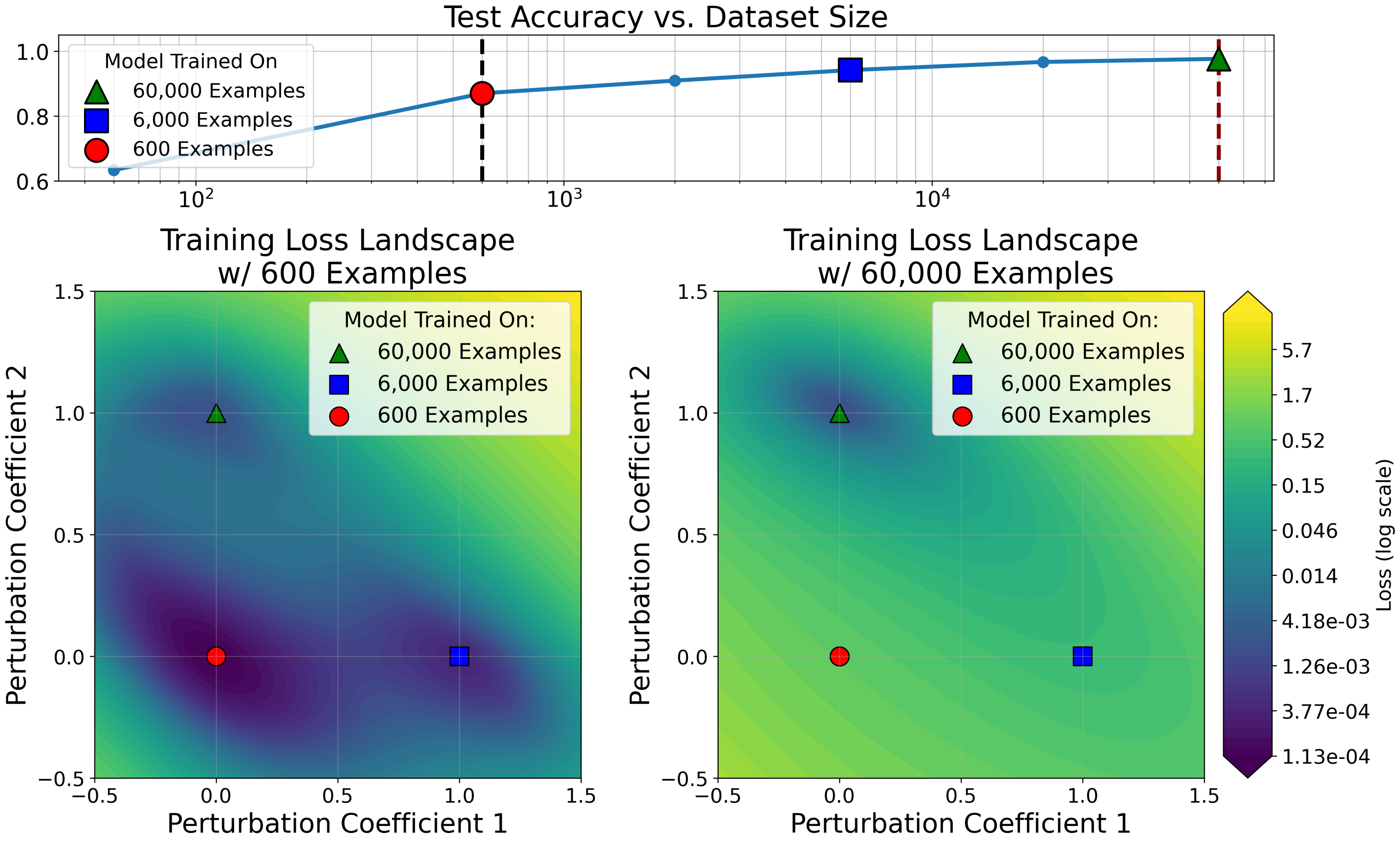
Why Is Generalization Hard?

Why Do We Need Data?


