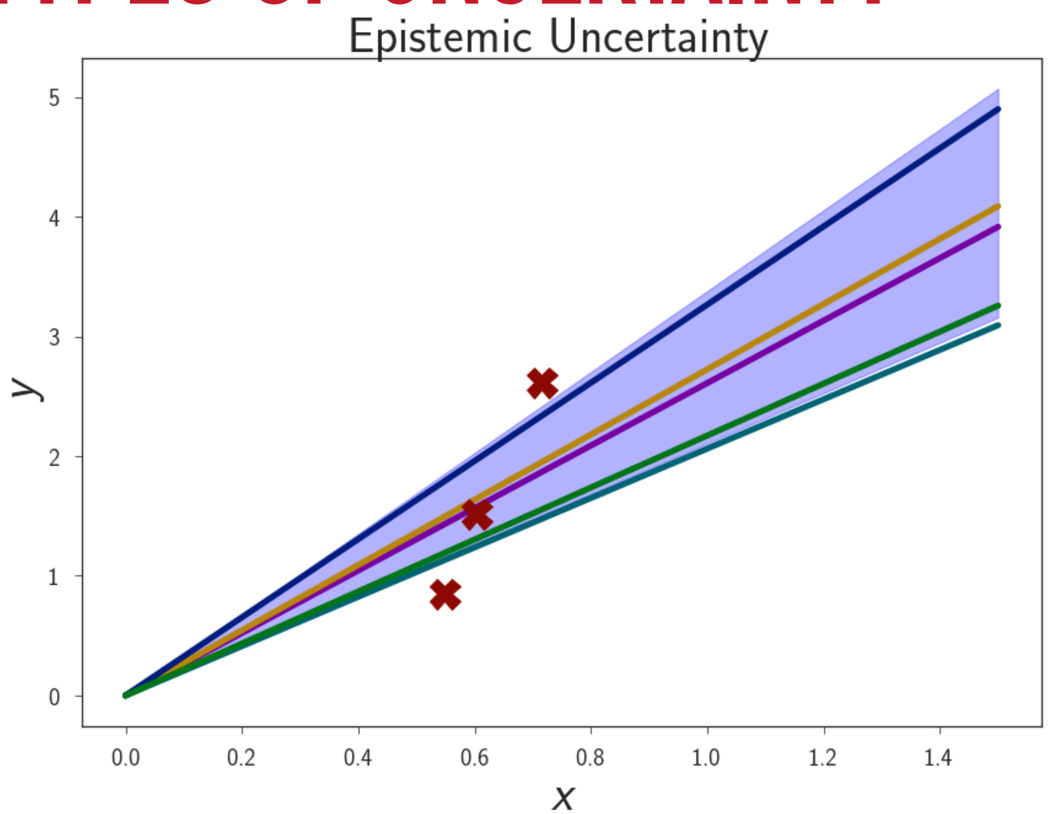




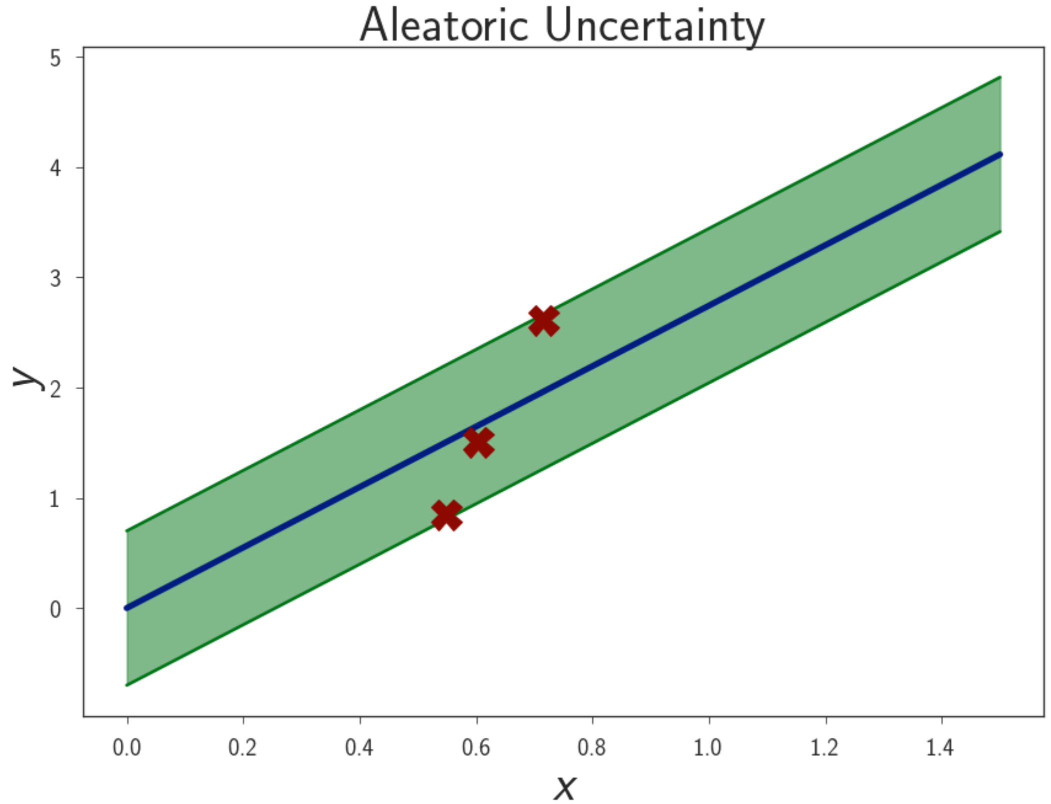
BAYESIAN MACHINE LEARNING: TWO TYPES OF UNCERTAINTY

Epistemic uncertainty is our uncertainty over the model

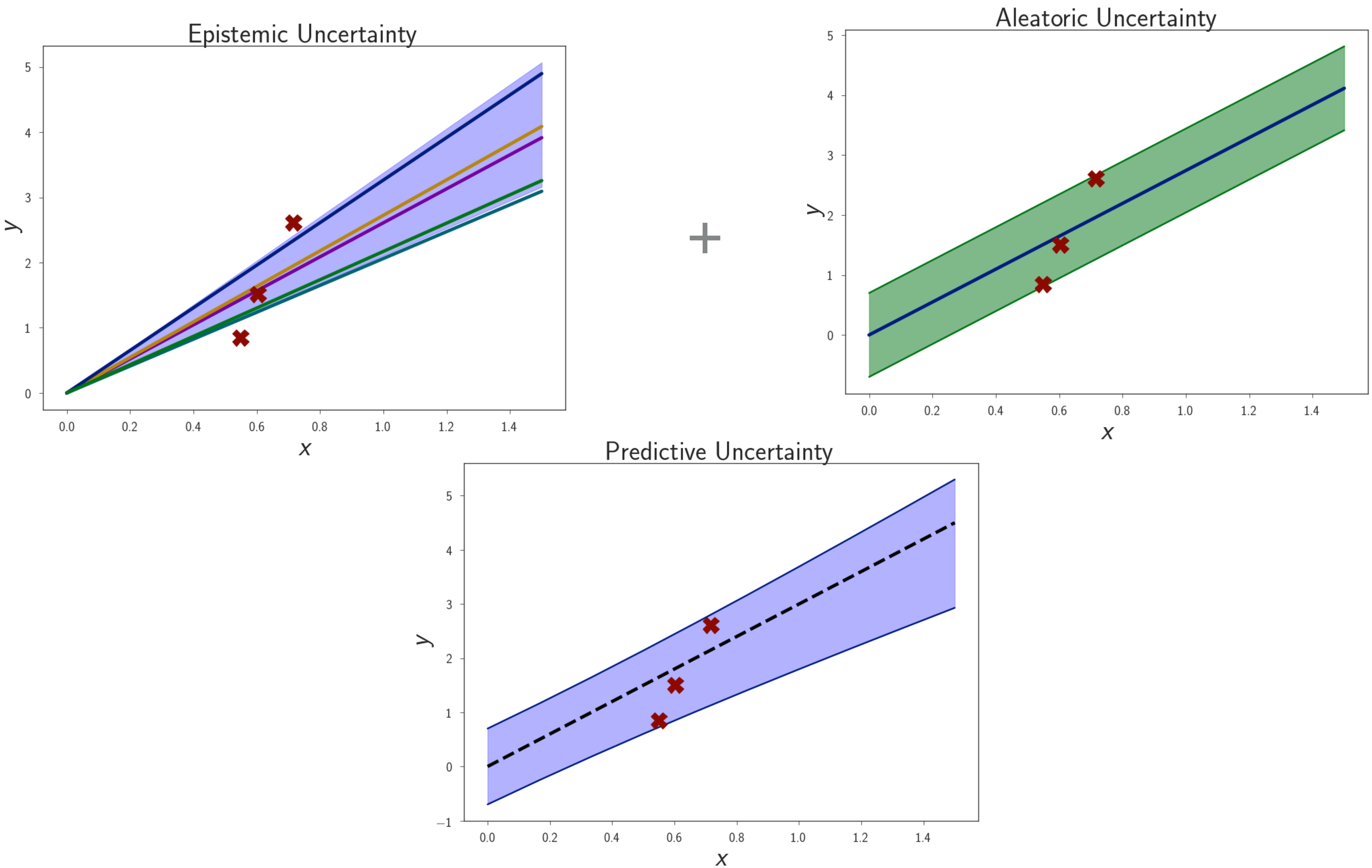
- Grows with x because uncertainty in w is multiplied by x



Aleatoric uncertainty is our uncertainty over the data for a fixed model, e.g. noise.

