D-semble: Efficient Diversity-Guided Search for Resilient ML Ensembles

Abraham Chan,

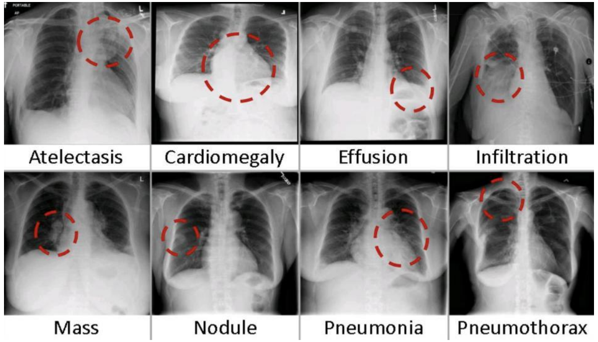
Arpan Gujarati, Karthik Pattabiraman, Sathish Gopalakrishnan

Training Data Faults in Practice

70% of Lyft dataset missing, mislabelled [Kang et al, 2022]

Autonomous Vehicles

20% of ChestX-ray mislabelled [Tang et al, 2021]



Healthcare

Training Data Faults

Image	Label
	~~Speed 80km/h~~ Stop sign
	~~Stop sign~~ Speed 80km/h
	No entry

Normal
Mislabeling

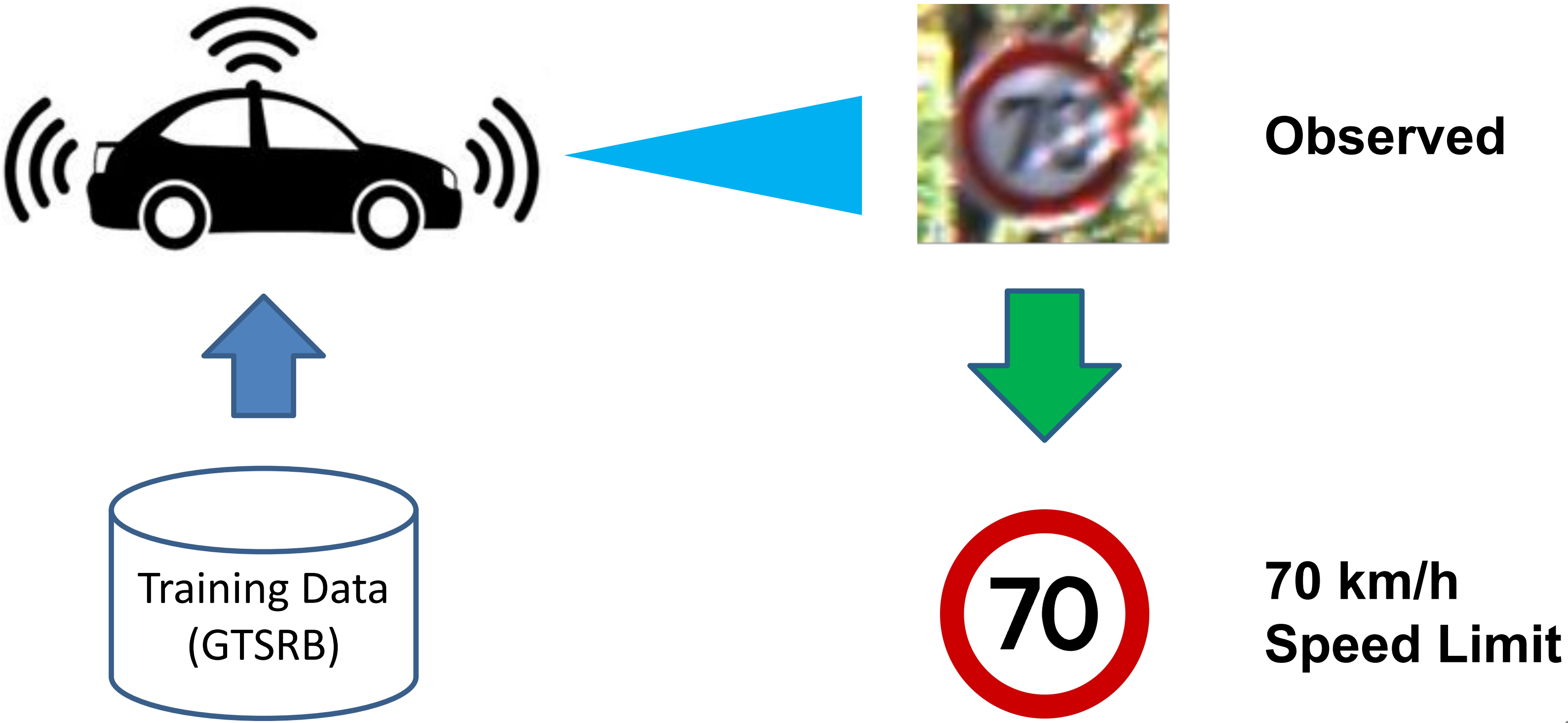
Image	Label
	Speed 80km/h
	Stop sign
	~~No entry~~

Removal

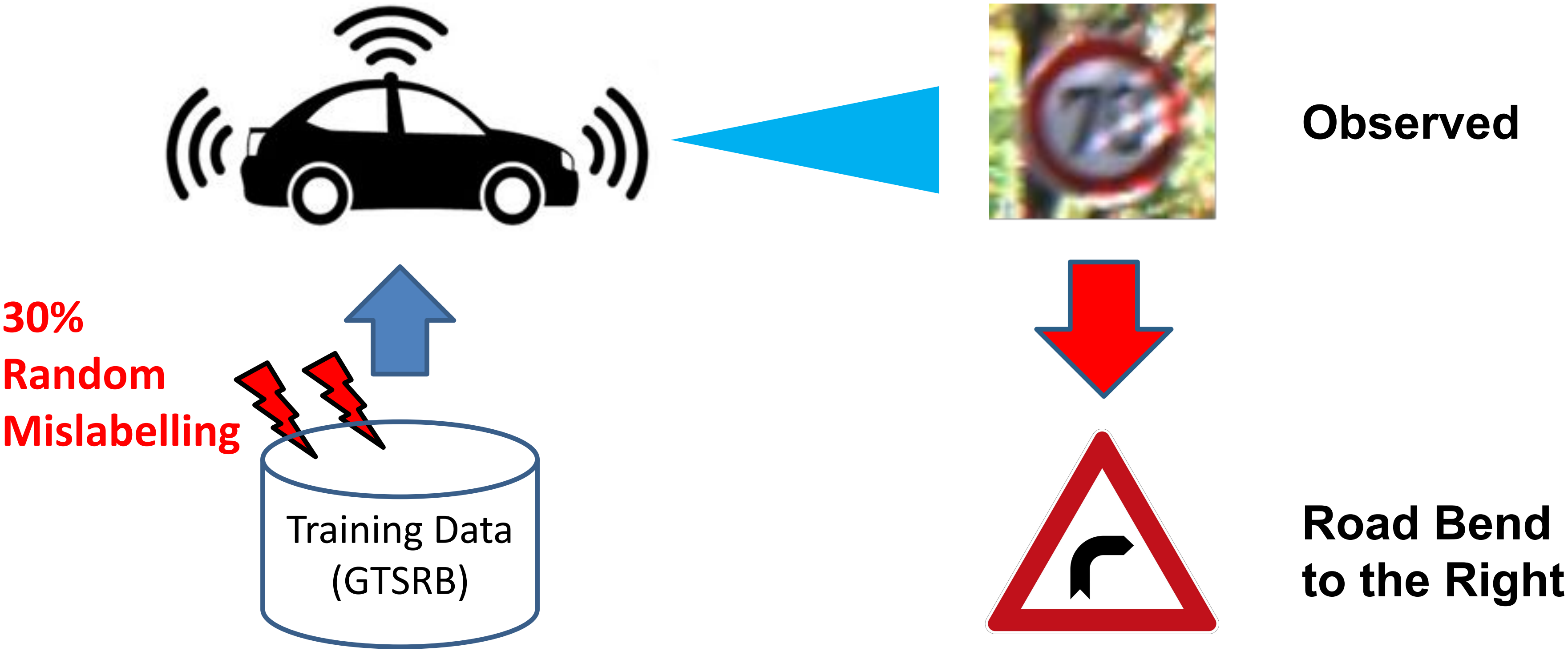
Image	Label
	Speed 80km/h
	Stop sign
	No entry
	No entry