Test Accuracy vs. Dataset Size
1.0
0.8
0.6
10^2
10^3
10^4
Model Trained On
▲ 60,000 Examples
■ 6,000 Examples
● 600 Examples
Training Loss Landscape w/ 600 Examples
Training Loss Landscape w/ 60,000 Examples
Model Trained On:
▲ 60,000 Examples
■ 6,000 Examples
● 600 Examples
Perturbation Coefficient 1
Perturbation Coefficient 2
−0.5
0.0
0.5
1.0
1.5
Loss (log scale)
5.7
1.7
0.52
0.15
0.046
0.014
4.18e-03
1.26e-03
3.77e-04
1.13e-04

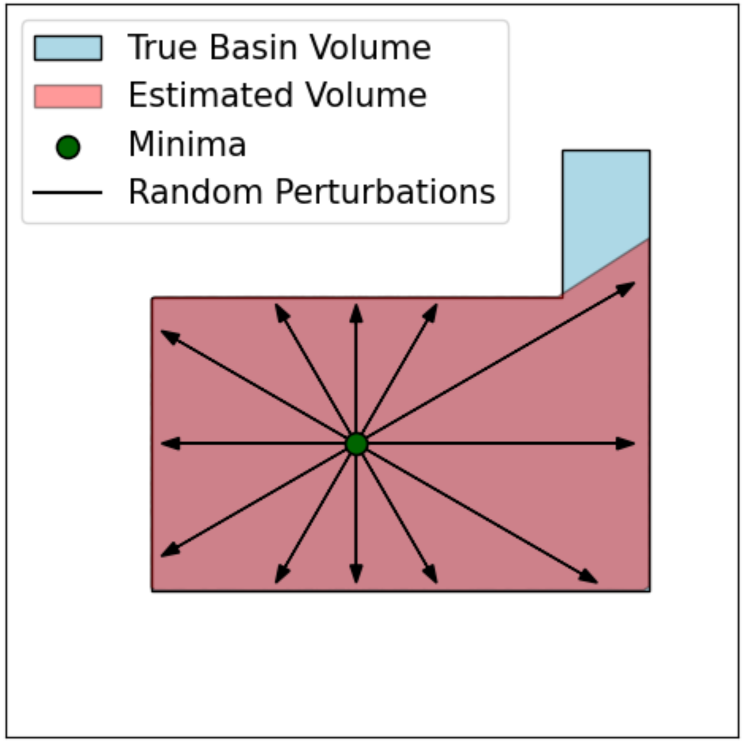
Generalization - Volume Hypothesis

- Claims:
  - Flat minima generalize well (1997)
  - Gradient Descent finds flat minima because they're large (2020)
- Generalizing is easy?

Hochreiter, Sepp and Schmidhuber, Jürgen. Flat minima. 1997
Huang, W. Ronny, et al. "Understanding generalization through visualizations." (2020)

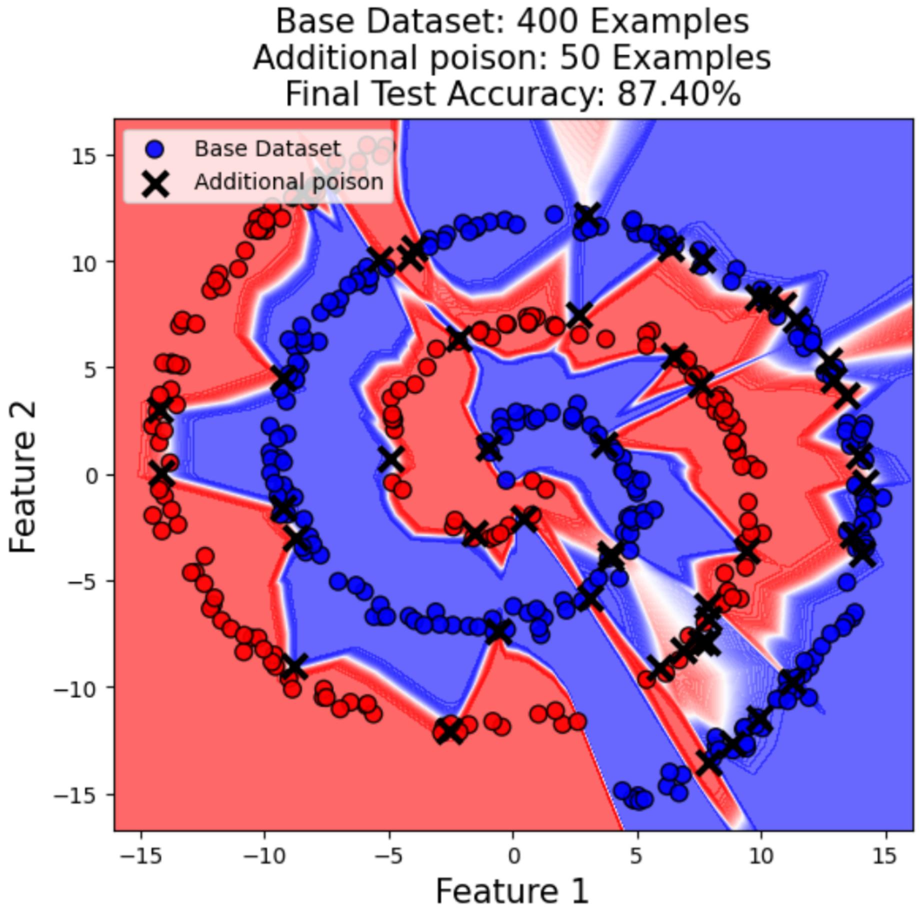
How Can We Measure Minima?