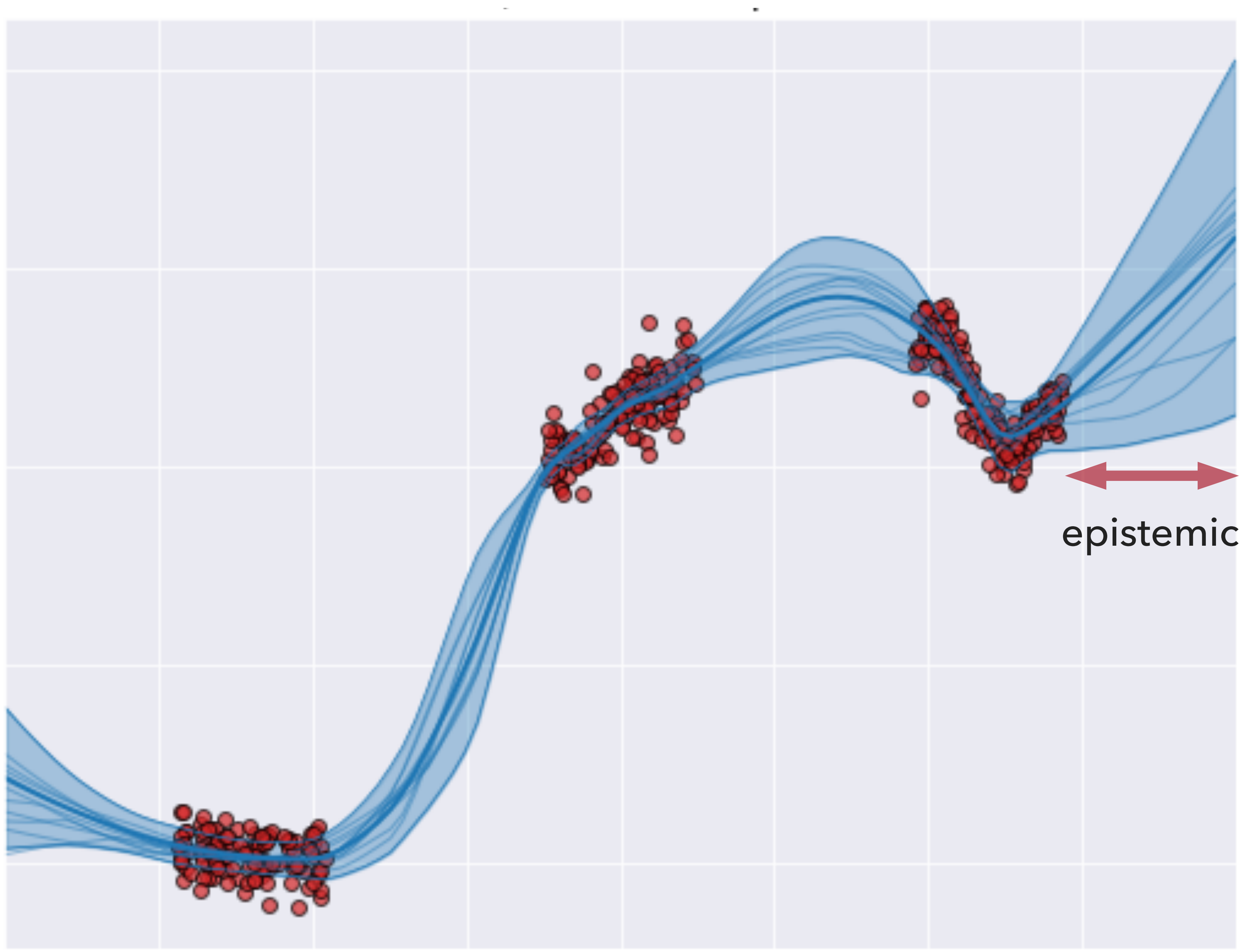


BAYESIAN MACHINE LEARNING: BAYESIAN MODEL AVERAGING



BAYESIAN MACHINE LEARNING: TWO TYPES OF UNCERTAINTY

Epistemic uncertainty: non-linear model



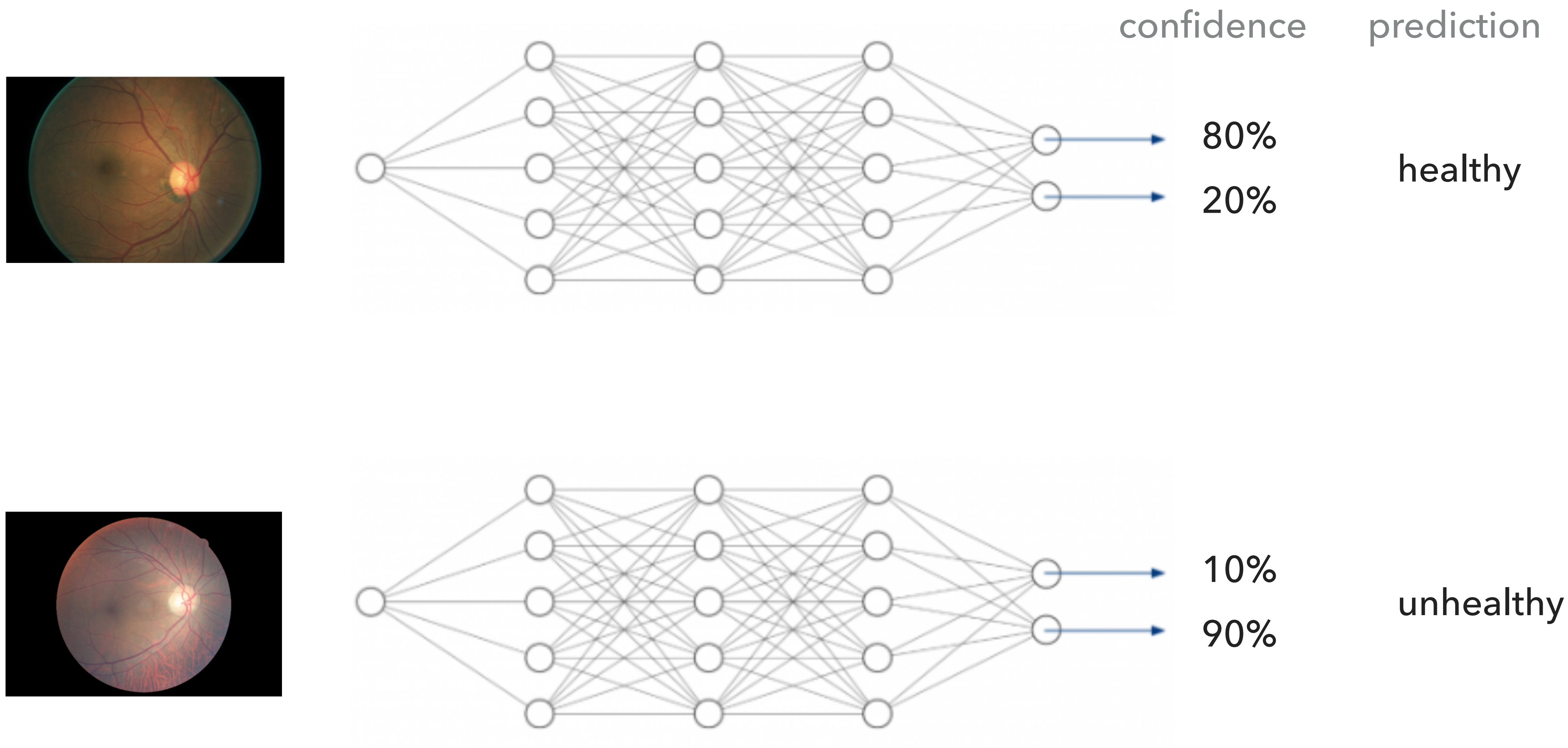
BAYESIAN MACHINE LEARNING: BAYESIAN MODEL AVERAGING

- We combine aleatoric and epistemic uncertainties via BMA:

$$p(y^*|x^*, D) = \int_w p(y^*|x^*, w)p(w|D)dw$$

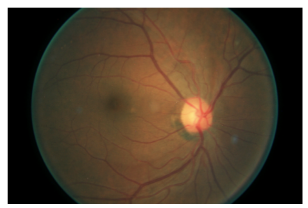
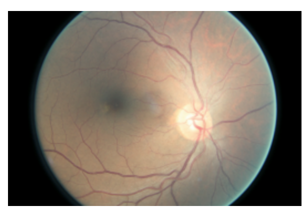
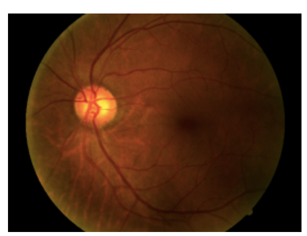
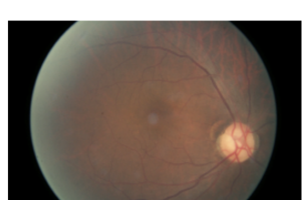
- Ignoring the uncertainty in the posterior over w leads to overconfident predictions

CALIBRATION



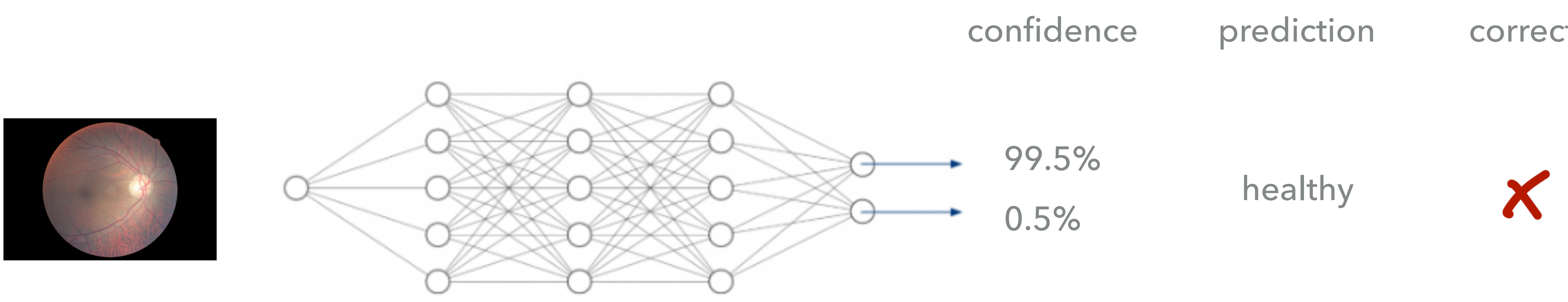
"Benchmarking Bayesian Deep Learning with Diabetic Retinopathy Diagnosis" by Angelos Filos et al.

CALIBRATION

	confidence	prediction	correct
	80% 20%	healthy	✓
	80% 20%	healthy	✓
	80% 20%	healthy	✓
	80% 20%	healthy	✗
	80% 20%	healthy	✓

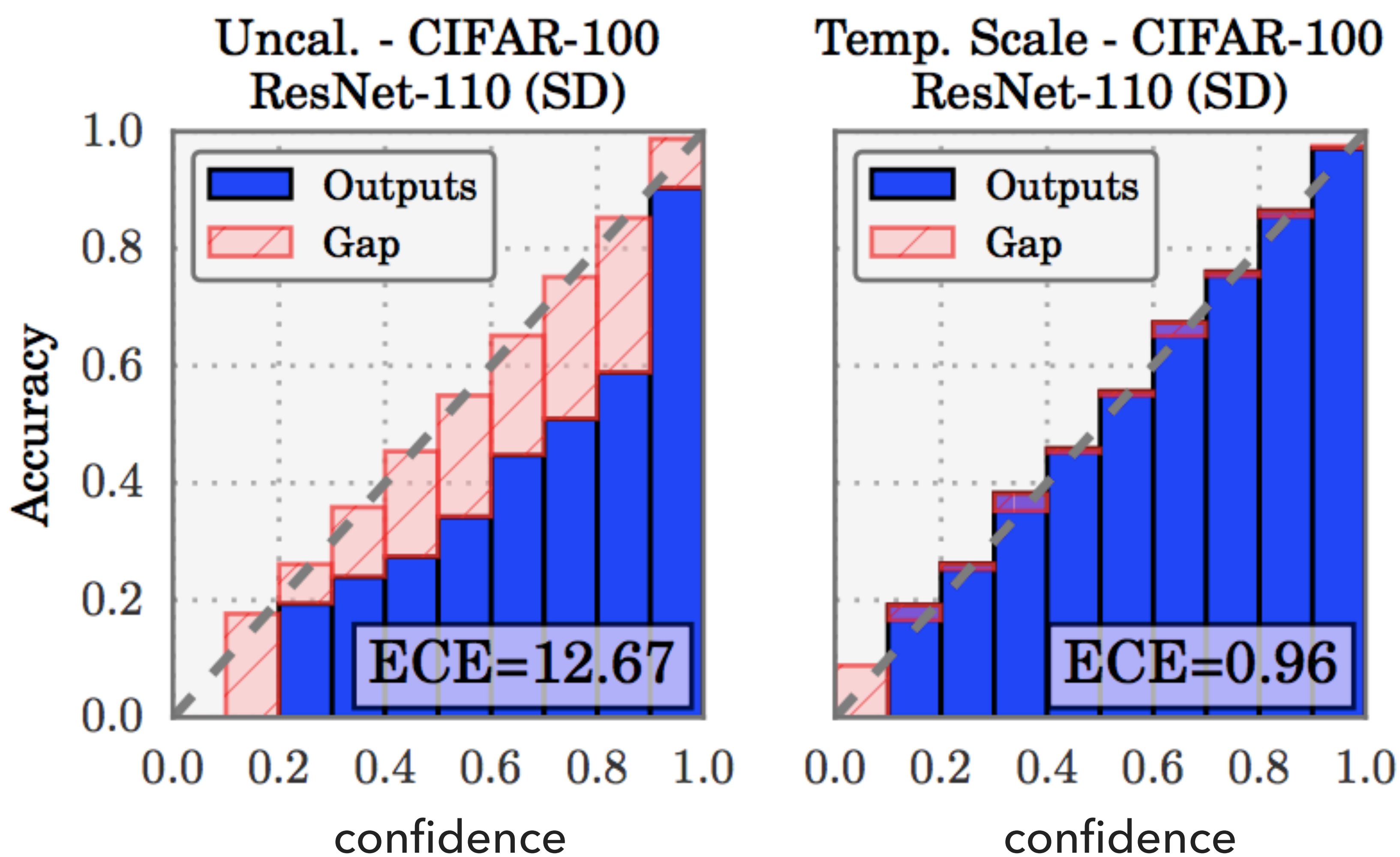
UNCERTAINTY: OVERCONFIDENCE IN NEURAL NETWORKS