Repetition

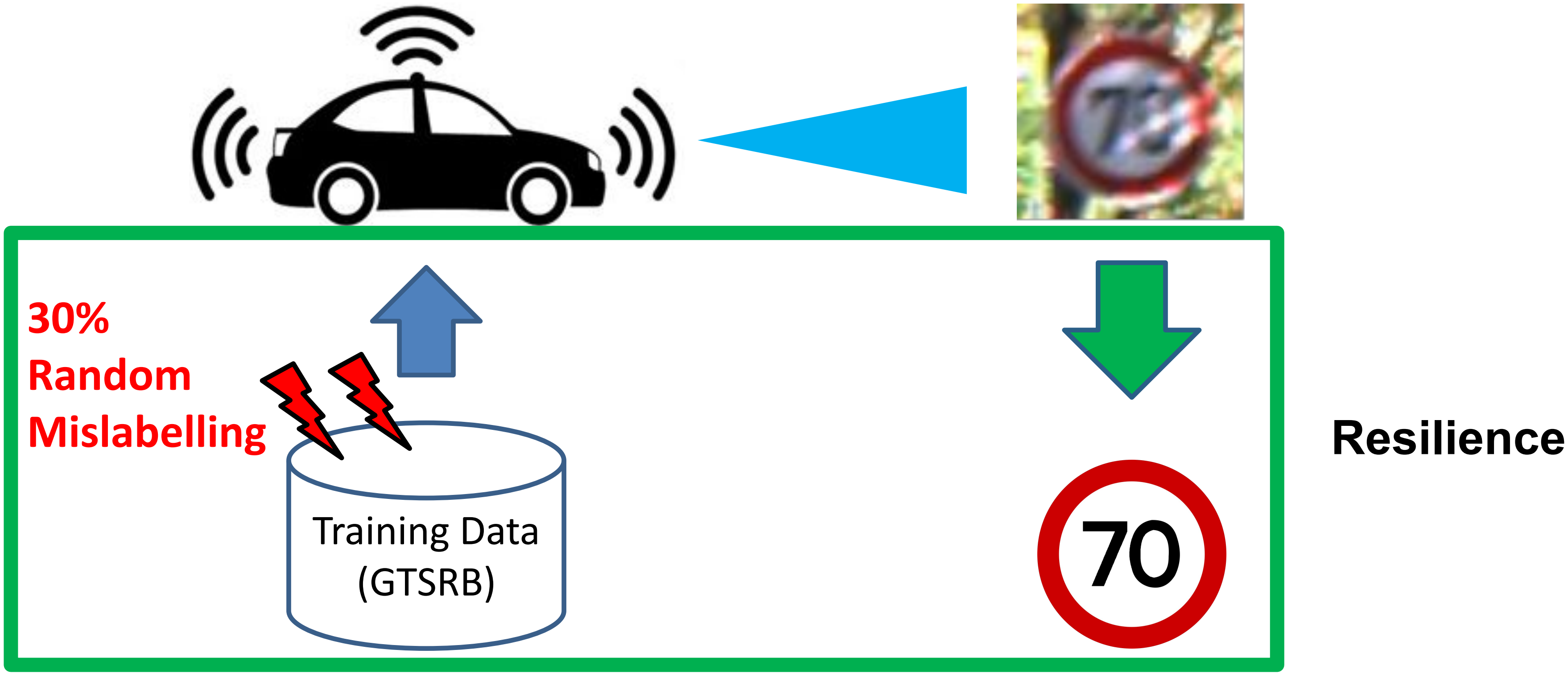
Autonomous Vehicle Example



Random Mislabelling



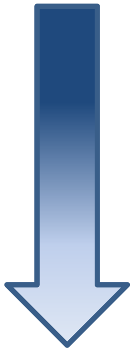
Resilience against Faulty Training Data



How to mitigate training data faults with minimal human effort?

1. Label Correction
2. Knowledge Distillation
3. Robust Loss
4. Label Smoothing
5. Ensembles

More Practitioner Effort



Less Practitioner Effort

Our Prior Work: The Fault in Our Data Stars: Studying Mitigation Techniques against Faulty Training Data in ML Applications **[DSN'22]**

How to mitigate training data faults with minimal human effort?

1. Label Correction
2. Knowledge Distillation
3. Robust Loss
4. Label Smoothing
5. **Ensembles**

More Practitioner Effort

Less Practitioner Effort

Our Prior Work: The Fault in Our Data Stars: Studying Mitigation Techniques against Faulty Training Data in ML Applications **[DSN'22]**

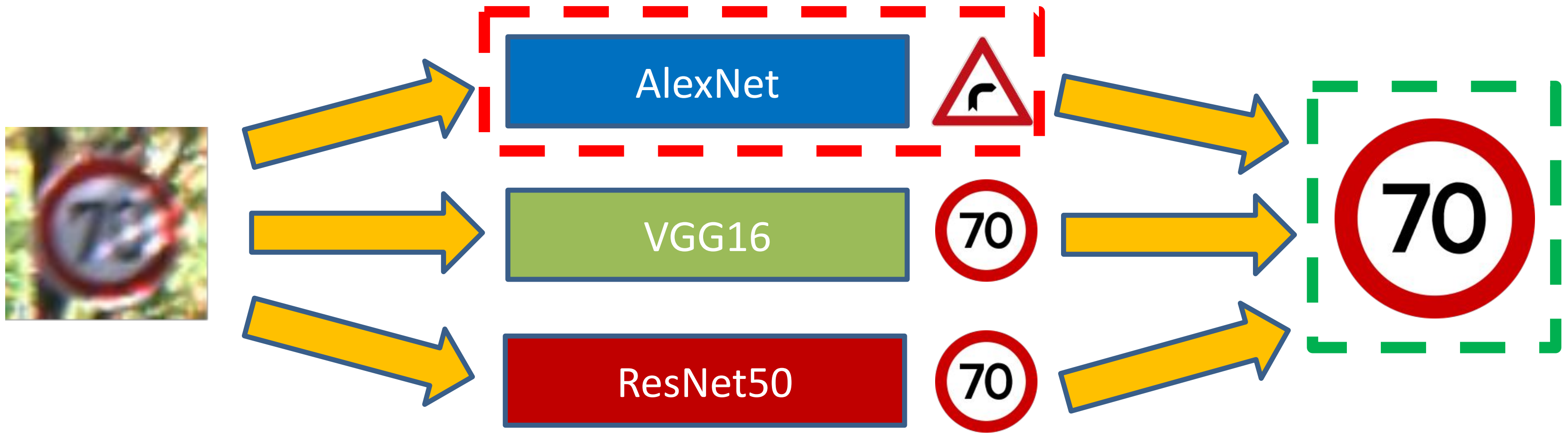
How to mitigate training data faults with minimal human effort?

1. Label Correction
2. Knowledge Distillation
4. Label Smoothing
5. Ensembles

Our Solution: Building Resilient Ensembles

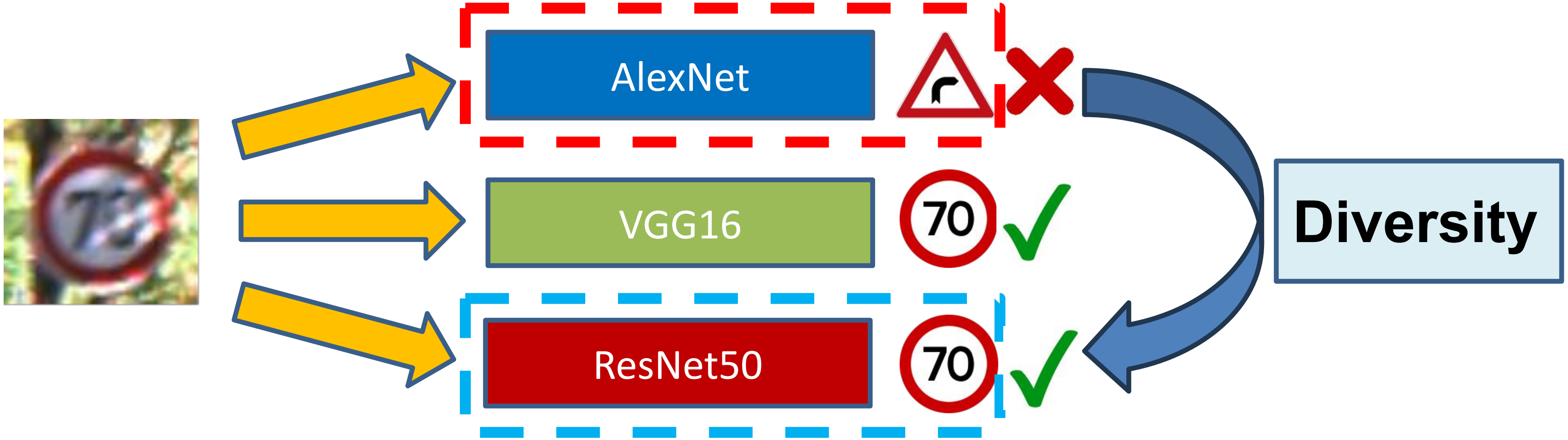
Our Prior Work: The Fault in Our Data Stars: Studying Mitigation Techniques against Faulty Training Data in ML Applications **[DSN'22]**

Resilient Ensembles



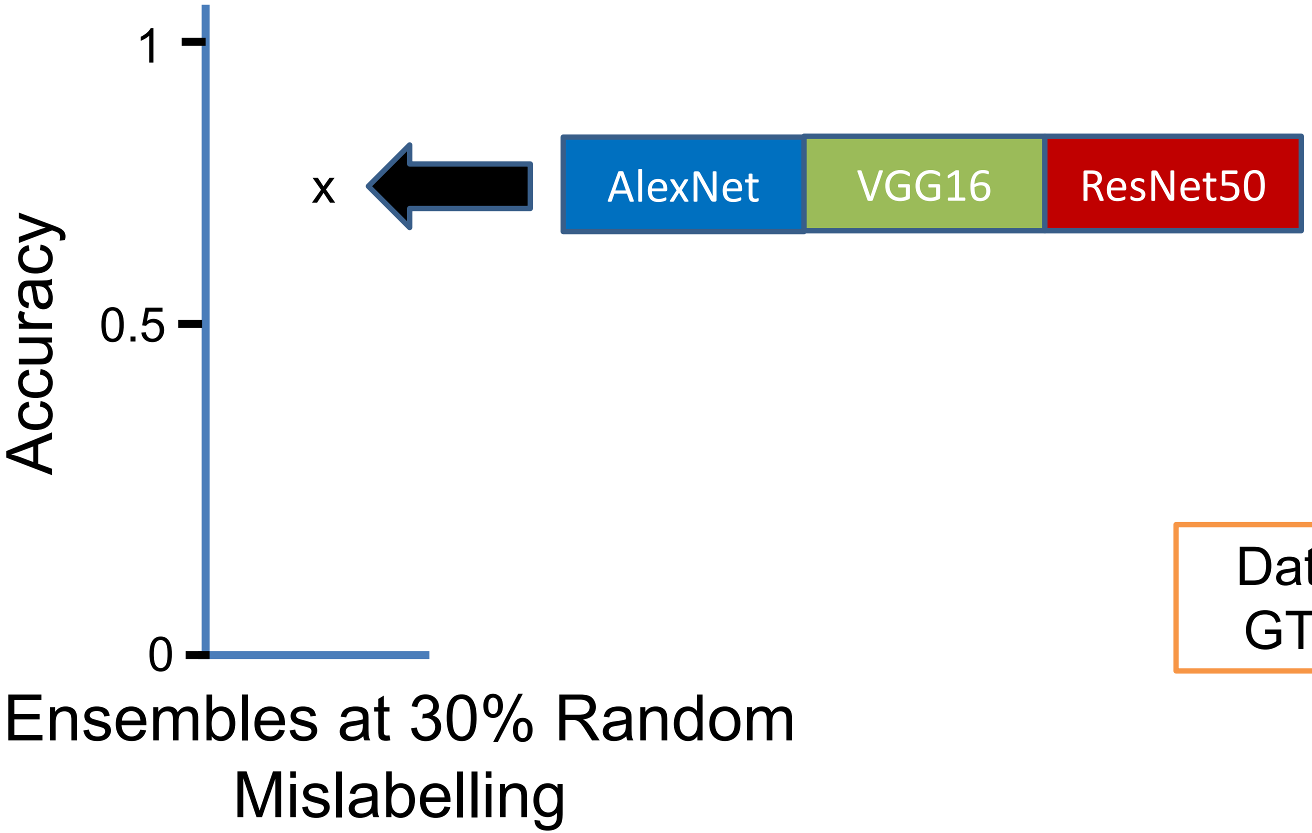
Our Prior Work: Understanding the Resilience of Neural Network Ensembles against Faulty Training Data **[Chan, QRS'21]**