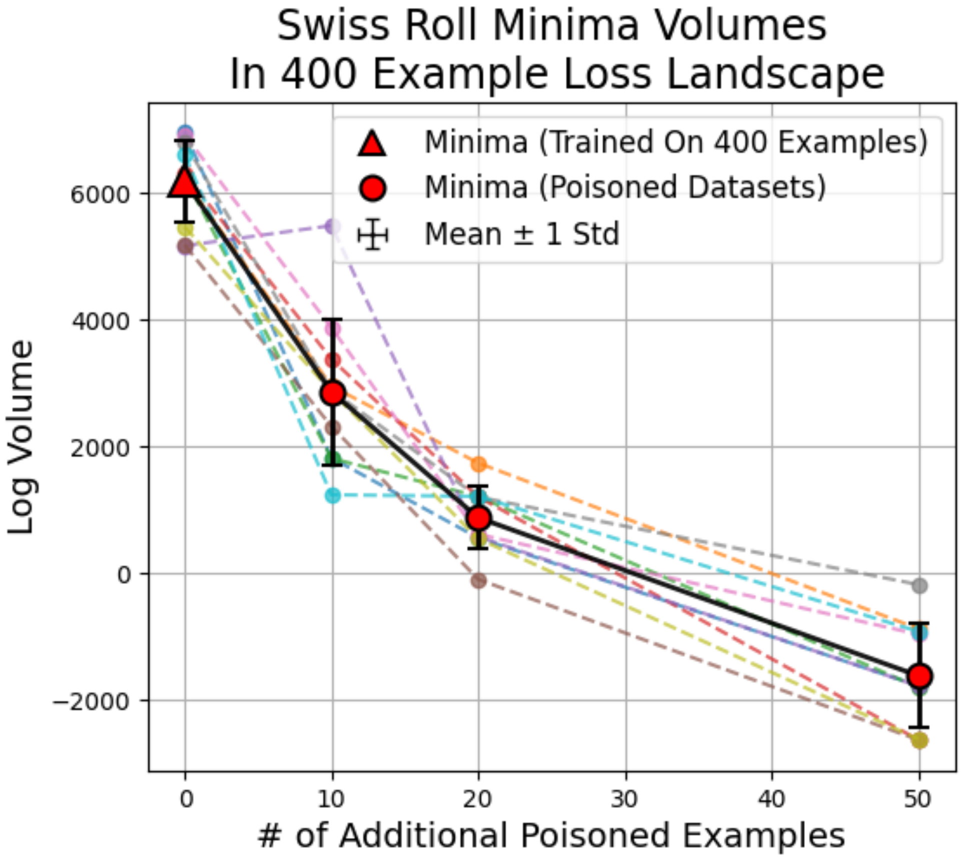
- How much can weights can be perturbed without increasing loss?
- **Basin Volume Estimation** - start from a minima, take random vectors, see loss increases
  - Monte Carlo = need many perturbs?
  - Underestimates true volume...



Existing Results – Poisoned Minima

- Generate bad minima, measure volumes
- Poisoned minima have much smaller volumes!