- $p(y|x)$ should represent probabilities of belonging to a class
- Neural networks are often over-confident in their predictions



EXPECTED CALIBRATION ERROR (ECE)

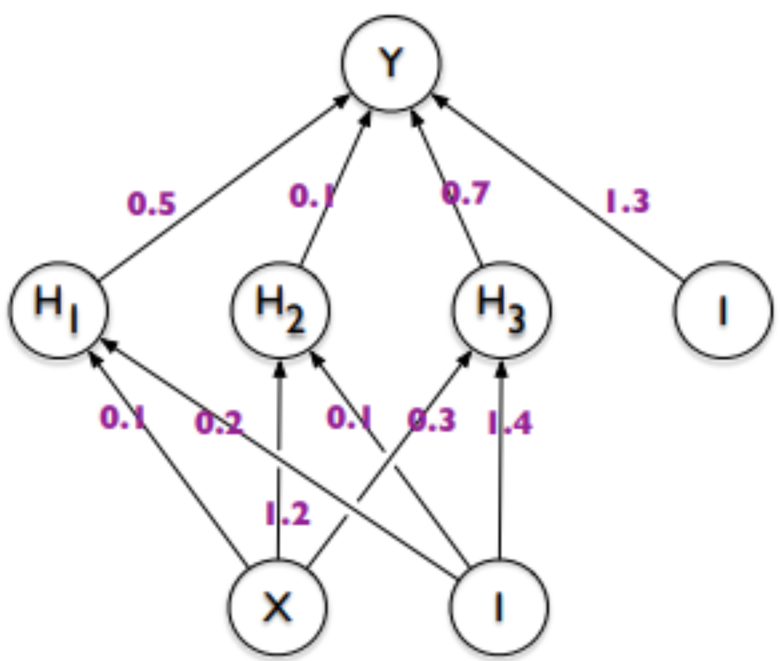
ECE is the expected difference between model's confidence and its accuracy



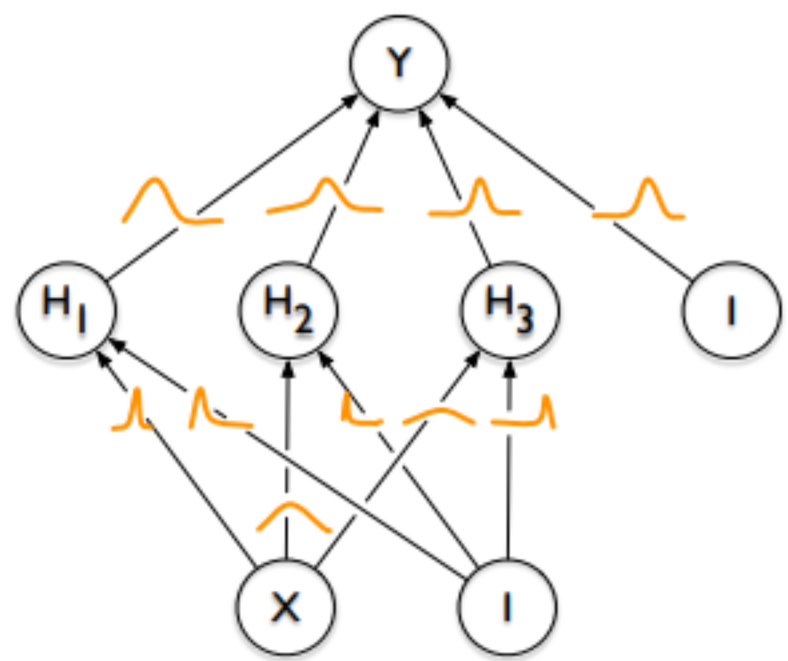
"On Calibration of Modern Neural Networks" by Chuan Guo, Geoff Pleiss, Yu Sun and Kilian Q. Weinberger

BAYESIAN DEEP LEARNING

- In Bayesian deep learning we model posterior distribution over the weights of neural networks
- In theory, leads to better predictions and well-calibrated uncertainty



Standard DNN



Bayesian DNN

"Weight Uncertainty in Neural Networks" by Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, Daan Wierstra

BAYESIAN DEEP LEARNING: CHALLENGES

Bayesian inference for deep neural networks is extremely challenging

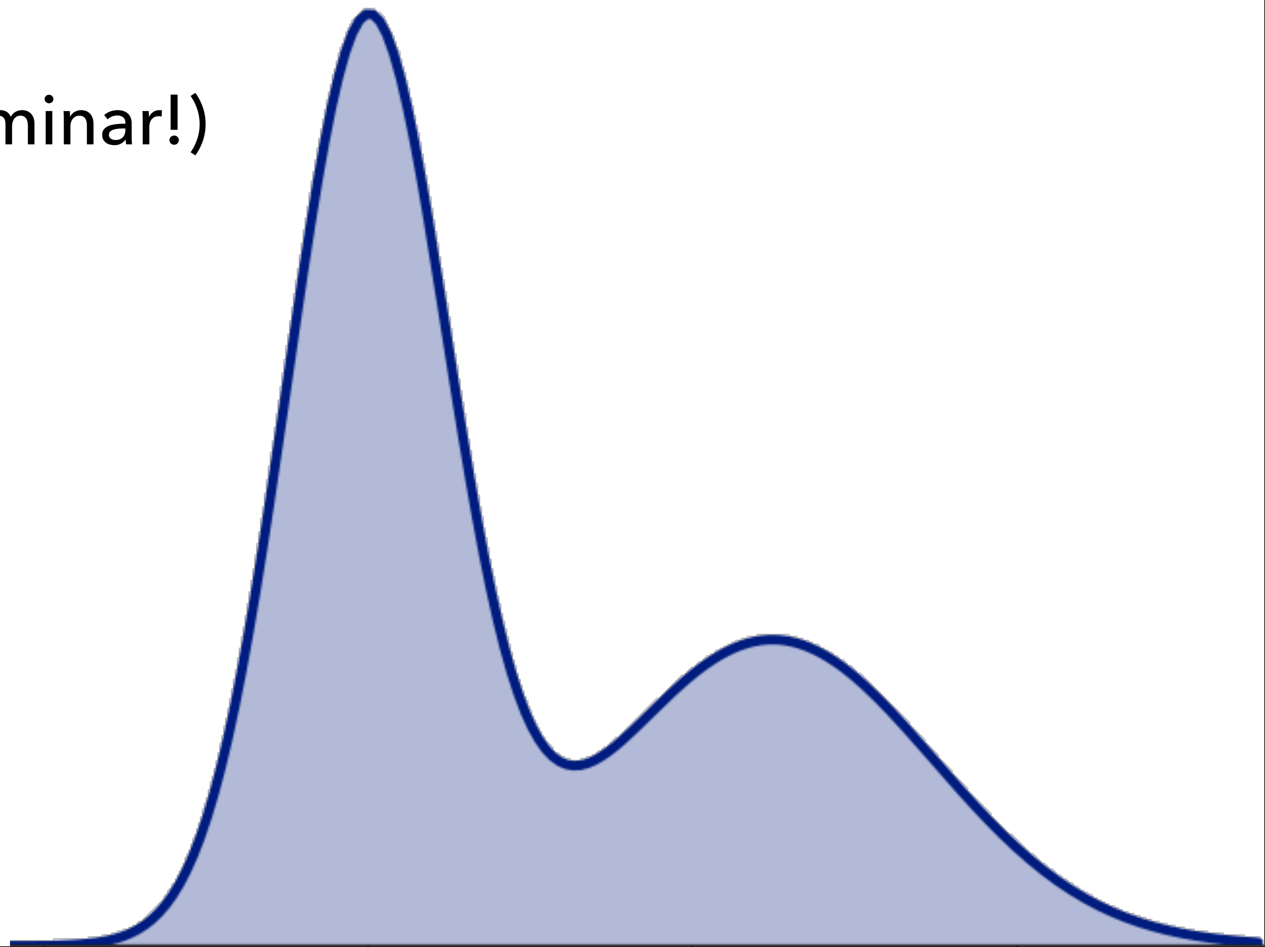
- Posterior is intractable
- Millions of parameters
- Large datasets
- Unclear which priors to use

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)} = \frac{p(D|w)p(w)}{\int_{w'} p(D|w')p(w')dw'}$$

HOW CAN WE DO APPROXIMATE BAYESIAN INFERENCE?

Posterior Approximation:

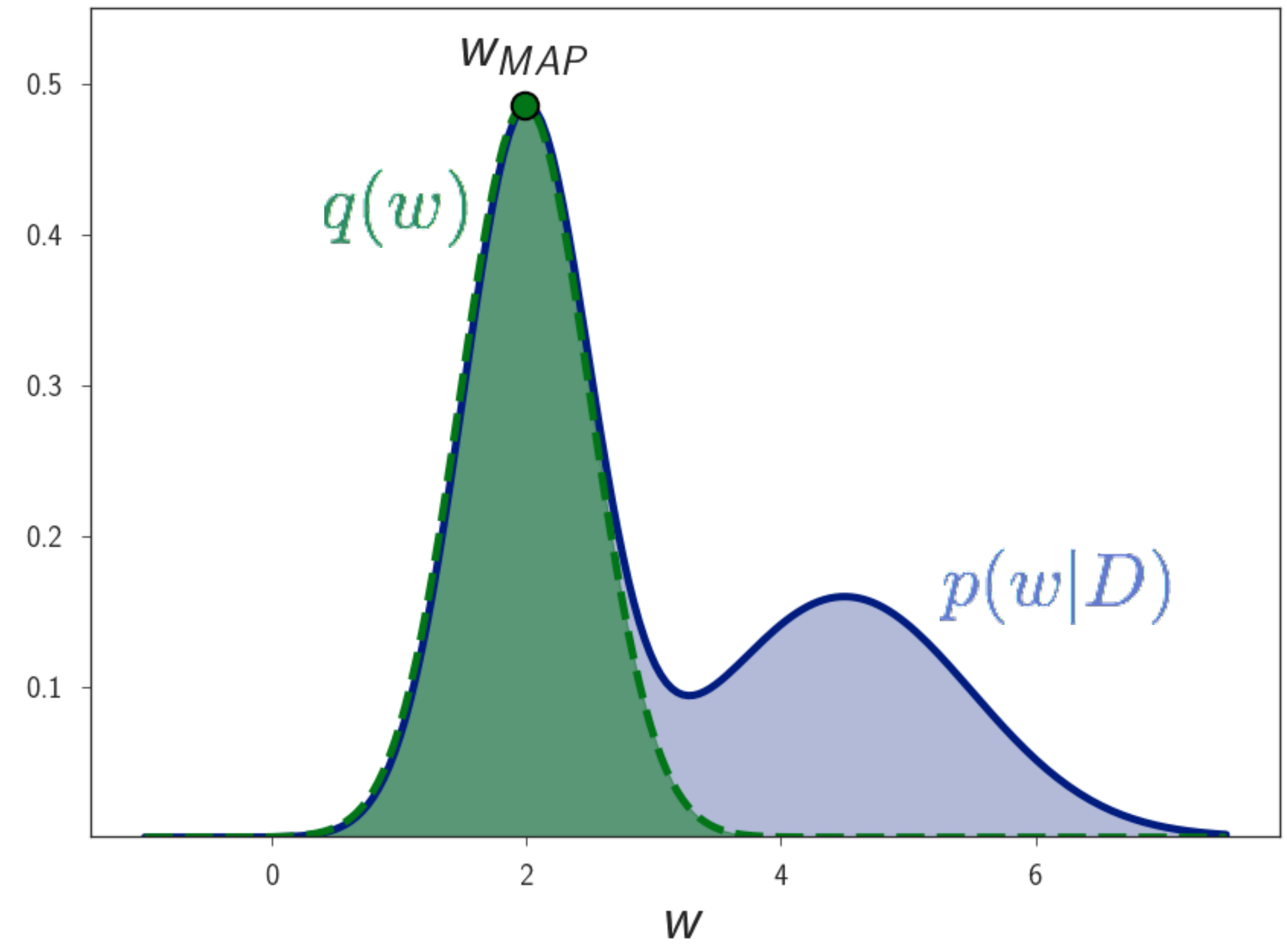
- Laplace Approximation
- Variational Inference
- Markov Chain Monte Carlo
- Geometrically Inspired Methods (see seminar!)