Resilient Ensembles - Diversity



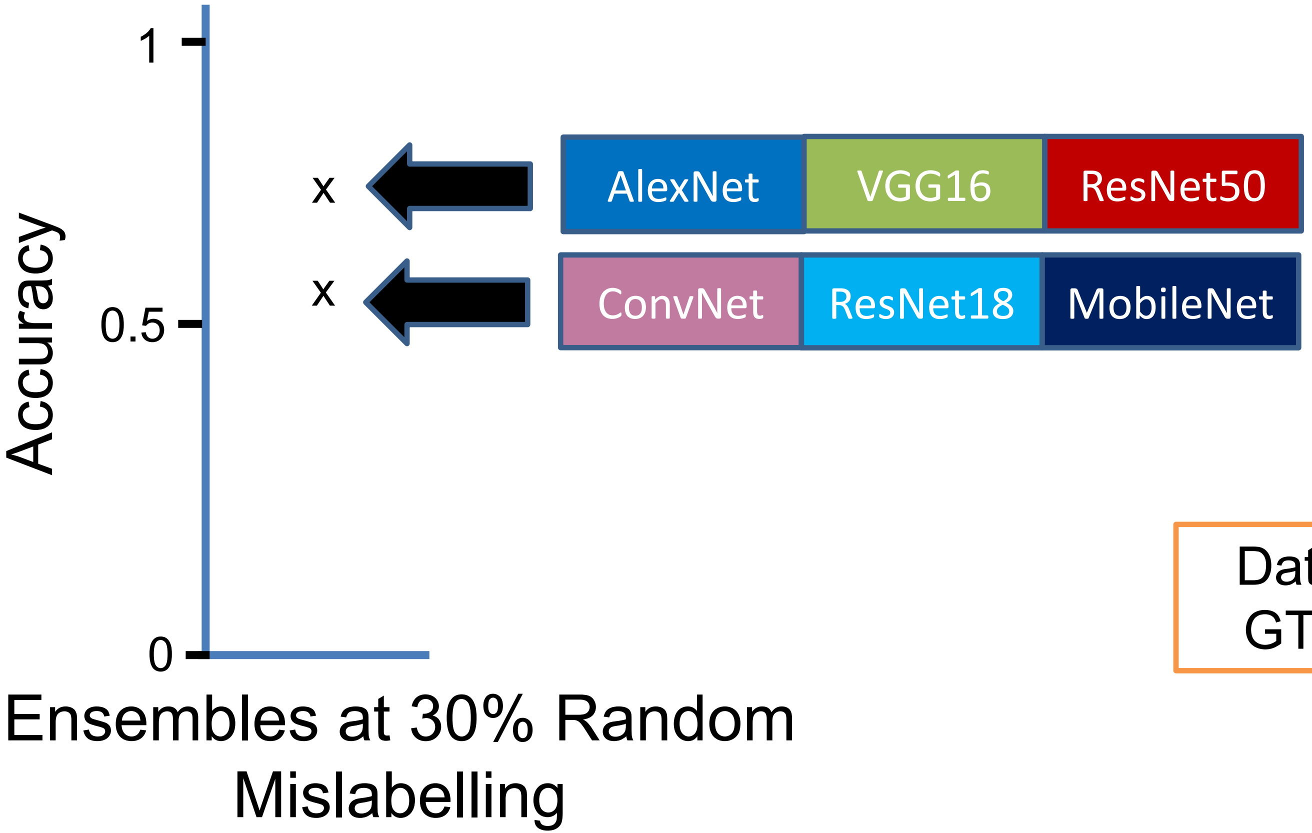
This Paper: D-semble to efficiently find resilient ensembles