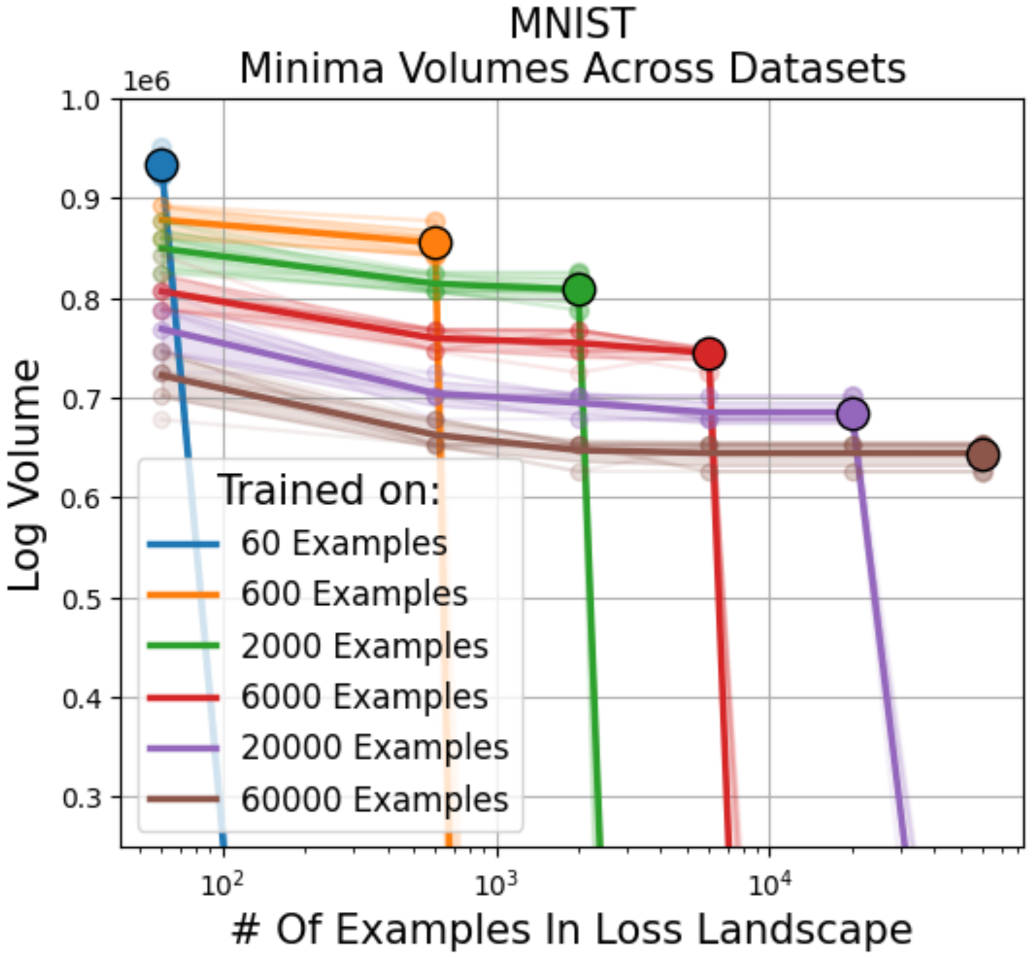


Limitations – Only Compared Poisoned Minima

- Explains why bad minima don’t happen
- But deep learning needs data. Why?
    - What are the volumes of good minima?
    - How does data change minima volumes?

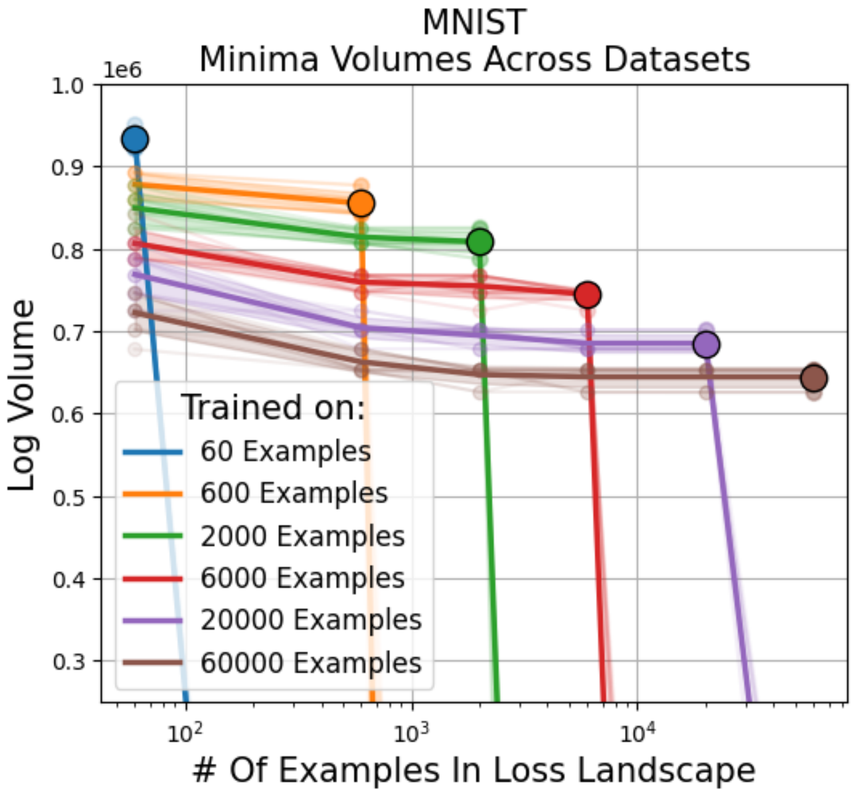
MNIST - GD Finds Large Minima, But Sharp Minima Are Better