LAPLACE APPROXIMATION

Approximate posterior with a Gaussian $\mathcal{N}(w|\mu, A^{-1})$

- $w = w_{MAP}$ mode (local maximum) of $p(w|D)$
- $A = -\nabla\nabla \log[p(D|w)p(w)]$
- Only captures a single mode

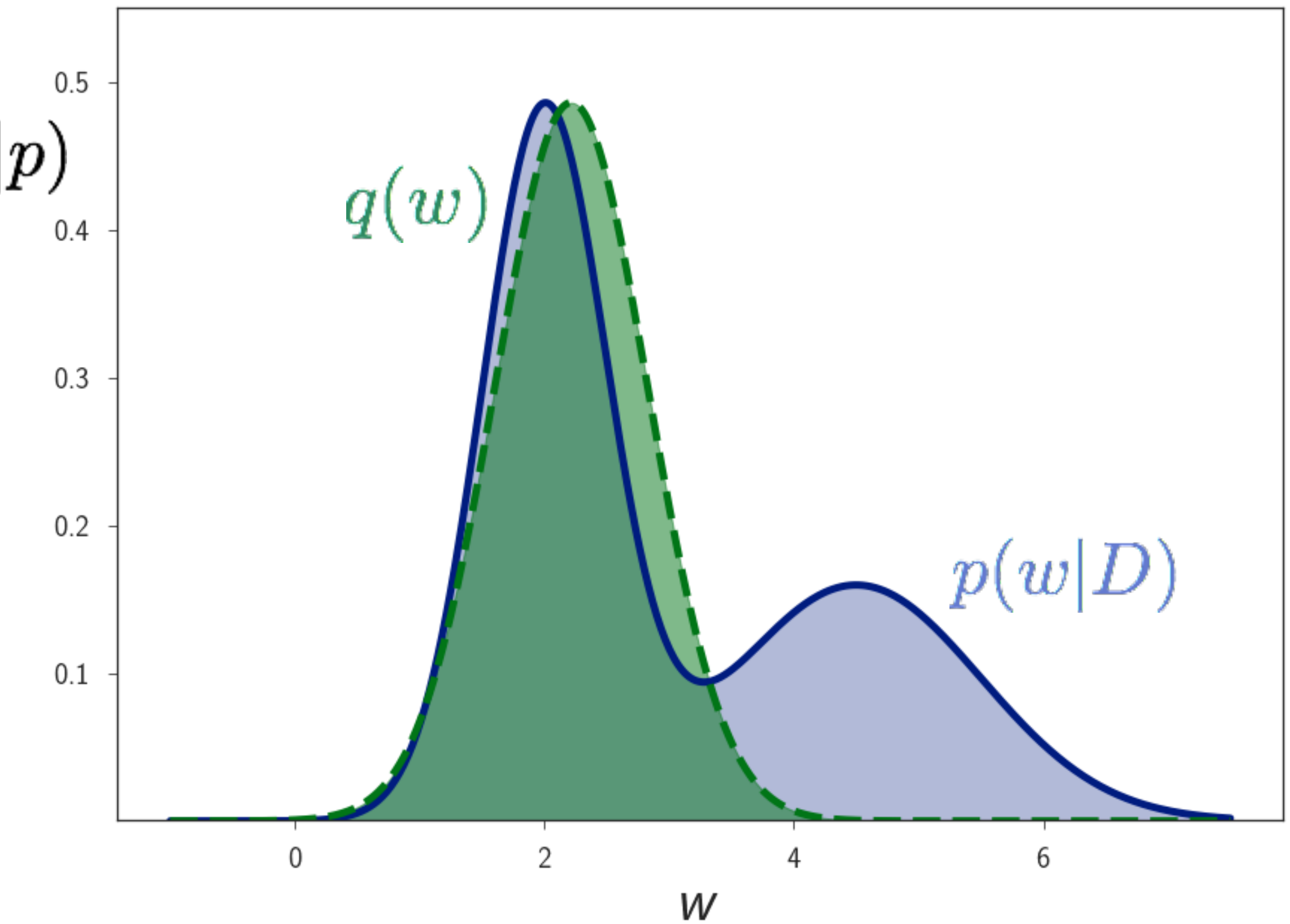


VARIATIONAL INFERENCE

We can find the best approximating distribution within a given family with respect to KL-divergence

- $KL(q||p) = \int_w q(w) \log \frac{q(w)}{p(w|D)} dw$

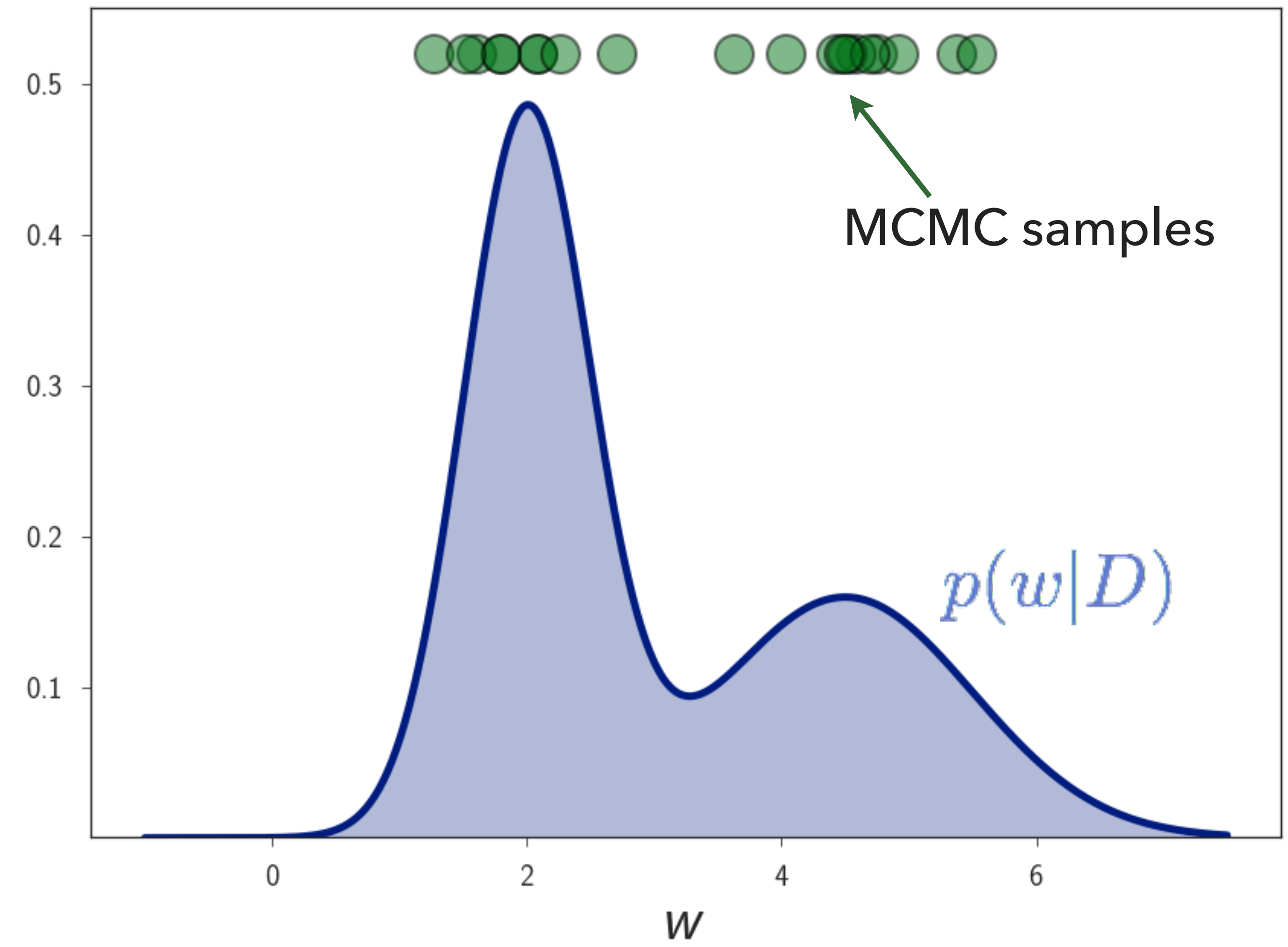
- If $q = \mathcal{N}(\mu, \Sigma)$, then $\min_{\mu,\Sigma} KL(q||p)$

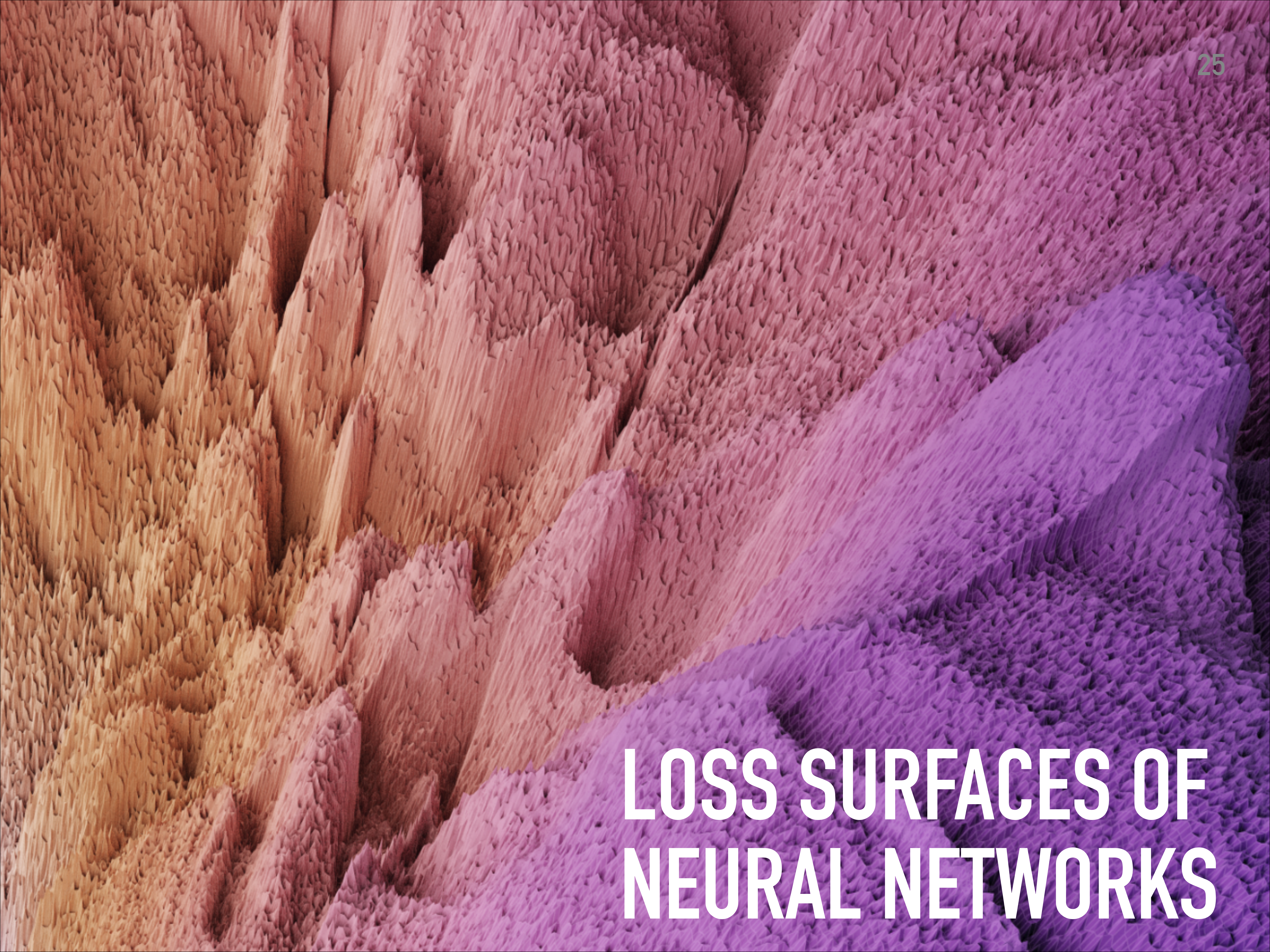


MARKOV CHAIN MONTE CARLO: SGLD

We can produce samples from the exact posterior by defining specific Markov Chains

