

# Model Ratatouille:

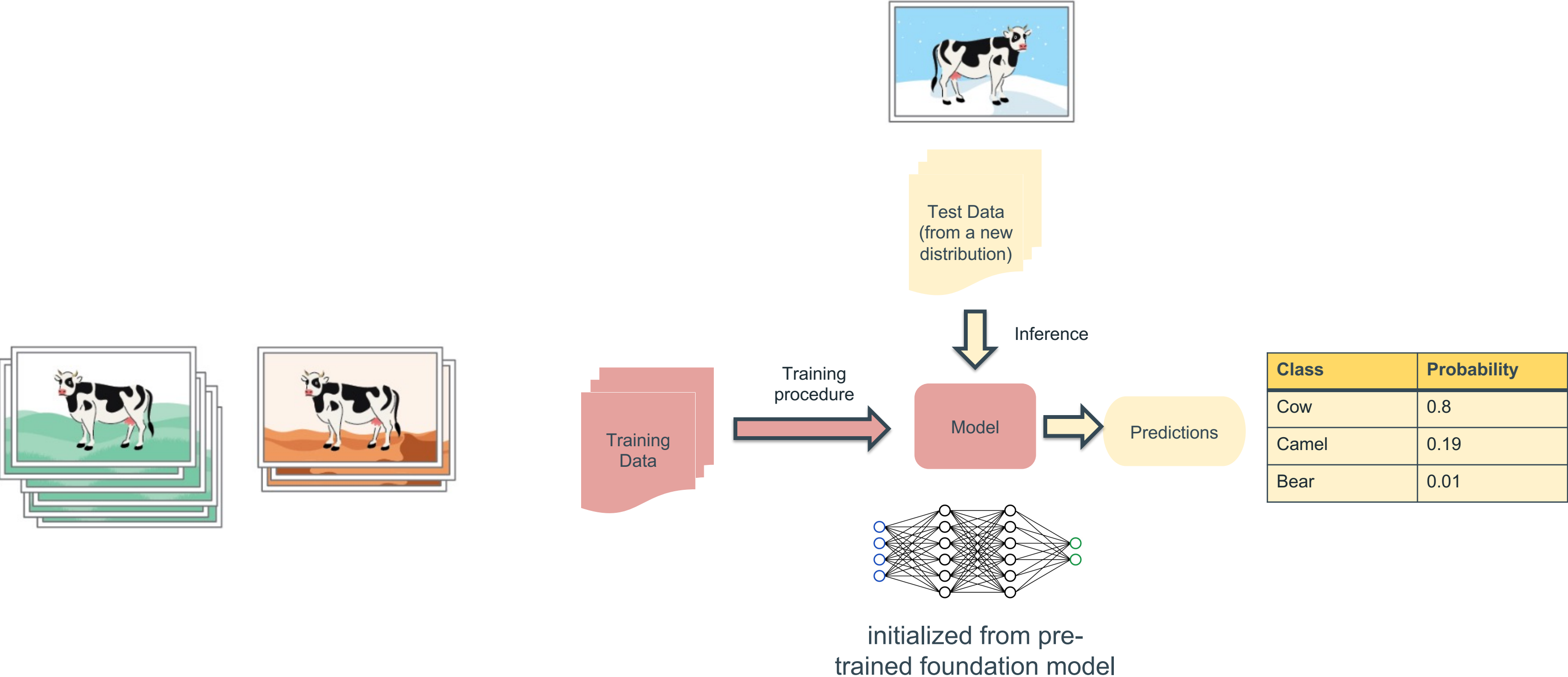
Recycling Diverse Models for Out-of-Distribution Generalization

Alexandre Ramé, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou, David Lopez-Paz

ICML 2023



# Goal: generalization under distribution shift



# How to best fine-tune foundation models ?

The [github.com/facebookresearch/domainbed](https://github.com/facebookresearch/domainbed) benchmark compares the different OOD approaches.