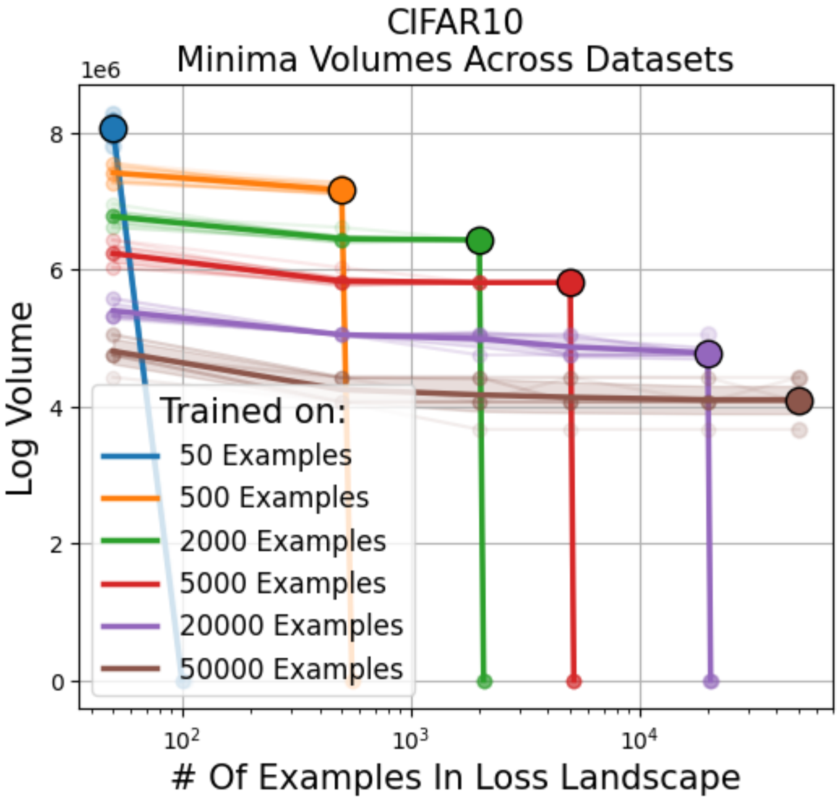
- Found minima are largest in its loss landscape
- Minima from larger datasets generalize better but are smaller
- Data shrinks all minima, largest minima disappears



MNIST and CIFAR10 – Volume Data Scaling Law?

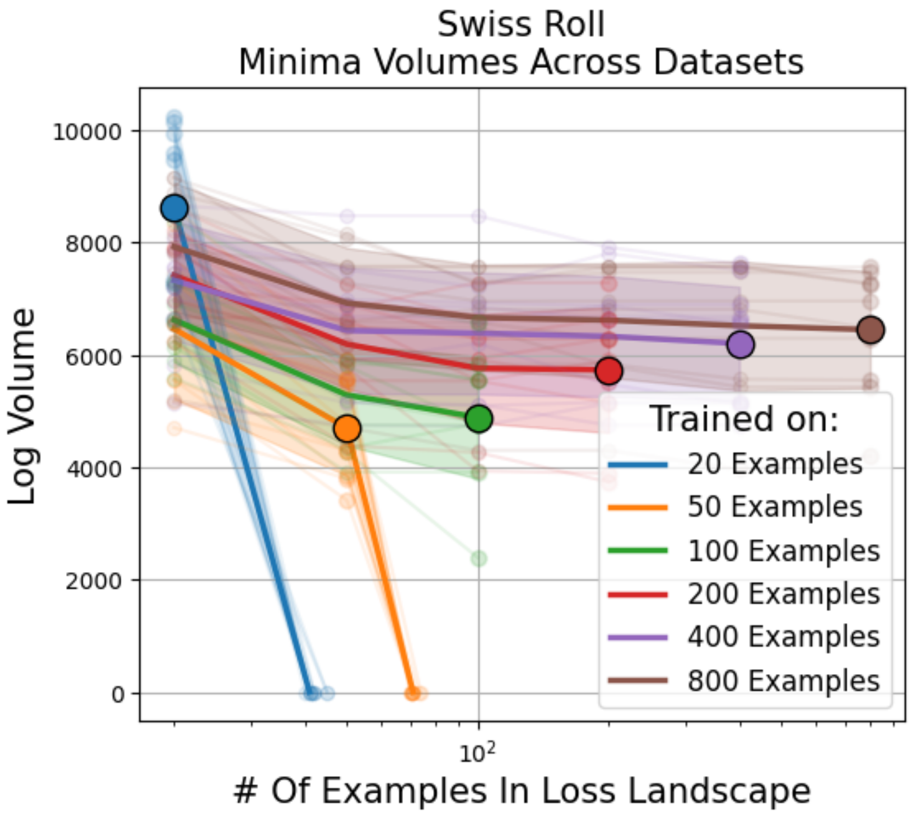
- Volume and data sizes are predictable?
- Scaling laws are model dependent (CNN vs MLP)