- We can modify SGD to define a scalable MCMC sampler





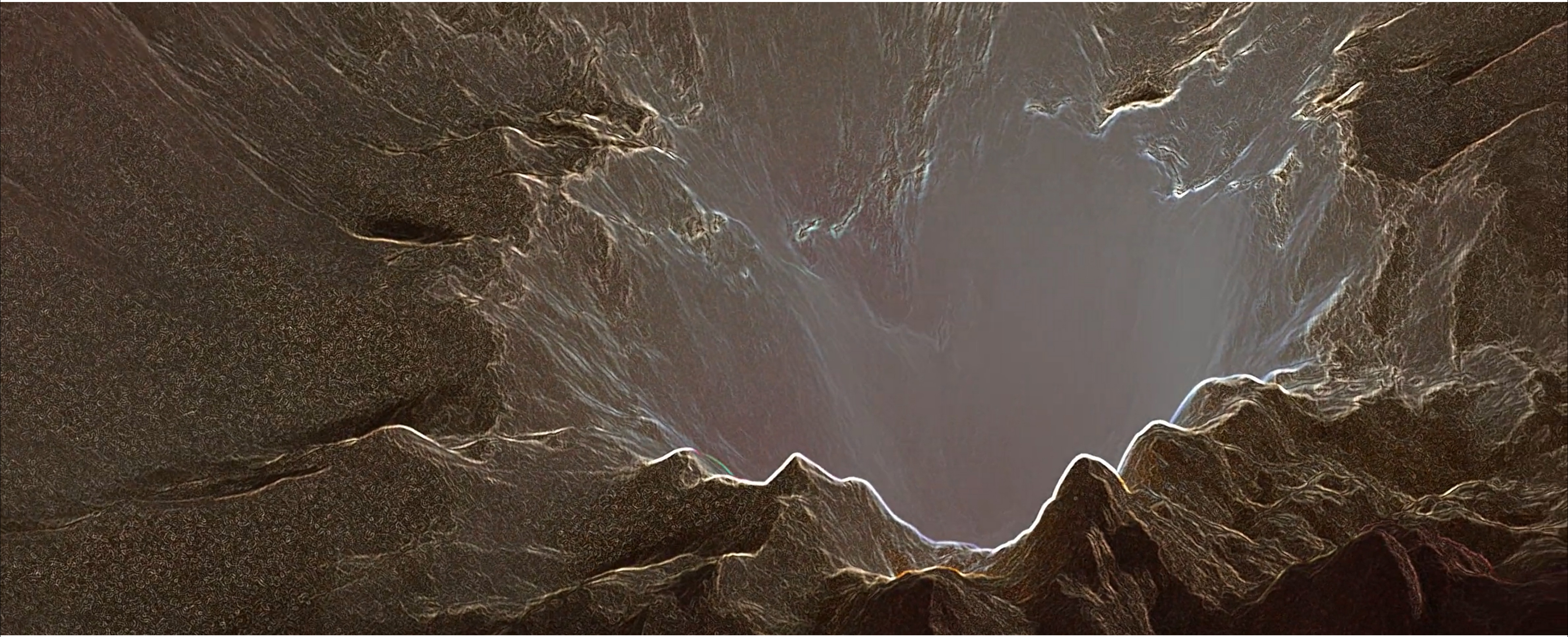
LOSS SURFACES OF NEURAL NETWORKS

LOSS SURFACES: WHY DO WE CARE?

- A tool for understanding generalization
- Better training methods motivated by geometric intuition
- Better approximate Bayesian Inference
  - $\text{loss} = -\log p(w|D)$, so understanding loss surfaces is crucial for approximate Bayesian inference