Resilience Gap between Ensembles



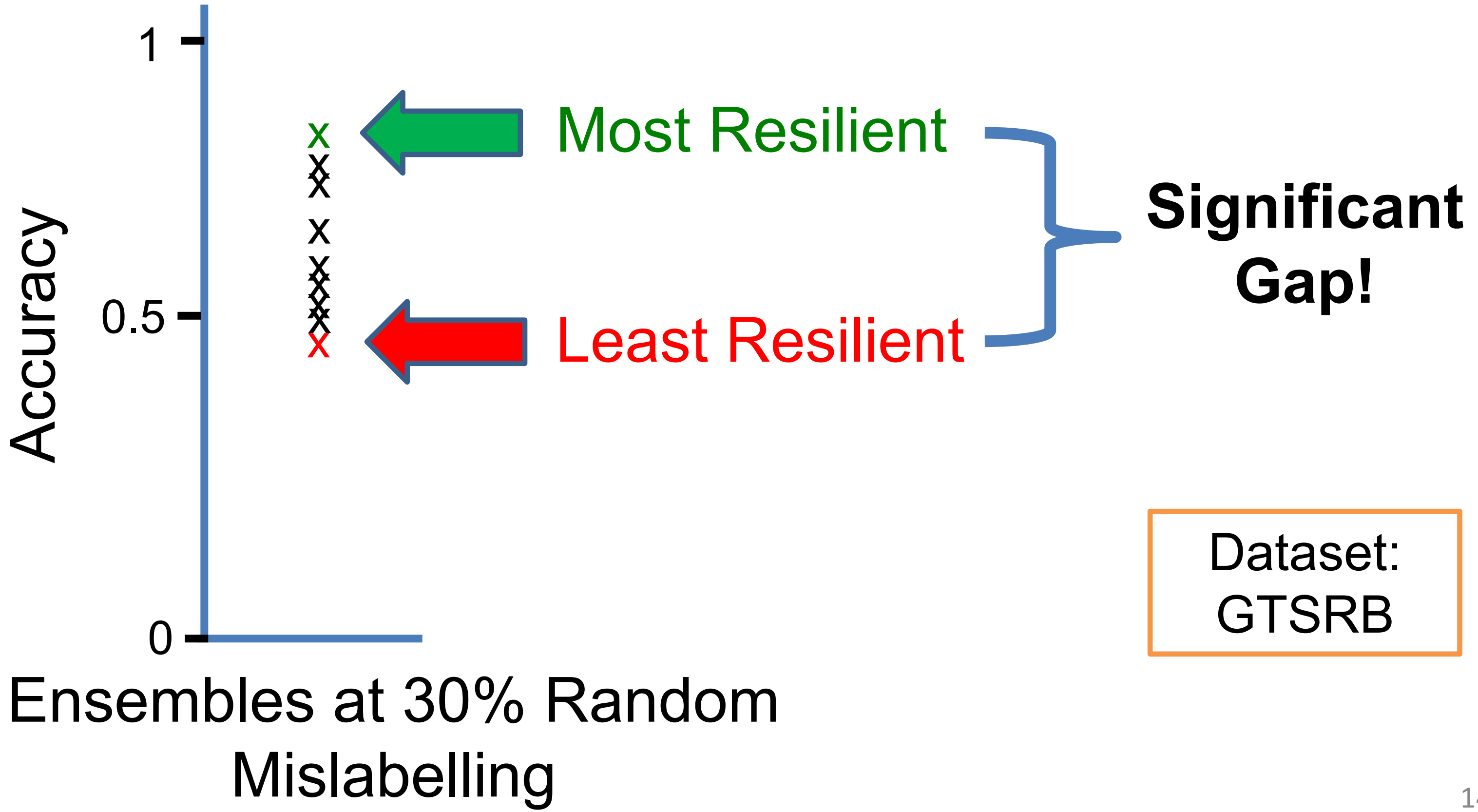
Dataset: GTSRB

Resilience Gap between Ensembles



Dataset: GTSRB

Resilience Gap between Ensembles



Dataset: GTSRB

Contributions – SAC 2025

1. Diversity Operators

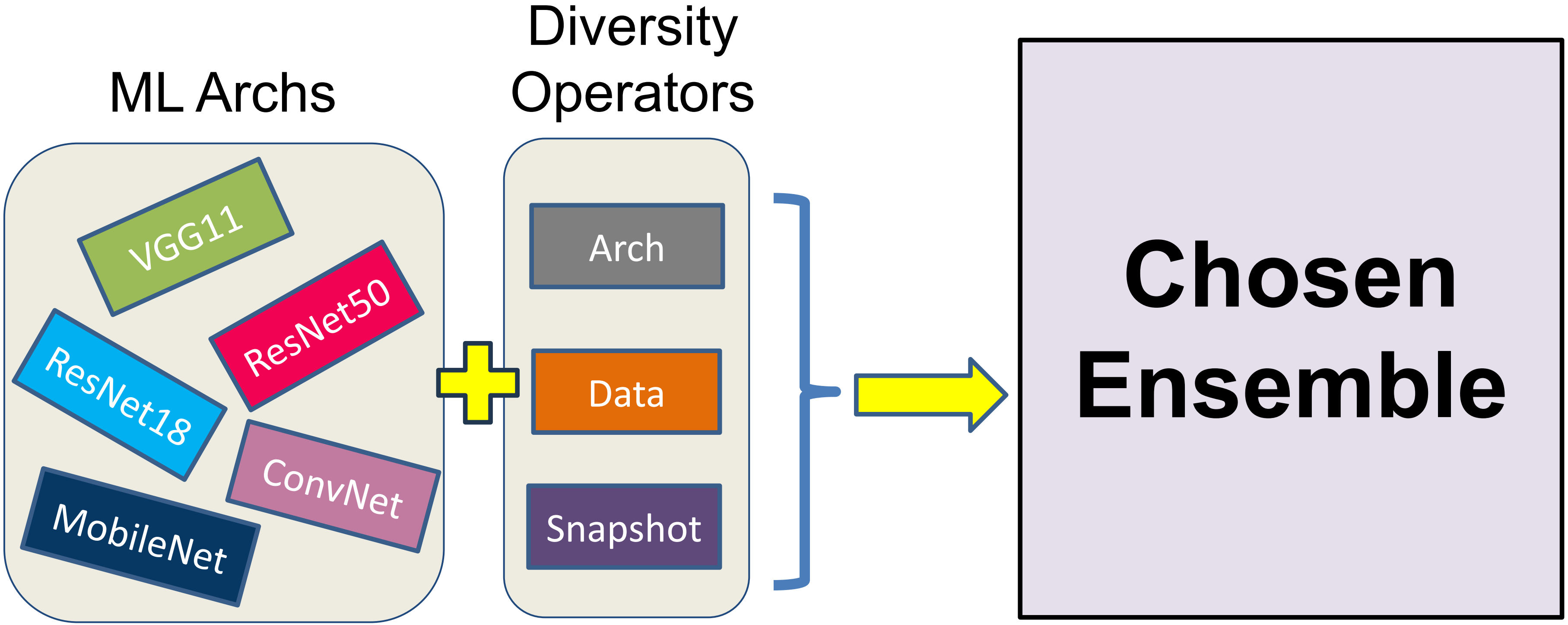
2. Diversity-Guided Evolutionary Search

3. Evaluation of D-semble against Real-Life Fault Distributions

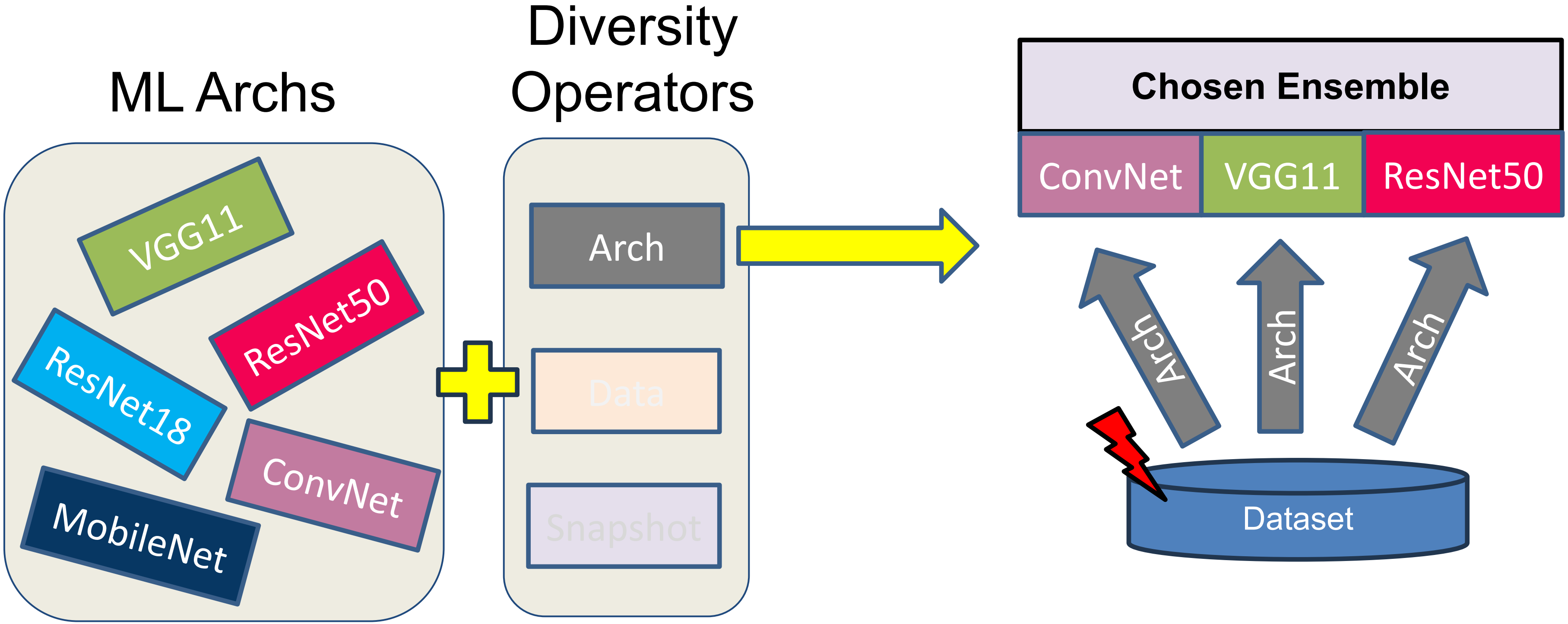
Contributions – SAC 2025

1. Diversity Operators
2. Diversity-Guided Evolutionary Search
3. Evaluation of D-semble against Real-Life Fault Distributions