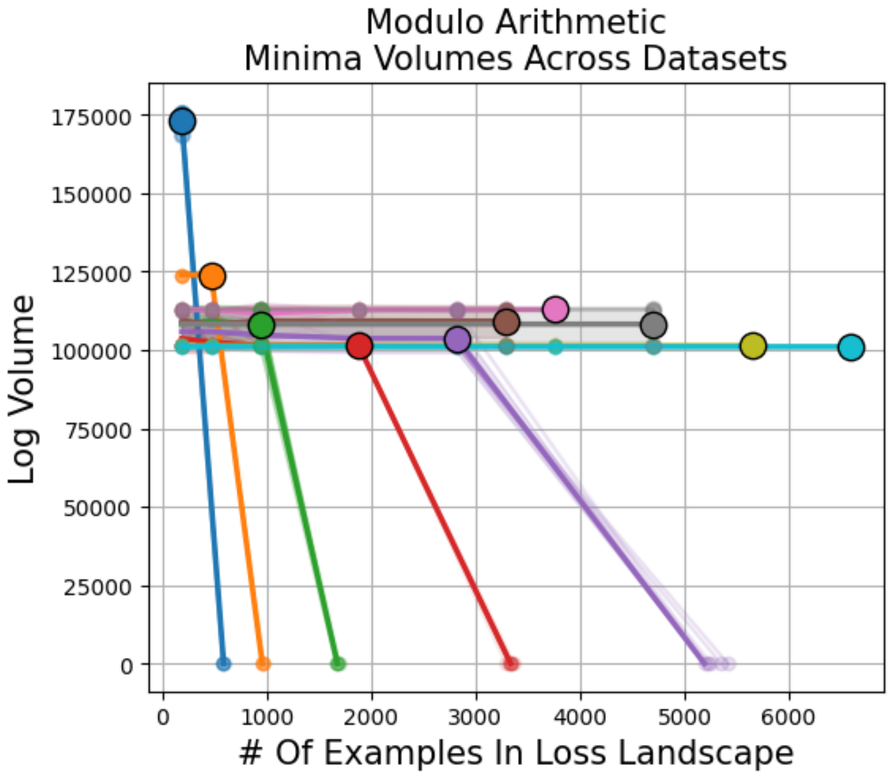




Sometimes You Find Small Minima Swiss Roll, Modulo Arithmetic

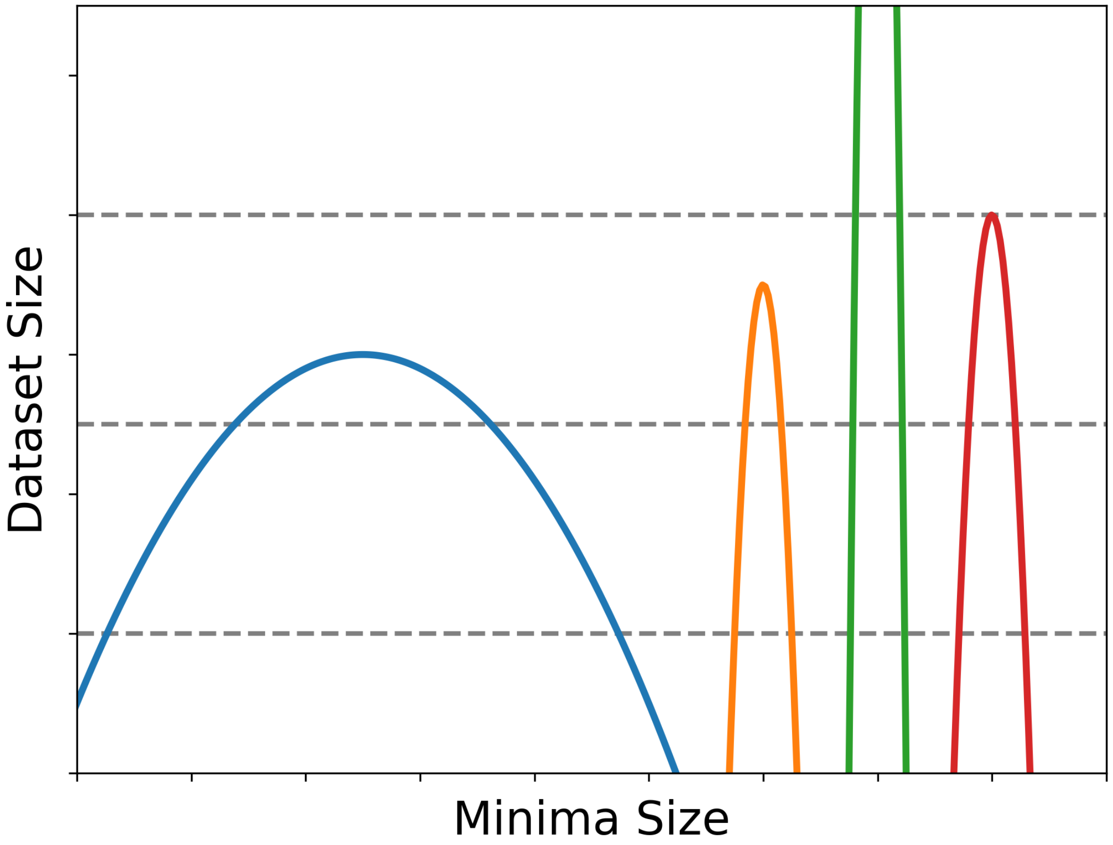