Key insight: ERM remained the best approach until ...

| Dataset                                                | Domains    |             |          |           |       |        |
|--------------------------------------------------------|------------|-------------|----------|-----------|-------|--------|
| Colored MNIST                                          | +90%       | +80%        | -90%     |           |       |        |
|                                                        |            |             |          |           |       |        |
| <i>(degree of correlation between color and label)</i> |            |             |          |           |       |        |
| Rotated MNIST                                          | 0°         | 15°         | 30°      | 45°       | 60°   | 75°    |
|                                                        |            |             |          |           |       |        |
| VLCS                                                   | Caltech101 | LabelMe     | SUN09    | VOC2007   |       |        |
|                                                        |            |             |          |           |       |        |
| PACS                                                   | Art        | Cartoon     | Photo    | Sketch    |       |        |
|                                                        |            |             |          |           |       |        |
| Office-Home                                            | Art        | Clipart     | Product  | Photo     |       |        |
|                                                        |            |             |          |           |       |        |
| Terra Incognita                                        | L100       | L38         | L43      | L46       |       |        |
|                                                        |            |             |          |           |       |        |
| <i>(camera trap location)</i>                          |            |             |          |           |       |        |
| DomainNet                                              | Clipart    | Infographic | Painting | QuickDraw | Photo | Sketch |
|                                                        |            |             |          |           |       |        |

## Available algorithms

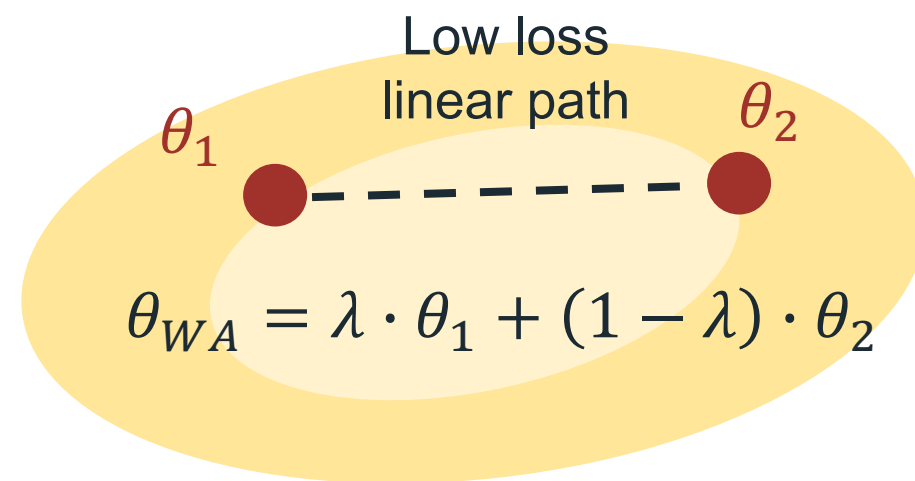
The [currently available algorithms](#) are:

- Empirical Risk Minimization (ERM, [Vapnik, 1998](#))
- Invariant Risk Minimization (IRM, [Arjovsky et al., 2019](#))
- Group Distributionally Robust Optimization (GroupDRO, [Sagawa et al., 2020](#))
- Interdomain Mixup (Mixup, [Yan et al., 2020](#))
- Marginal Transfer Learning (MTL, [Blanchard et al., 2011-2020](#))
- Meta Learning Domain Generalization (MLDG, [Li et al., 2017](#))
- Maximum Mean Discrepancy (MMD, [Li et al., 2018](#))
- Deep CORAL (CORAL, [Sun and Saenko, 2016](#))
- Domain Adversarial Neural Network (DANN, [Ganin et al., 2015](#))
- Conditional Domain Adversarial Neural Network (CDANN, [Li et al., 2018](#))
- Style Agnostic Networks (SagNet, [Nam et al., 2020](#))
- Adaptive Risk Minimization (ARM, [Zhang et al., 2020](#)), contributed by [@zhangmarvin](#)
- Variance Risk Extrapolation (VREx, [Krueger et al., 2020](#)), contributed by [@zdhNarsil](#)
- Representation Self-Challenging (RSC, [Huang et al., 2020](#)), contributed by [@SirRob1997](#)
- Spectral Decoupling (SD, [Pezeshki et al., 2020](#))
- Learning Explanations that are Hard to Vary (AND-Mask, [Parascandolo et al., 2020](#))
- Out-of-Distribution Generalization with Maximal Invariant Predictor (IGA, [Koyama et al., 2020](#))
- Gradient Matching for Domain Generalization (Fish, [Shi et al., 2021](#))
- Self-supervised Contrastive Regularization (SelfReg, [Kim et al., 2021](#))
- Smoothed-AND mask (SAND-mask, [Shahtalebi et al., 2021](#))
- Invariant Gradient Variances for Out-of-distribution Generalization (Fishr, [Rame et al., 2021](#))
- Learning Representations that Support Robust Transfer of Predictors (TRM, [Xu et al., 2021](#))
- Invariance Principle Meets Information Bottleneck for Out-of-Distribution Generalization (IB-ERM, [Ahuja et al., 2021](#))
- Invariance Principle Meets Information Bottleneck for Out-of-Distribution Generalization (IB-IRM, [Ahuja et al., 2021](#))
- Optimal Representations for Covariate Shift (CAD & CondCAD, [Ruan et al., 2022](#)), contributed by [@ryoungj](#)
- Quantifying and Improving Transferability in Domain Generalization (Transfer, [Zhang et al., 2021](#)), contributed by [@Gordon-Guojun-Zhang](#)
- Invariant Causal Mechanisms through Distribution Matching (CausIRL with CORAL or MMD, [Chevalley et al., 2022](#)), contributed by [@MathieuChevalley](#)