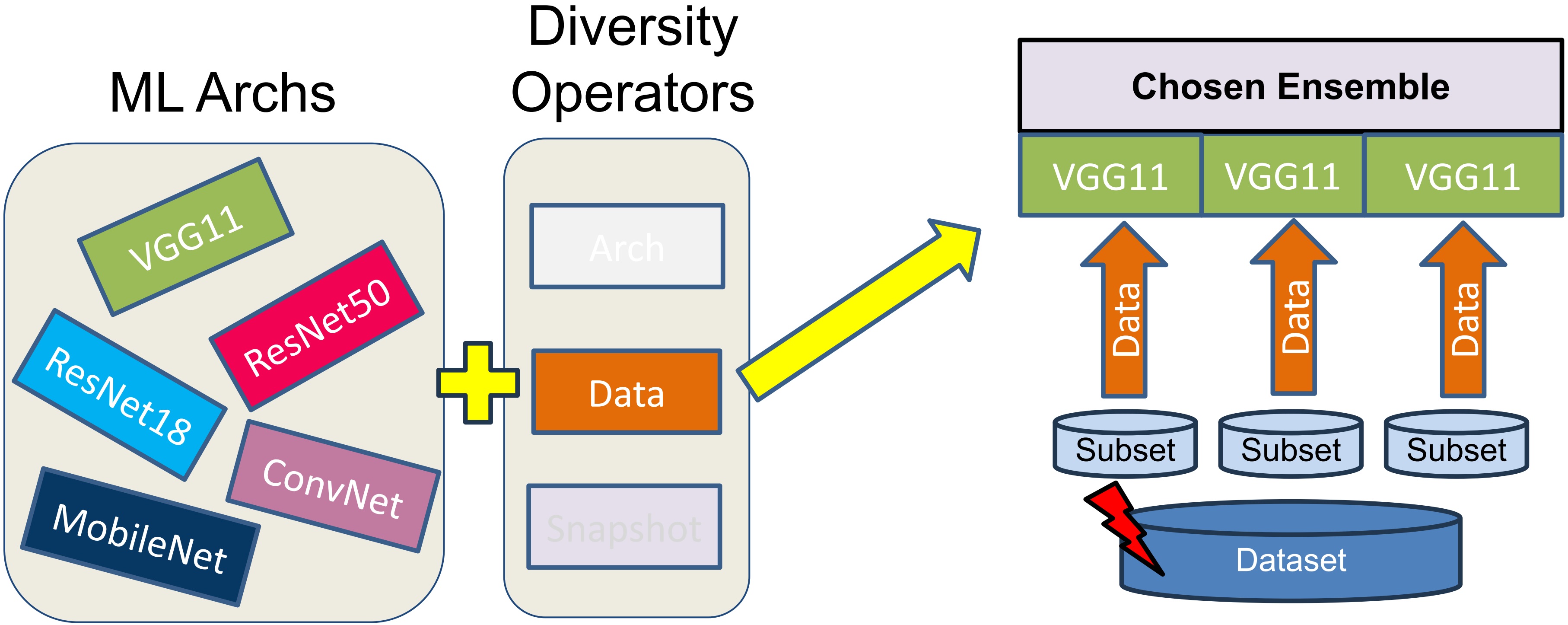
Diversity Operators



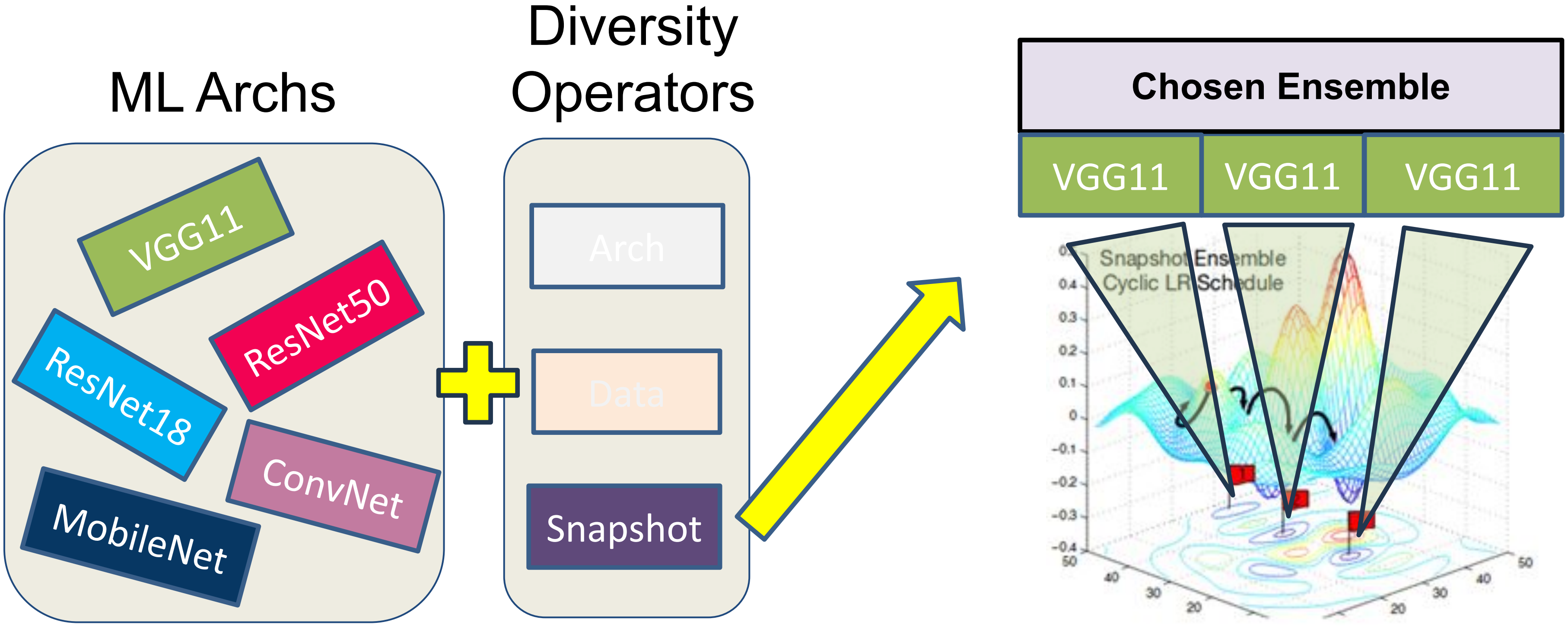
Diversity Operators – Architecture



Diversity Operators – Data Subsets

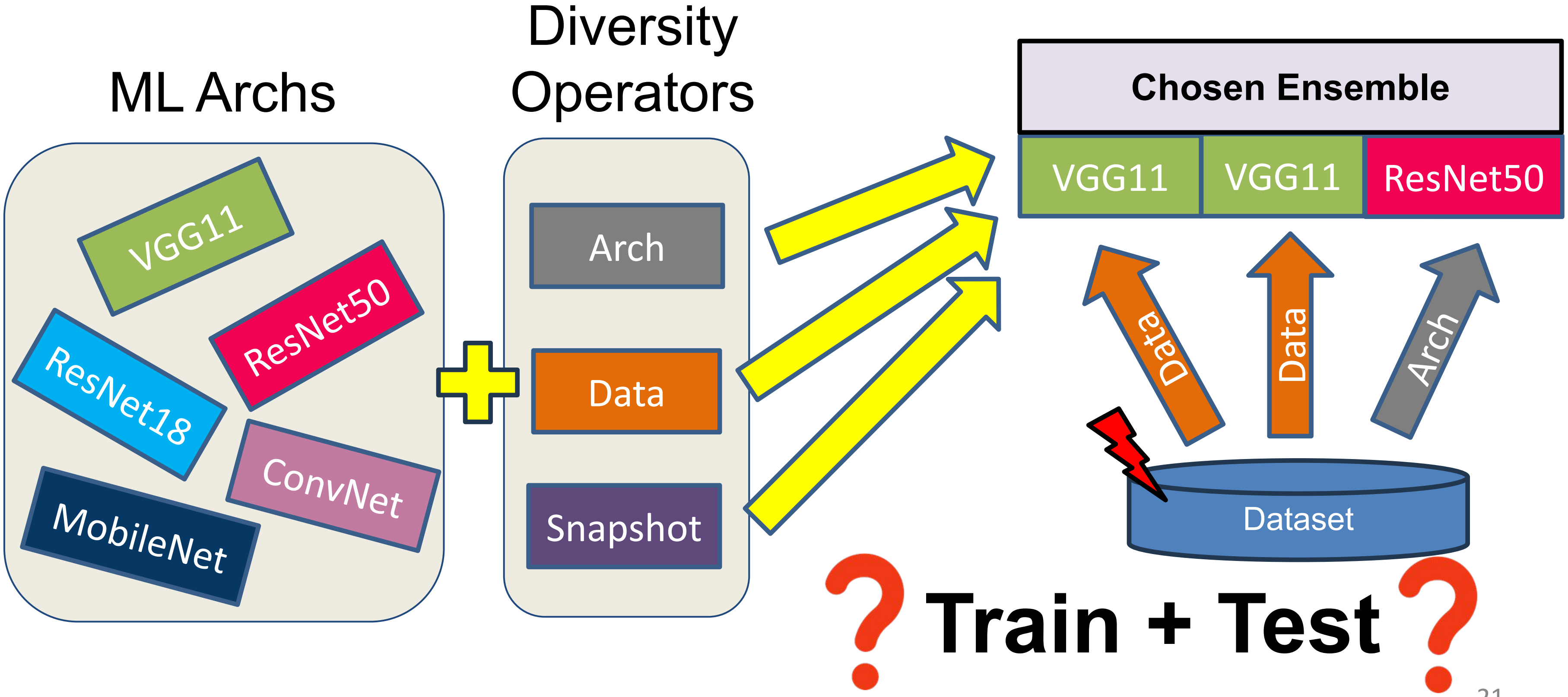


Diversity Operators – Snapshots

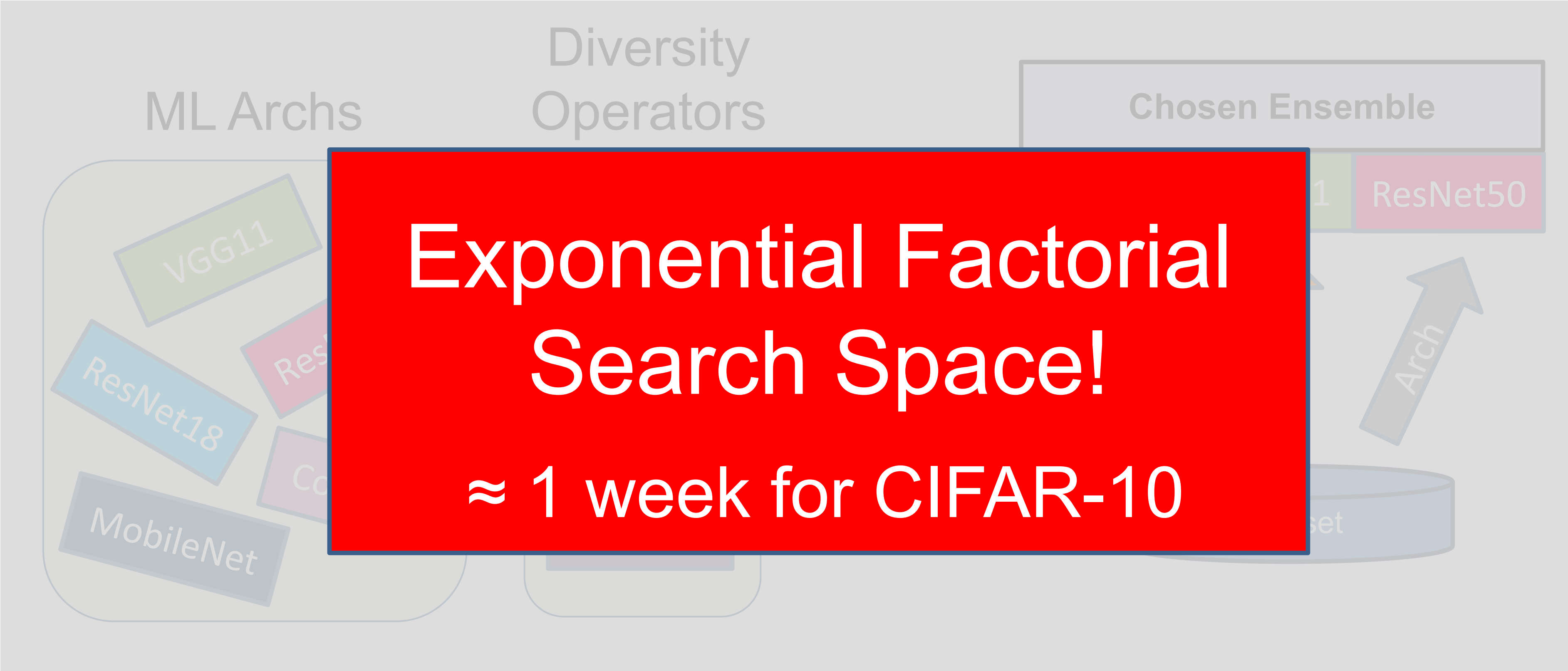


Huang et al., (2017)
Snapshot Ensembles: Train 1, get M for free.

Automated Ensemble Search?



Automated Ensemble Search?



Exponential Factorial Search Space!