- Relationship for MNIST + CIFAR10 is not universal





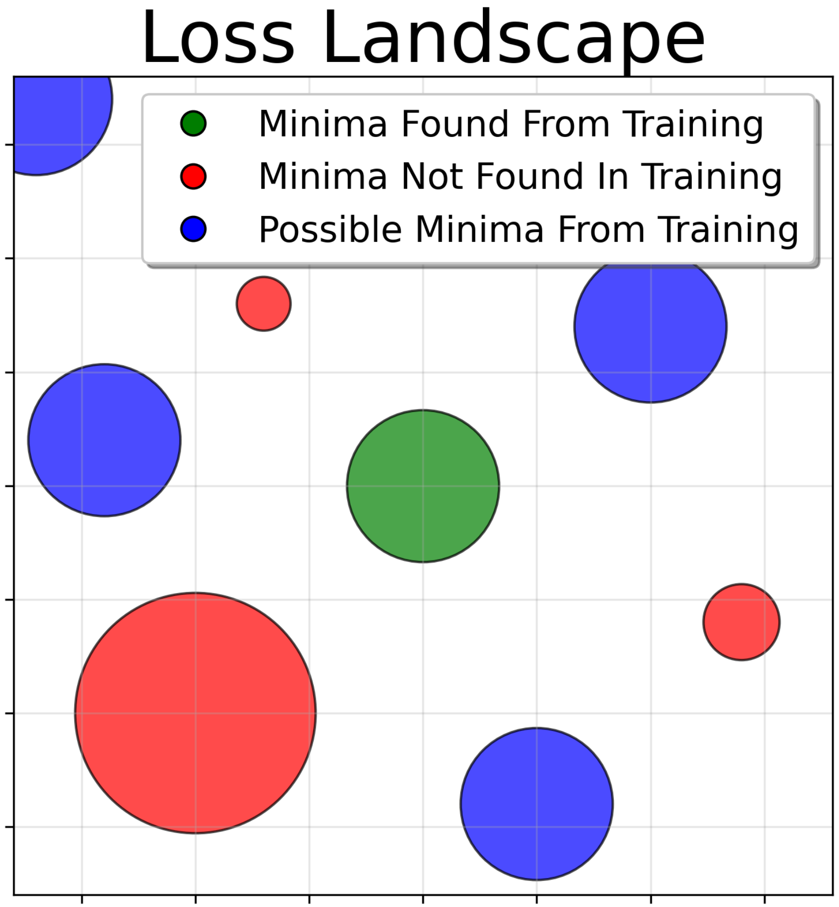
Take Away: Mild Picture Of Generalization