LOSS SURFACE VISUALIZATIONS

- We can use loss surface visualizations to better understand properties of DNNs

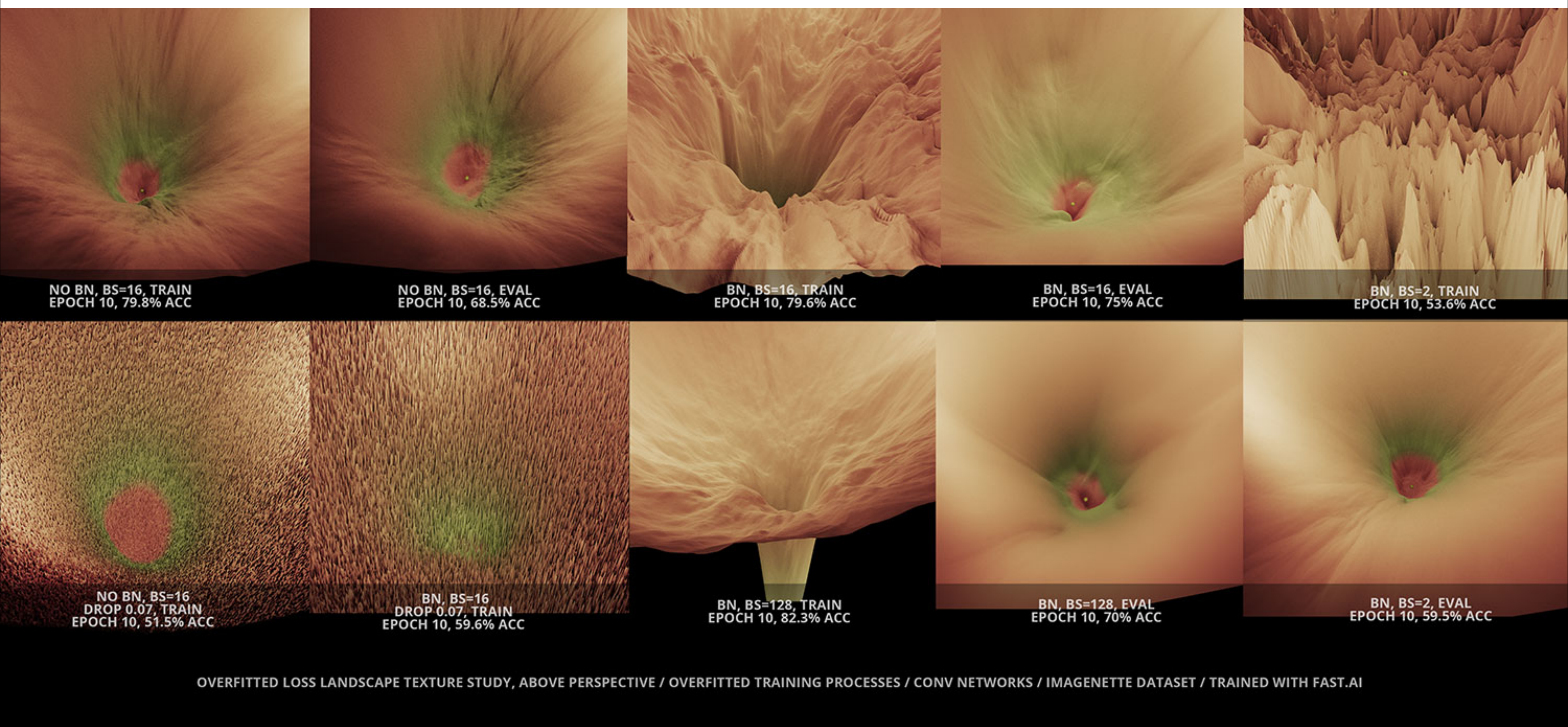


Visualizations created by Javier Ideami

More great visualizations available at https://losslandscape.com/

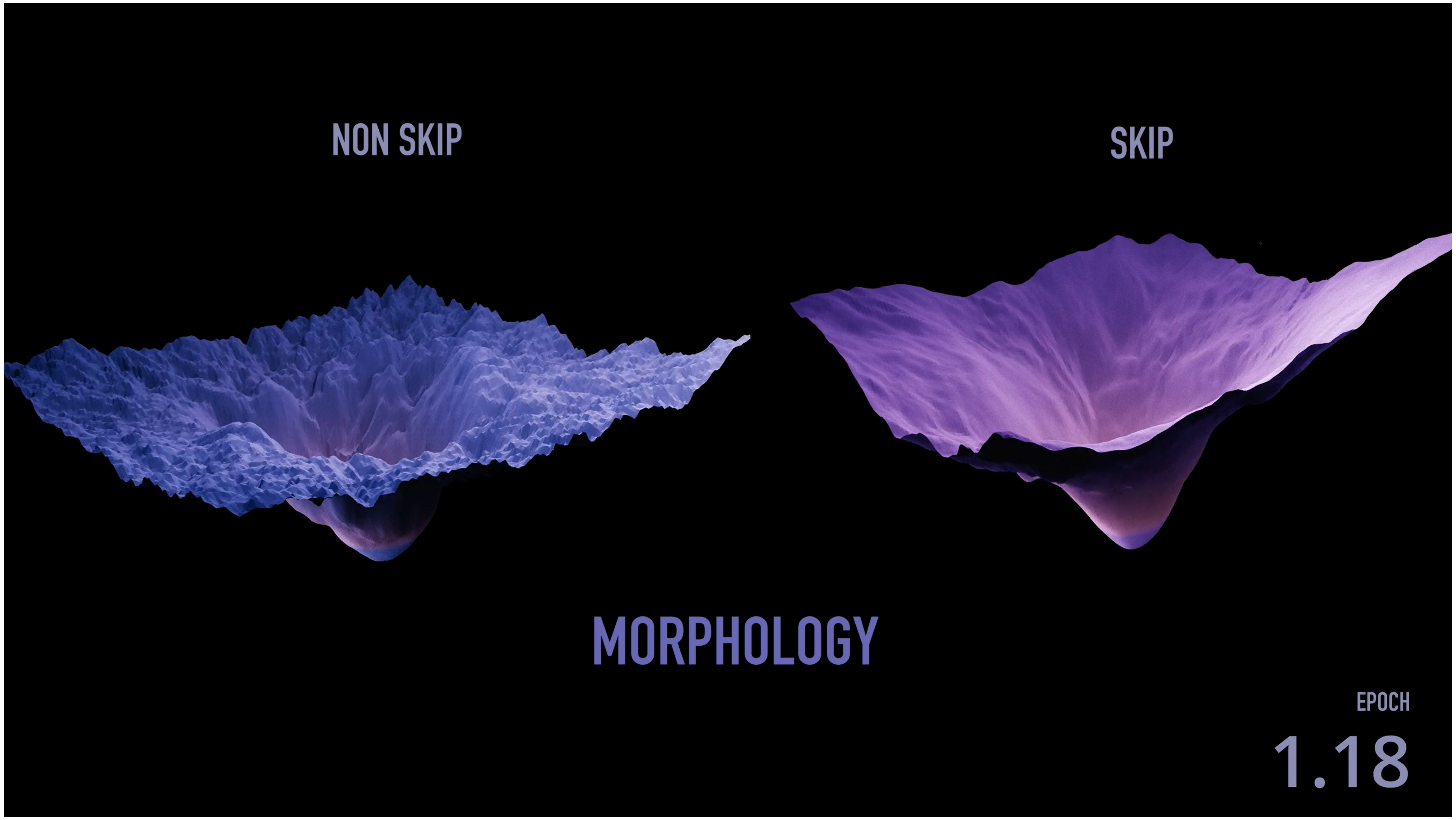
LOSS SURFACE VISUALIZATIONS

More great visualizations available at https://losslandscape.com/



LOSS SURFACE VISUALIZATIONS

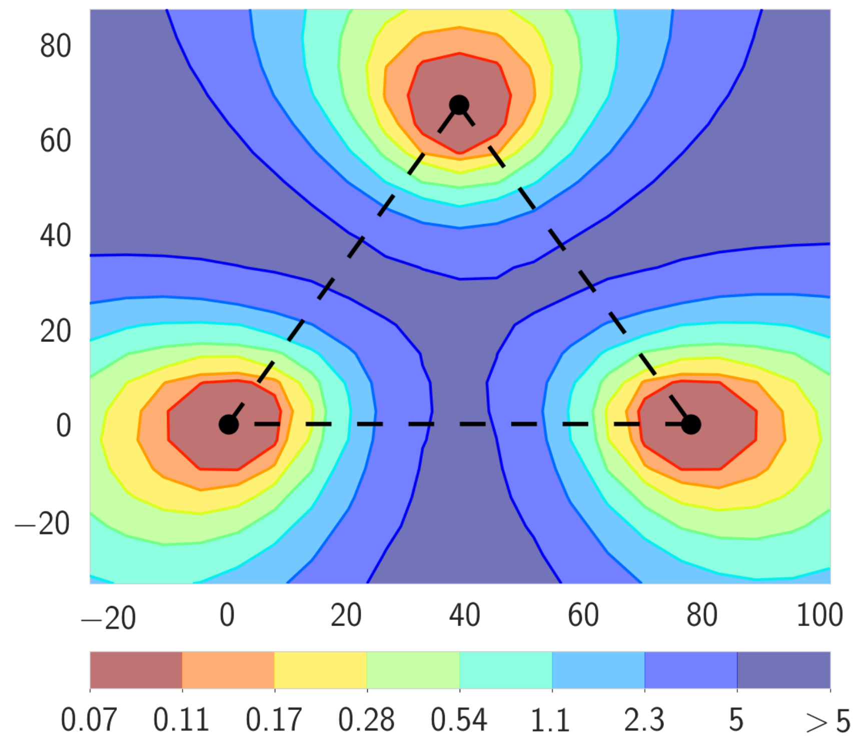
- Adding skip connections makes the loss landscape more smooth



First reported in "Visualizing the Loss Landscape of Neural Nets" by Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer and Tom Goldstein

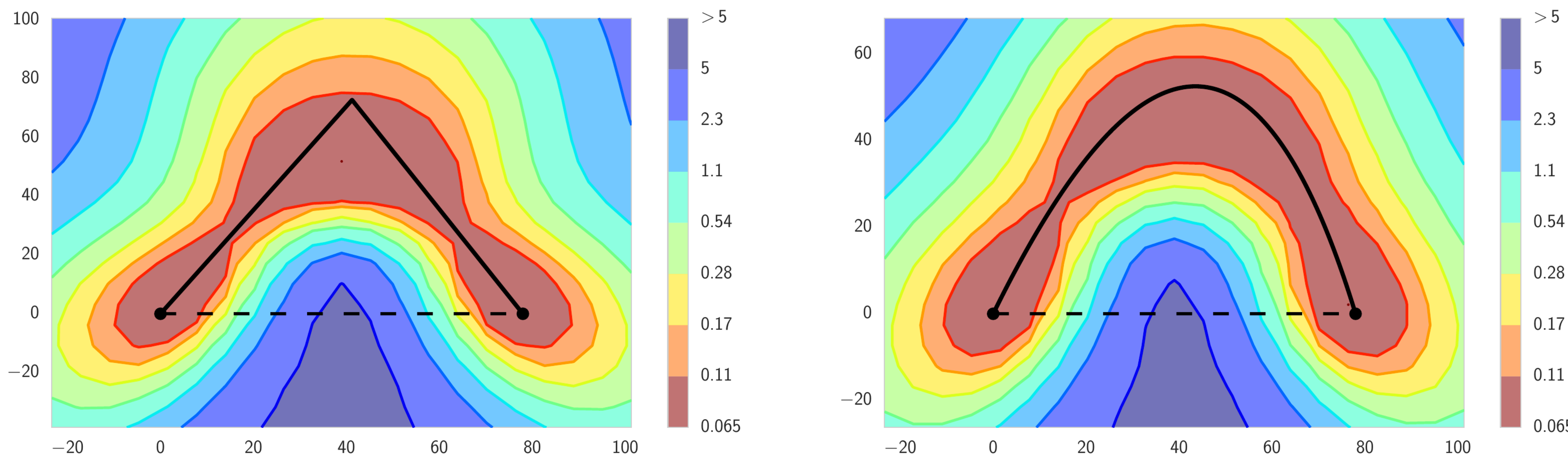
MODE CONNECTIVITY AND GLOBAL STRUCTURE OF LOSS LANDSCAPES

- When we train networks from different initializations, we get different solutions
- If we look along line segment connecting independently trained solutions, loss goes up a lot
- Does this mean local optima are isolated from each other?

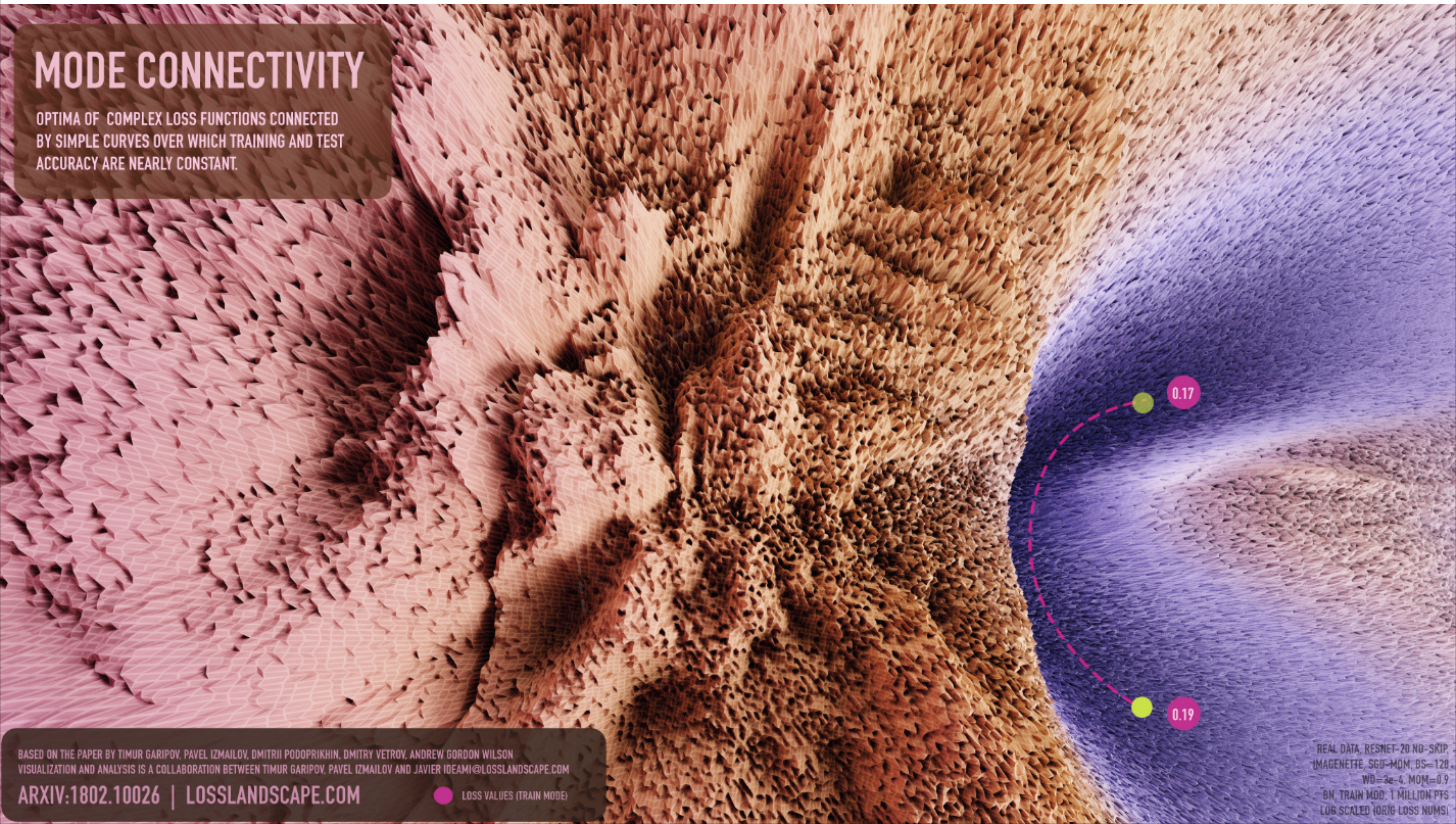


MODE CONNECTIVITY:

- Turns out independently trained solutions can be connected by paths of low loss
- These paths are very simple to find and can take very simple shapes



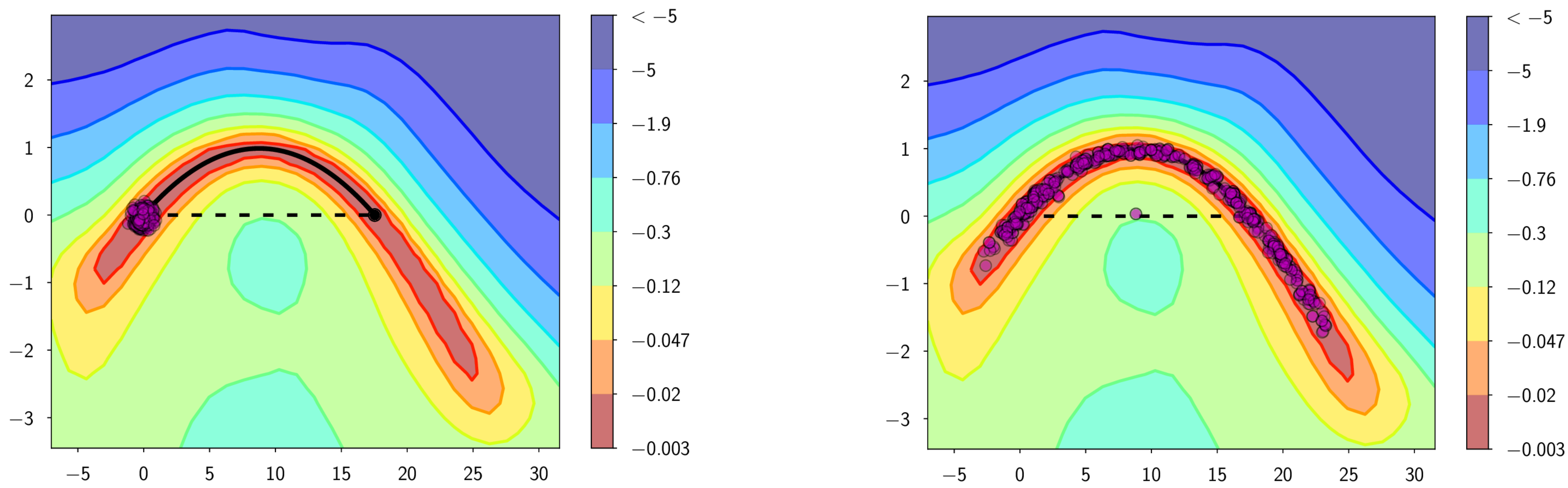
MODE CONNECTIVITY VISUALIZATION



MODE CONNECTIVITY VISUALIZATION

LOSS SURFACES: WHY DO WE CARE?