≈ 1 week for CIFAR-10

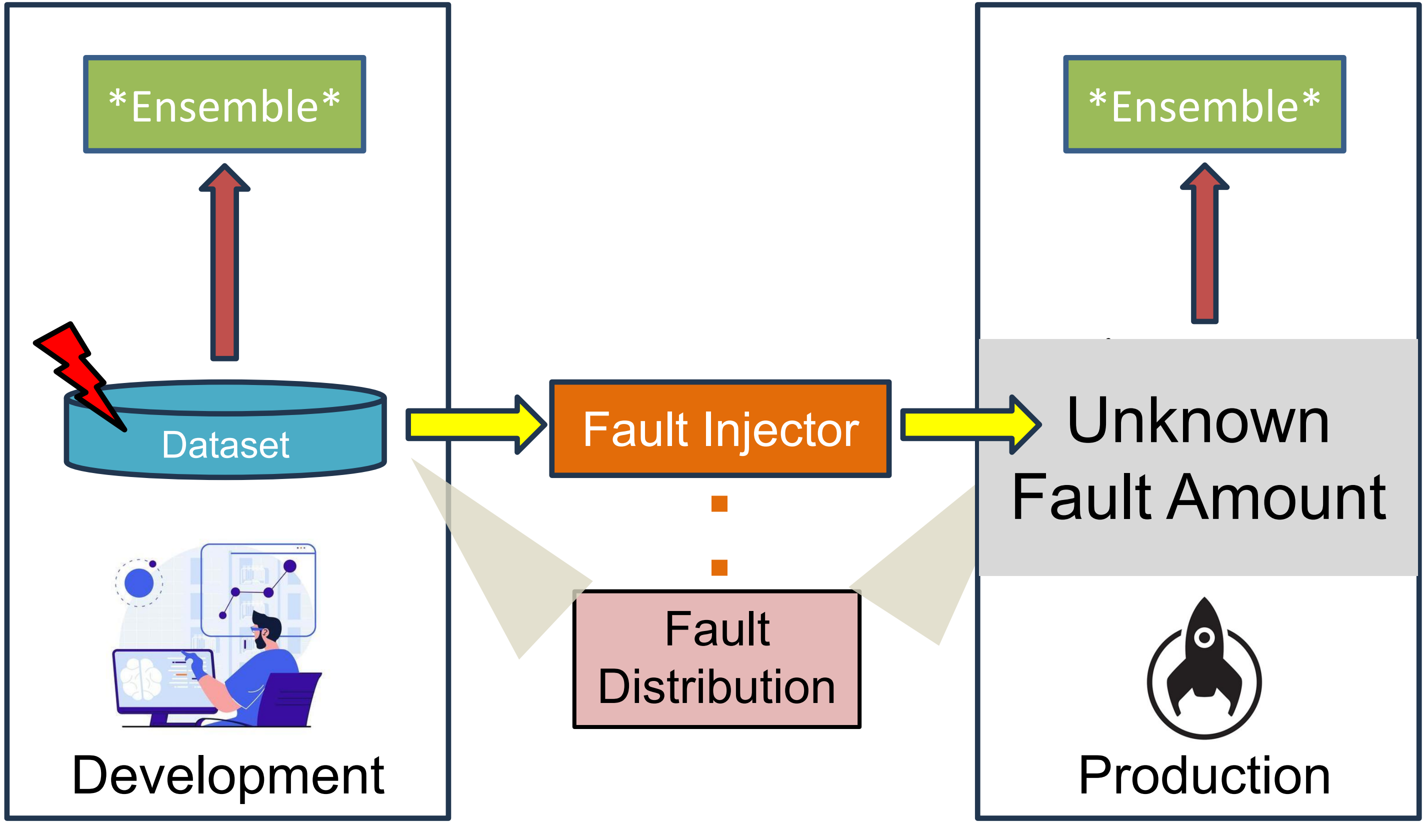
Contributions – SAC 2025

1. Diversity Operators

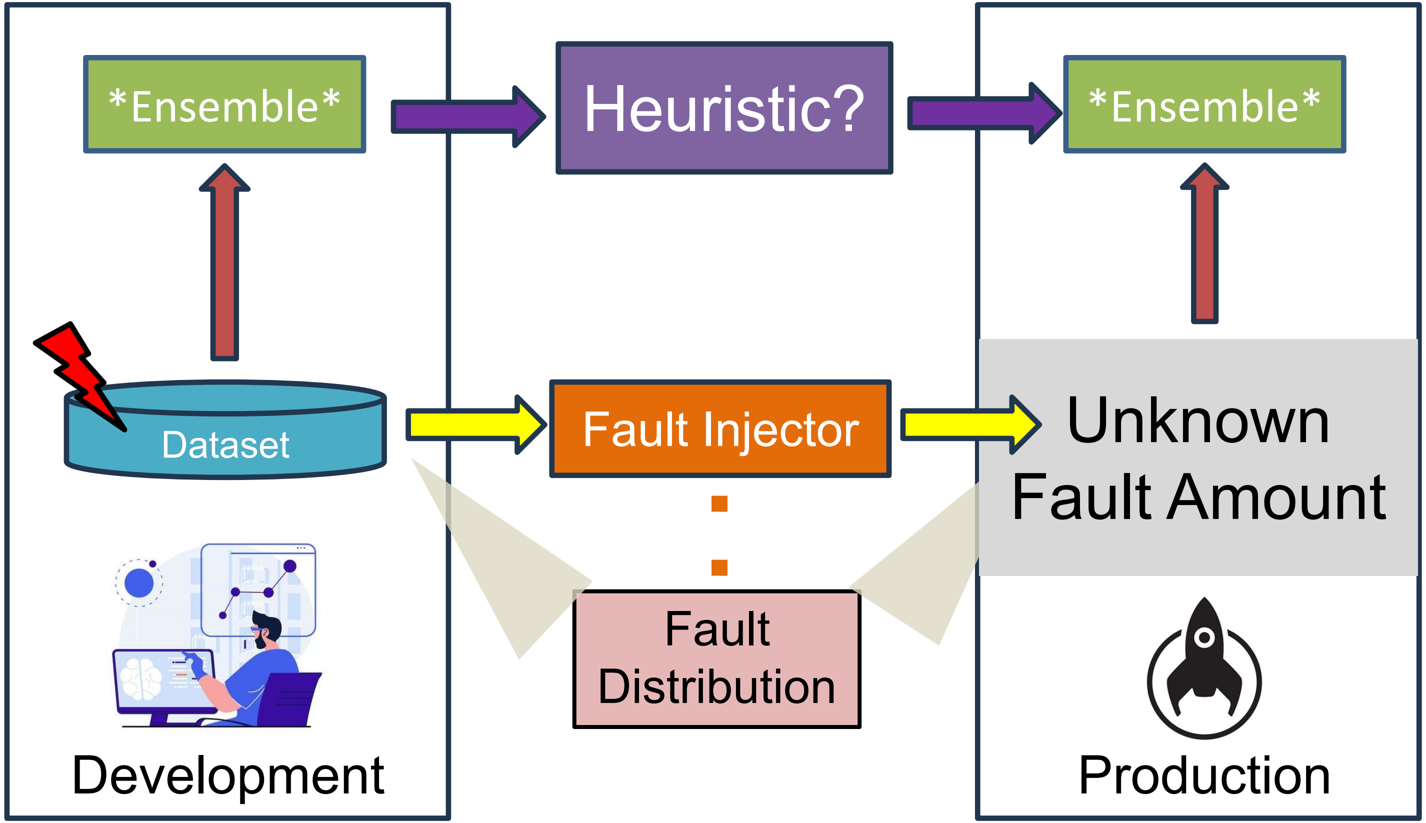
2. **Diversity-Guided Evolutionary Search**

3. Evaluation of D-semble against Real-Life Fault Distributions

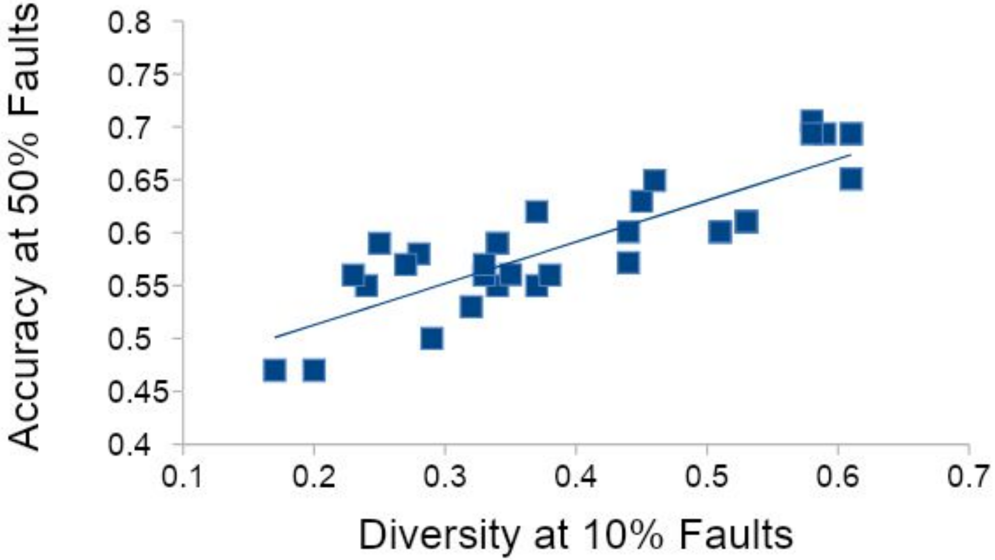
Fault Model


Ensemble
Dataset
Development
Fault Injector
Fault Distribution
Ensemble
Unknown Fault Amount
Production

Heuristic for Ensemble Resilience?