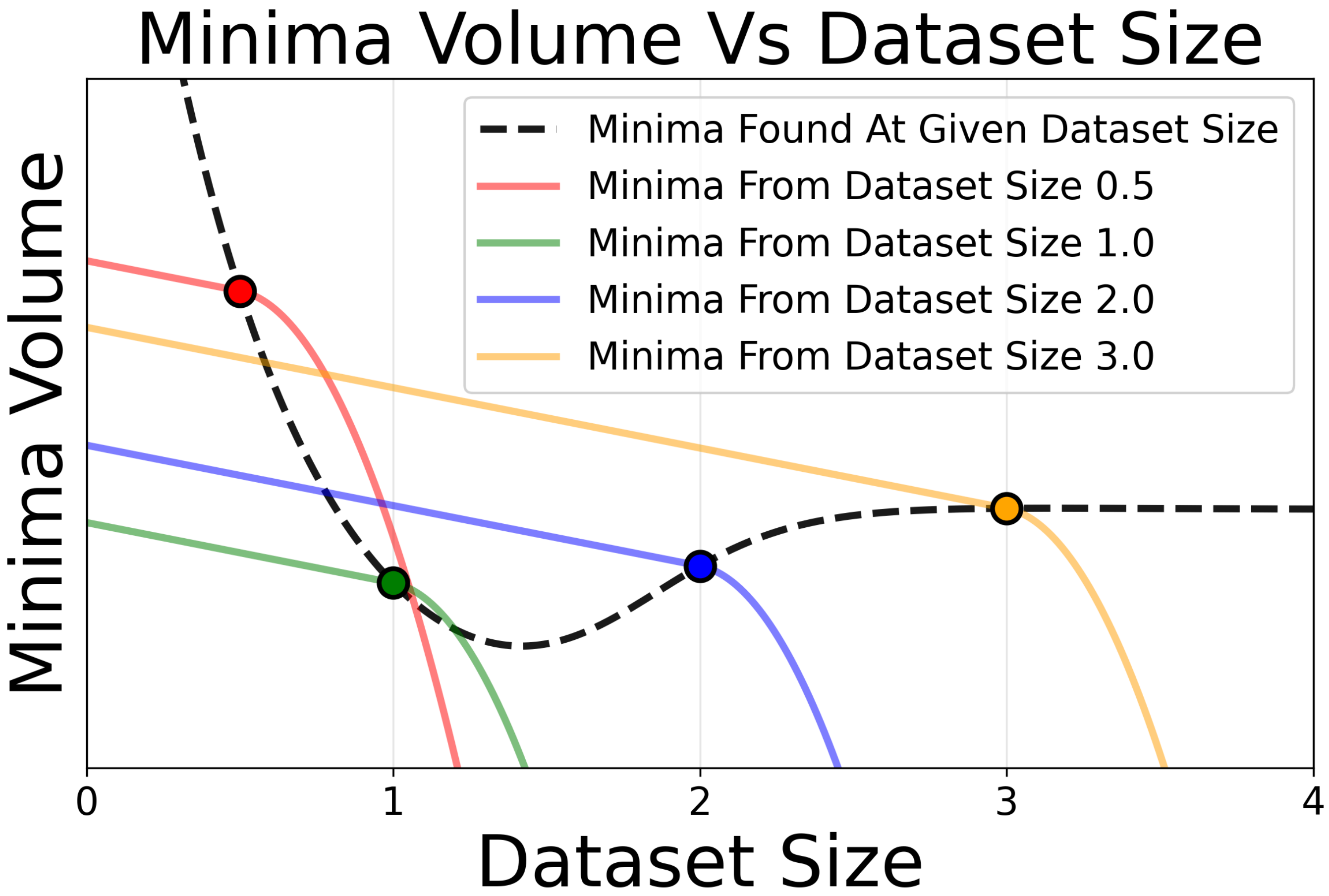
- Flat minima generalize well
- We often find large minima
- BUT there are sharp minima that generalize well
- Unlikely to find because of small volume
- Adding data changes the largest minima so we start to find them



Extra: Overall Volume != Individual Volume

- Maybe some minima classes are common, and have large volume overall


Loss Landscape
● Minima Found From Training
● Minima Not Found In Training
● Possible Minima From Training


Minima Volume Vs Dataset Size