- $\text{loss} = -\log p(w|D)$, so understanding loss surfaces is crucial for approximate Bayesian inference



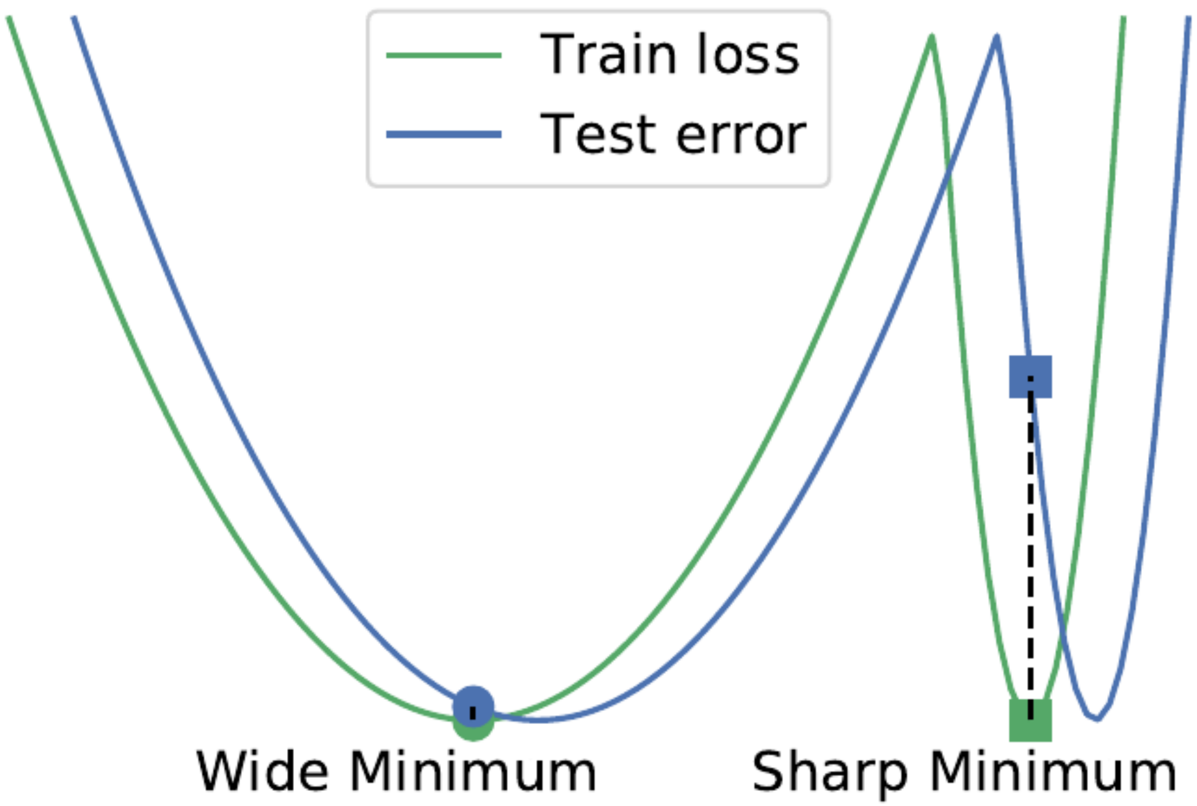
MODE CONNECTIVITY: IMPLICATIONS

- Fast ensembling methods
- Better training methods
- Better approximate Bayesian deep learning

See the seminar talk!

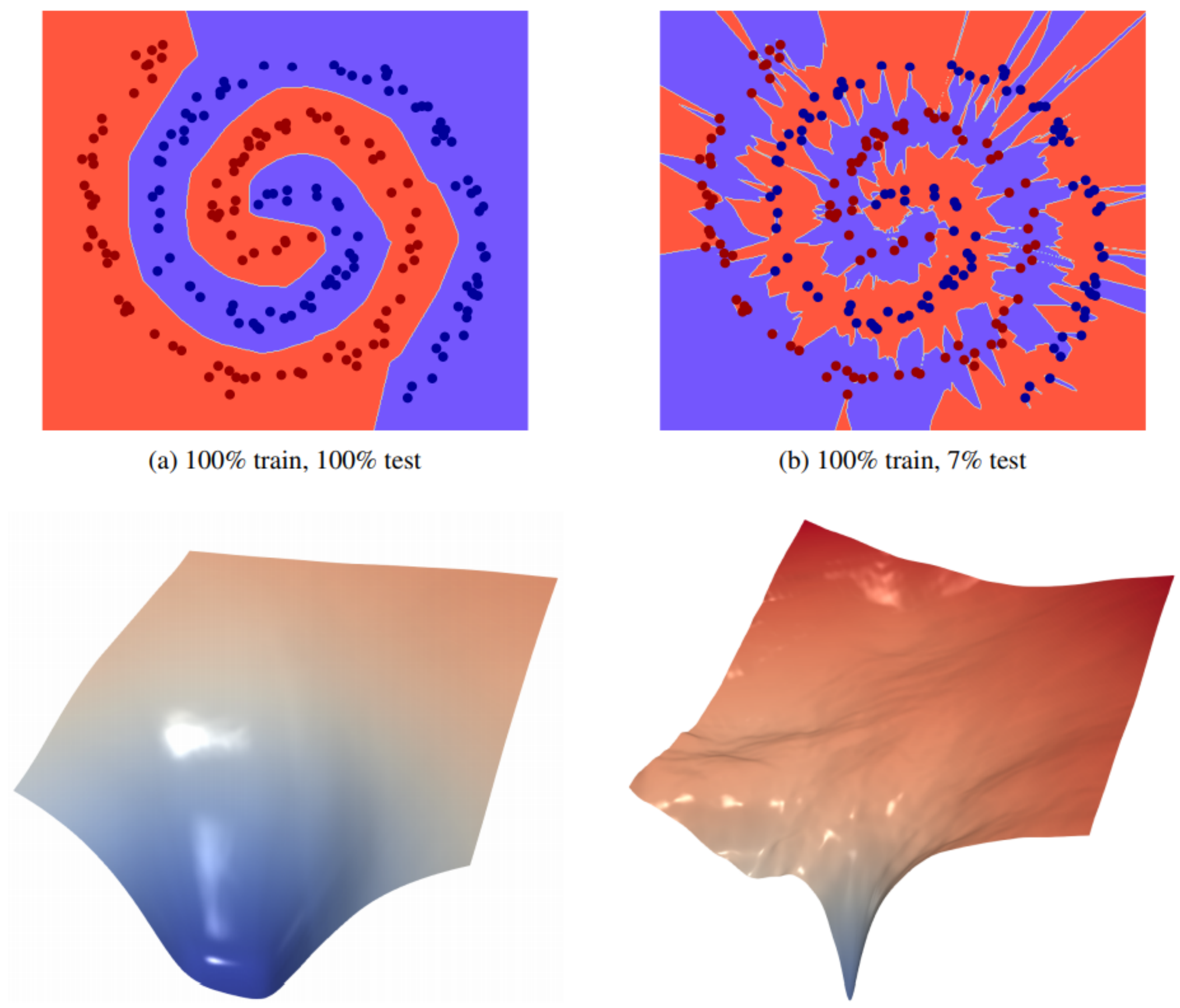
FLATNESS AND GENERALIZATION

- Intuitively, flat solutions that lie in flat regions of the loss surface should generalize better



FLATNESS AND GENERALIZATION

- Intuitively, flatness corresponds to higher margin



(a) 100% train, 100% test

(b) 100% train, 7% test

(c) Minimizer of network in (a) above

(d) Minimizer of network in (b) above

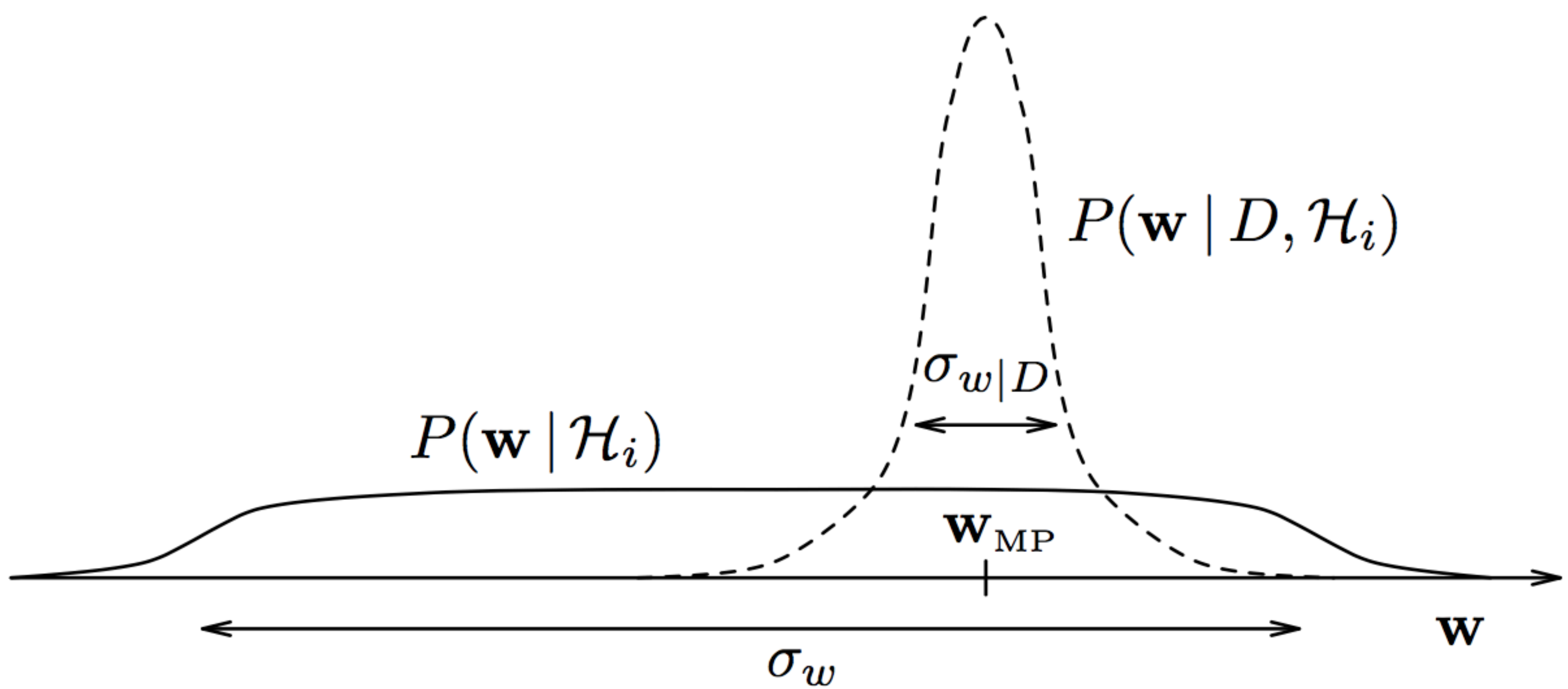
Results from "Understanding Generalization through Visualizations" by W. Ronny Huang, Zeyad Emam, Micah Goldblum, Liam Fowl, Justin K. Terry, Furong Huang, Tom Goldstein