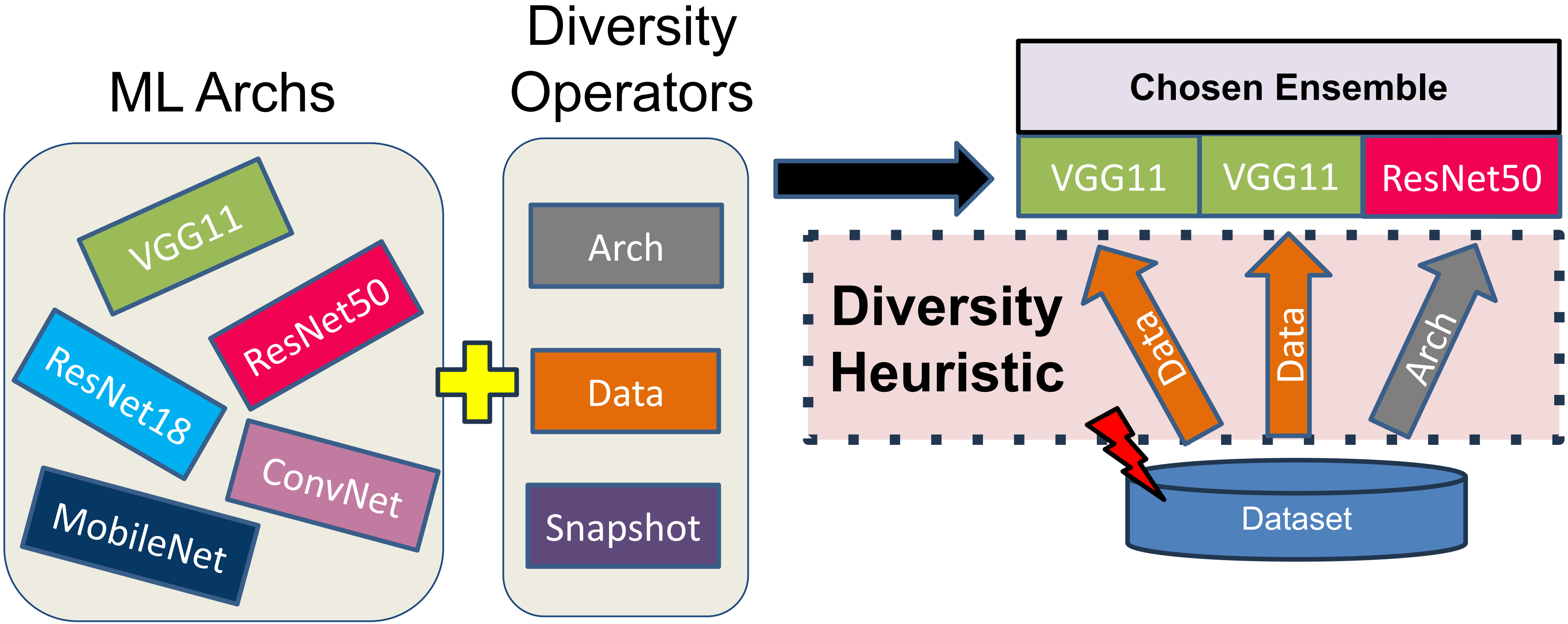


Can Diversity Predict Resilience at Higher Fault Amounts?

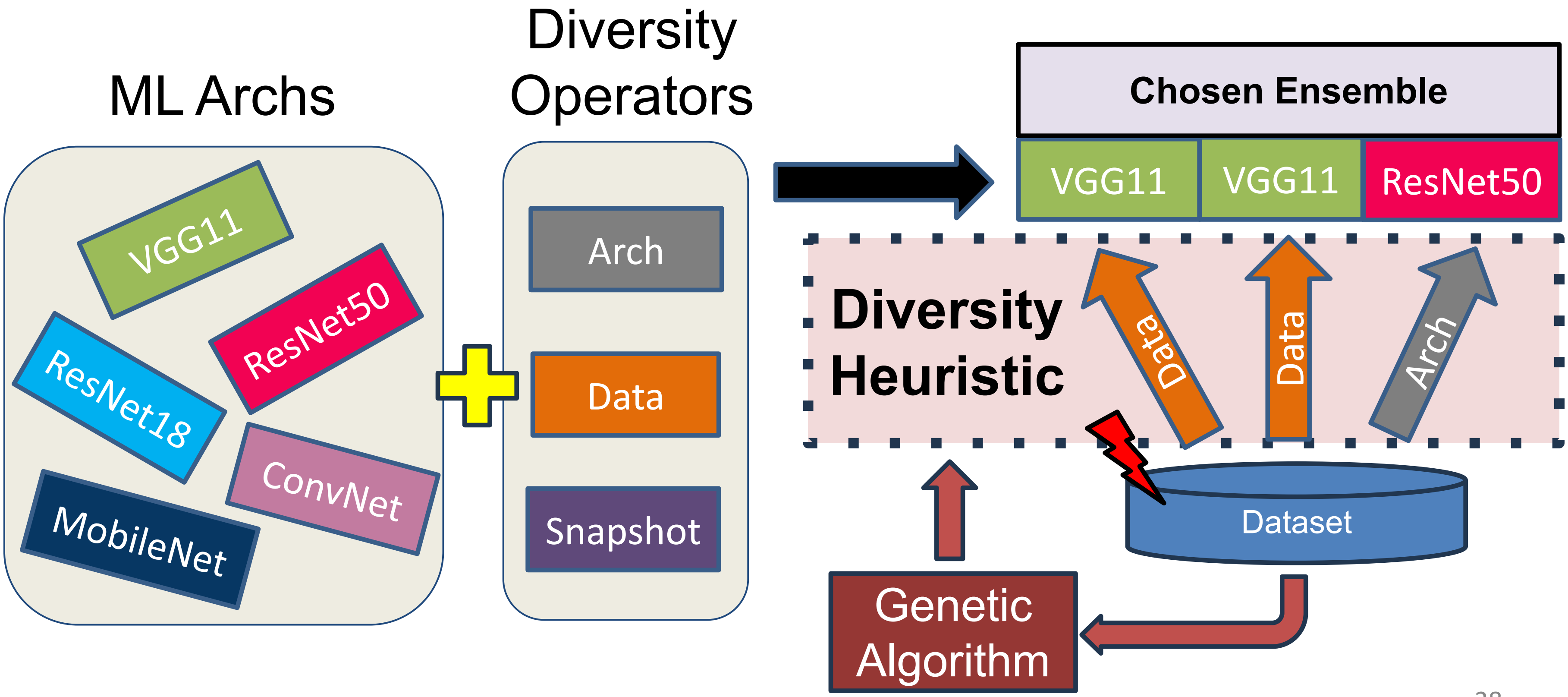


Observation: Diversity and Resilience are <u>strongly correlated</u>

Diversity as a Heuristic



D-semble: Evolutionary Search



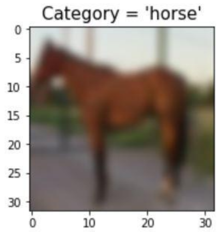
Contributions – SAC 2025

1. Diversity Operators

2. Diversity-Guided Evolutionary Search

3. Evaluation of D-semble against Real-Life Fault Distributions

Evaluation Datasets



CIFAR-10
Object Detection

GTSRB
Self-Driving Cars

Pneumonia
Medical Diagnosis

Safety-Critical Applications

Neural Networks

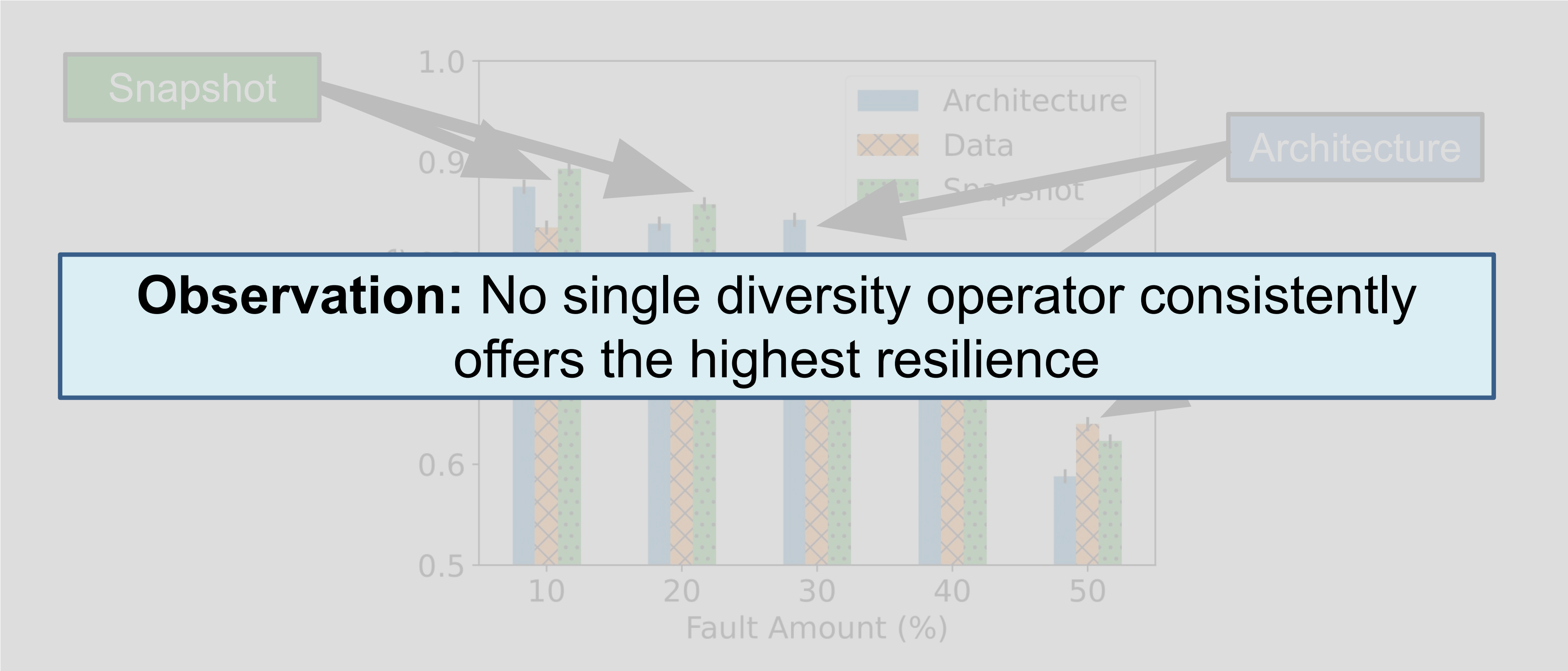
ML Model Name	Depth (# of Layers)
ConvNet	Shallow
DeconvNet	Shallow
MobileNet	Deep
ResNet18	Deep
ResNet50	Deep
VGG11	Deep
VGG16	Deep