FLATNESS AND GENERALIZATION: BAYESIAN PERSPECTIVE

$$p(D) = \int p(D|w)p(w)dw$$

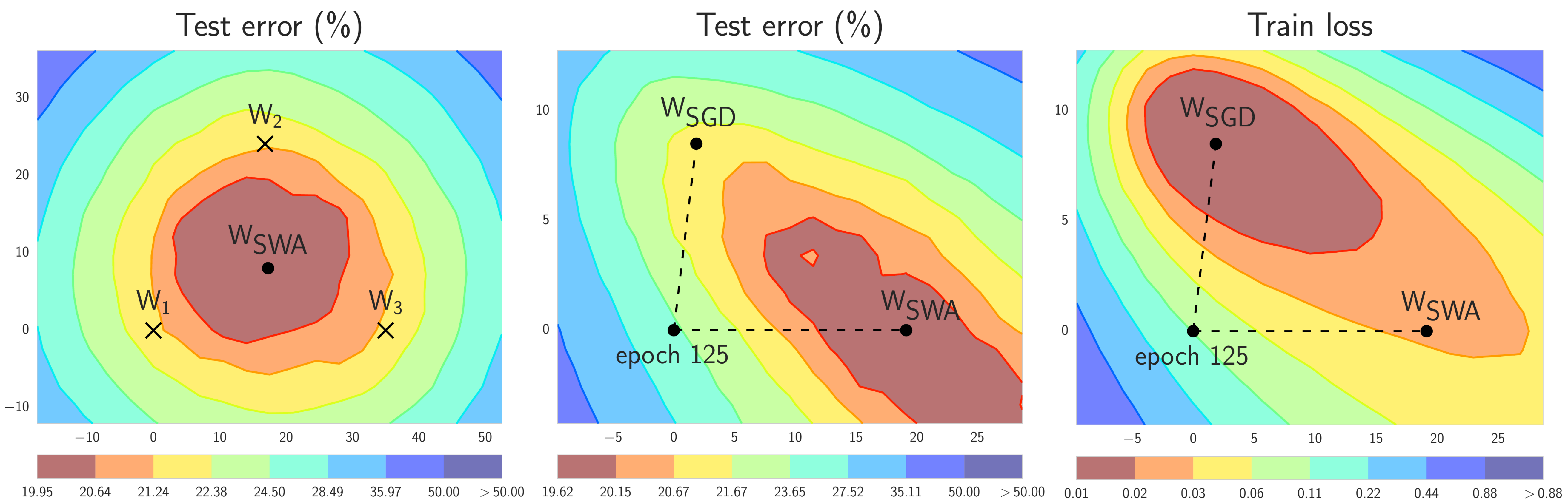
$$p(D) \approx p(D|w_{MP})\underbrace{p(w_{MP})}_{1/\sigma_w}\sigma_{w|D}$$

$$\text{Occam factor} = \frac{\sigma_{w|D}}{\sigma_w}$$



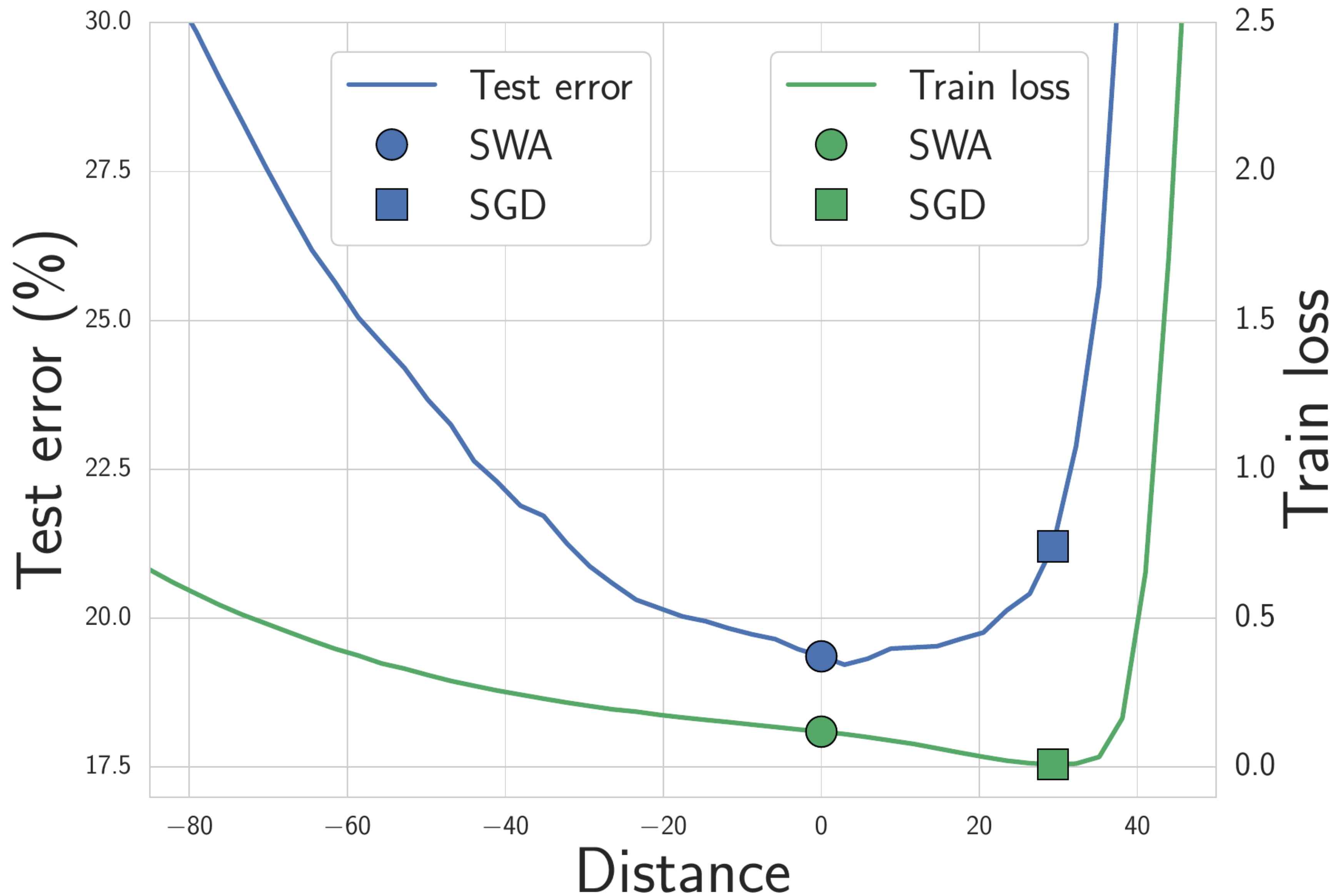
SWA: GEOMETRICALLY MOTIVATED TRAINING METHOD

- SWA is a training method for deep learning, motivated by mode connectivity.



SWA: GEOMETRICALLY MOTIVATED TRAINING METHOD

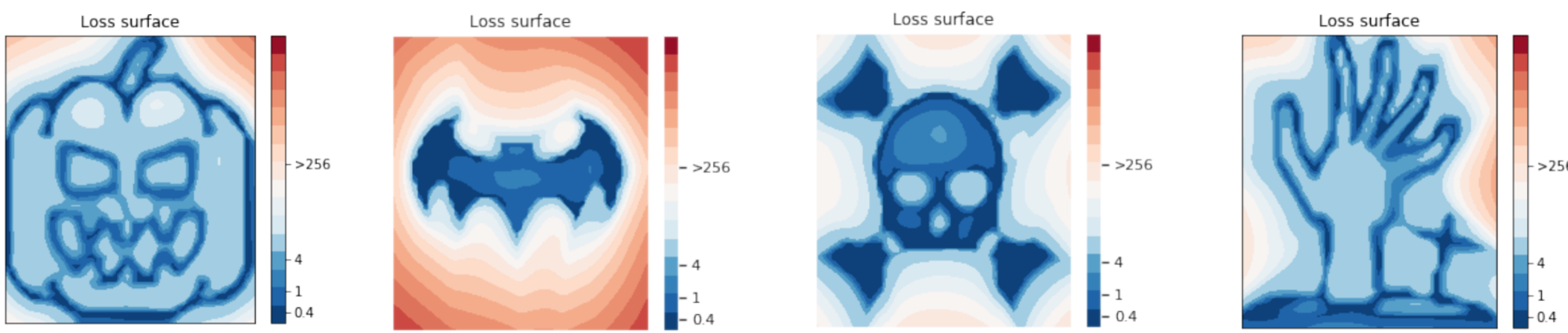
- SWA finds a flatter solution in the loss surface, and achieves better generalization



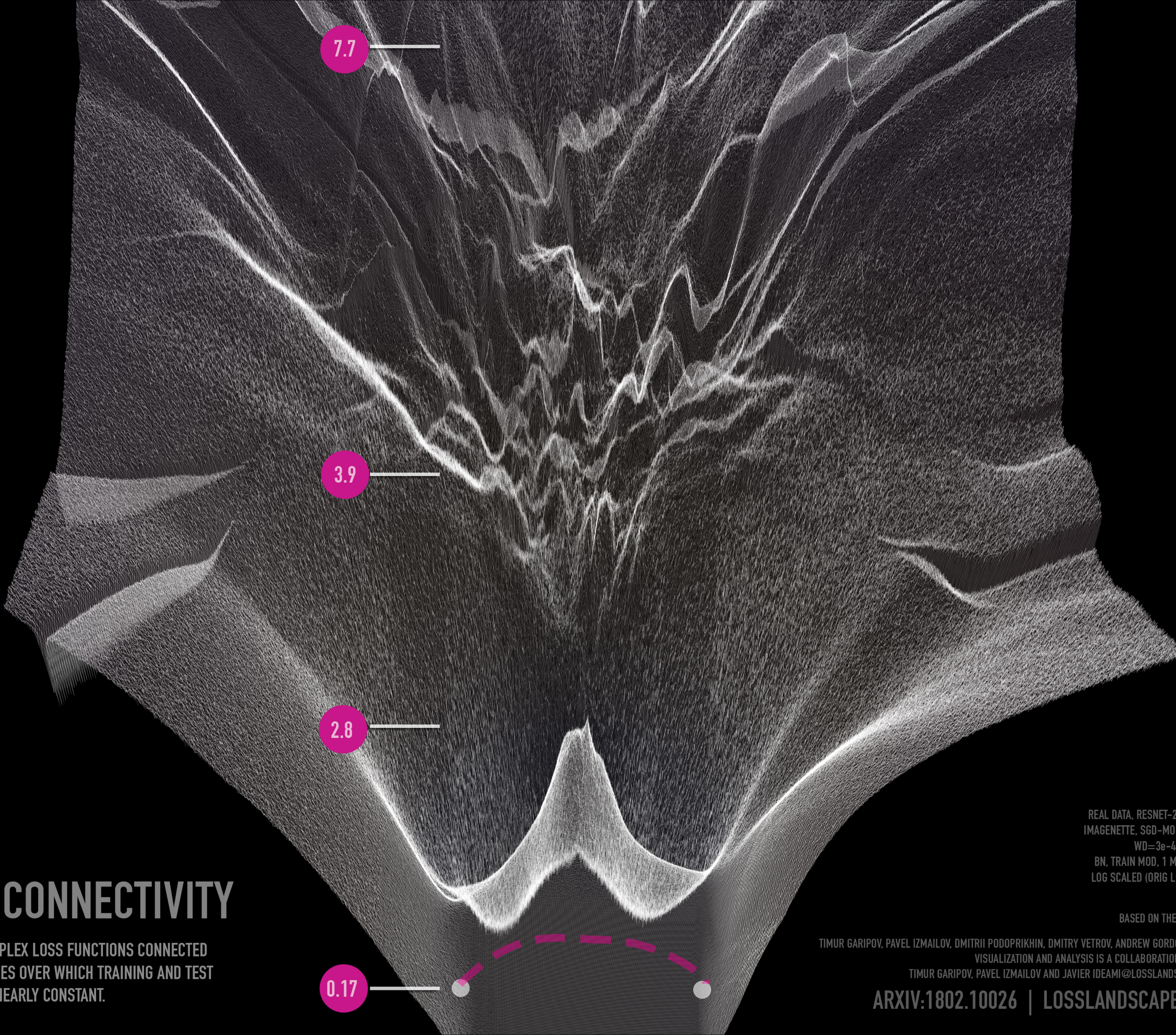
LOSS SURFACES: WHAT DO WE KNOW?

Loss Surfaces of Neural Networks are extremely complex:

- Live in million-dimensional parameter spaces
- Highly Non-convex
- Very Multimodal



Results from "Loss Landscape Sightseeing with Multi-Point Optimization" by Ivan Skorokhodov and Mikhail Burtsev



MODE CONNECTIVITY

OPTIMA OF COMPLEX LOSS FUNCTIONS CONNECTED BY SIMPLE CURVES OVER WHICH TRAINING AND TEST ACCURACY ARE NEARLY CONSTANT.

REAL DATA, RESNET-20 NO-SKIP,
IMAGENETTE, SGD-MOM, BS=128
WD=3e-4, MOM=0.9
BN, TRAIN MOD, 1 MILLION PTS
LOG SCALED (ORIG LOSS NUMS)

BASED ON THE PAPER BY

TIMUR GARIPOV, PAVEL IZMAILOV, DMITRII PODOPRIKHIN, DMITRY VETROV, ANDREW GORDON WILSON
VISUALIZATION AND ANALYSIS IS A COLLABORATION BETWEEN
TIMUR GARIPOV, PAVEL IZMAILOV AND JAVIER IDEAMI@LOSSLANDSCAPE.COM

ARXIV:1802.10026 | LOSSLANDSCAPE.COM