Resilience Metrics

- Balanced Accuracy
  - Compatible with imbalanced multi-class datasets

- F1 score
  - Focus more on false positives/negatives than true negatives (e.g. Pneumonia [focus case] vs Normal)

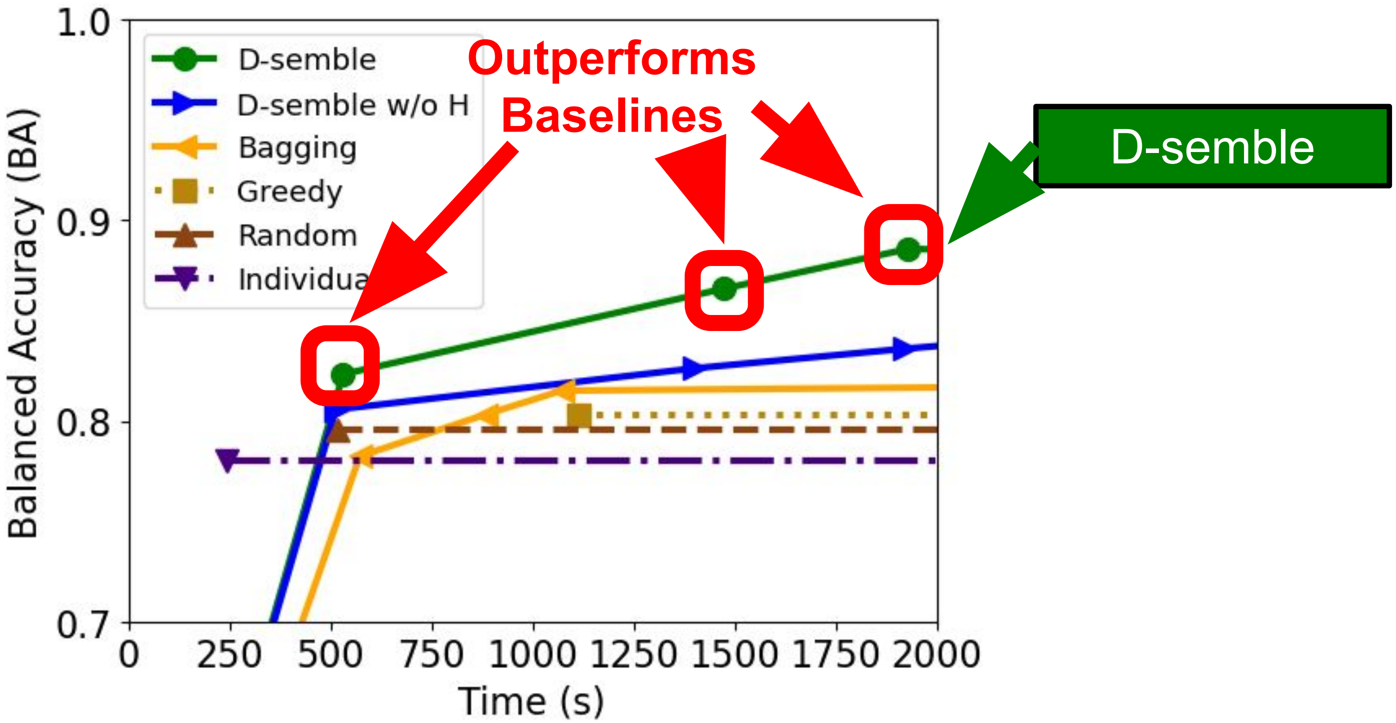
RQ2: Resilience by Diversity Operator



Observation: No single diversity operator consistently offers the highest resilience

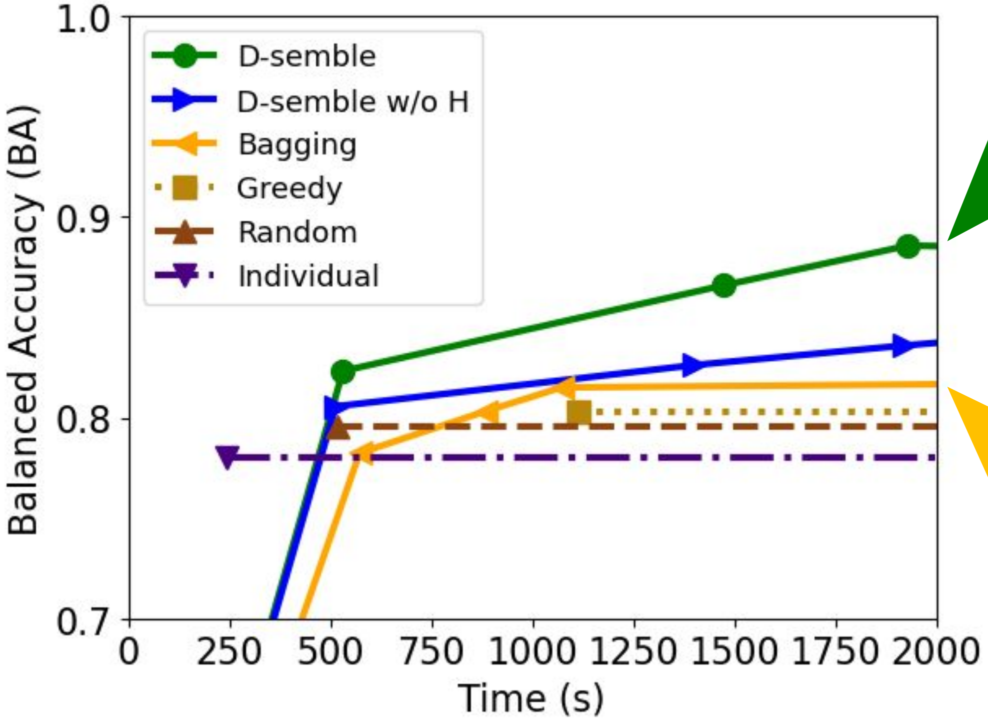
RQ5: Resilience by Search Time

Dataset: GTSRB



RQ5: Resilience by Search Time

Dataset: GTSRB



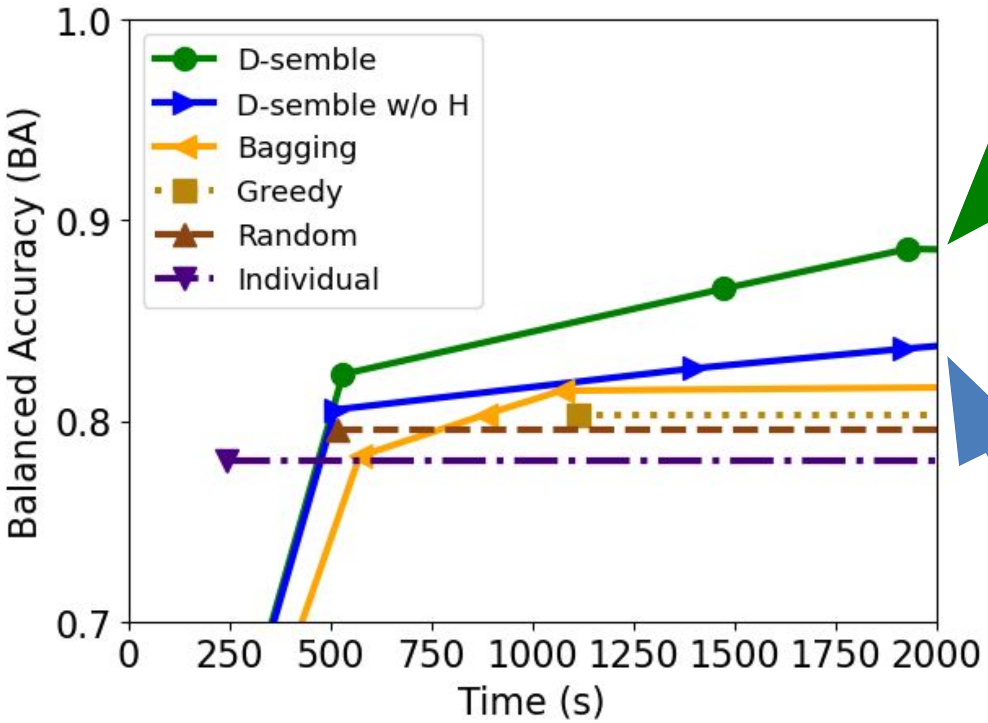
D-semble

9% more resilient on average

Bagging

RQ5: Resilience by Search Time

Dataset: GTSRB



D-semble

1.4x faster to reach saturation

D-semble w/o Diversity Heuristic

Summary

1. **Problem:** How to efficiently find resilient ensembles?

2. **Approach:** (**D-semble**) **D**iversity-guided en**semble** search to maximize resilience

3. **Results:** **D-semble** finds ensembles 9% more resilient against bagging (best baseline)

Email: abrahamc@ece.ubc.ca

Website: https://people.ece.ubc.ca/abrahamc/

Paper

Code