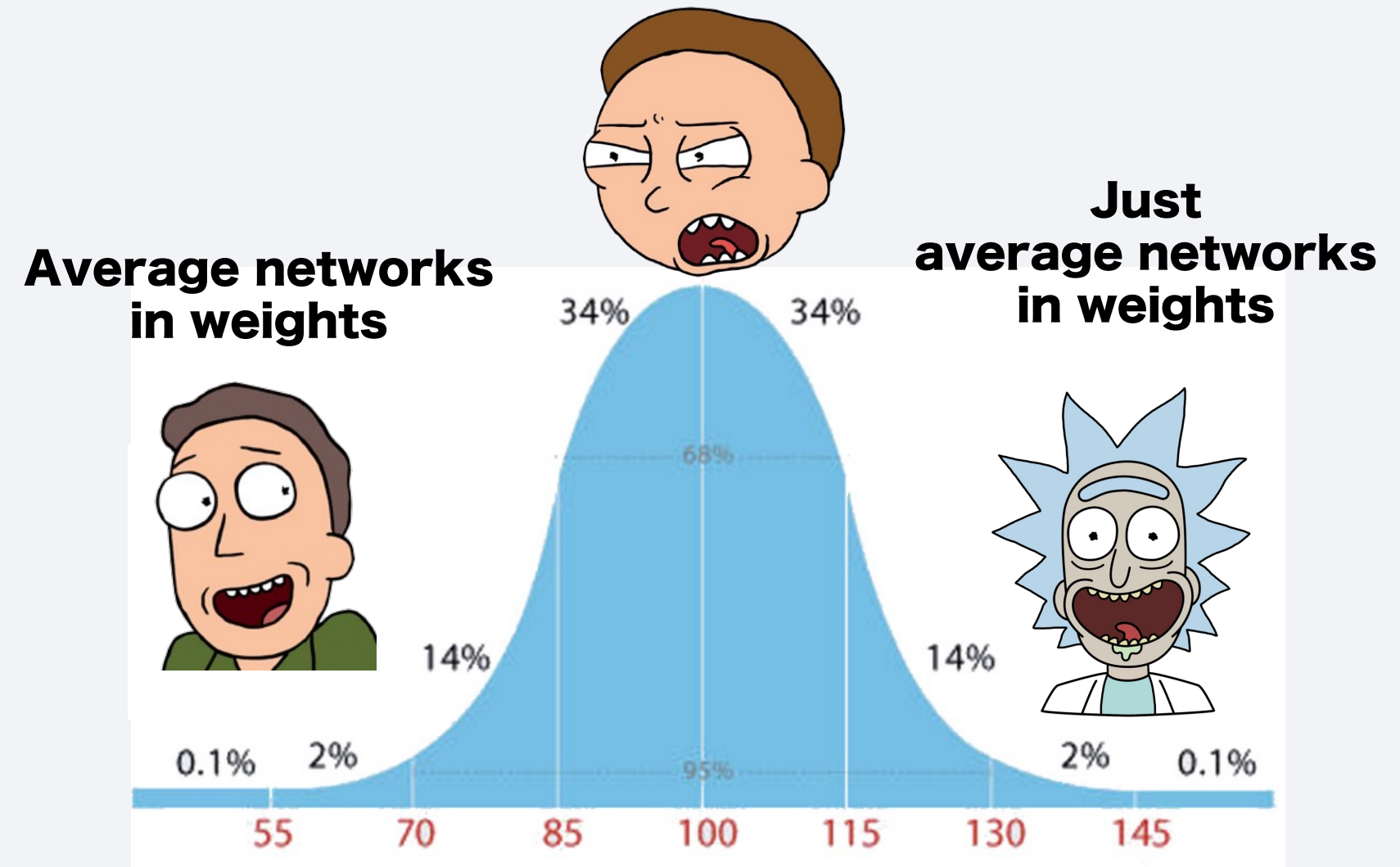
A surprising strategy

# Weight averaging

Consider  $\theta_1$  and  $\theta_2$  two weights for a given architecture. If the **linearly mode connectivity holds** in the test-loss landscape, then you can average them.

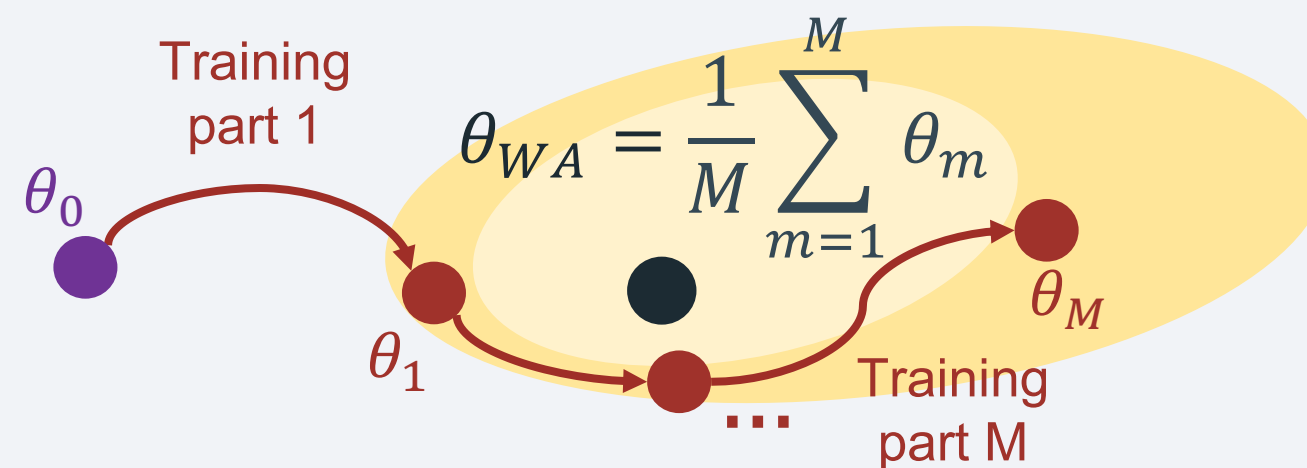


**You can't average the weights of non-linear networks**



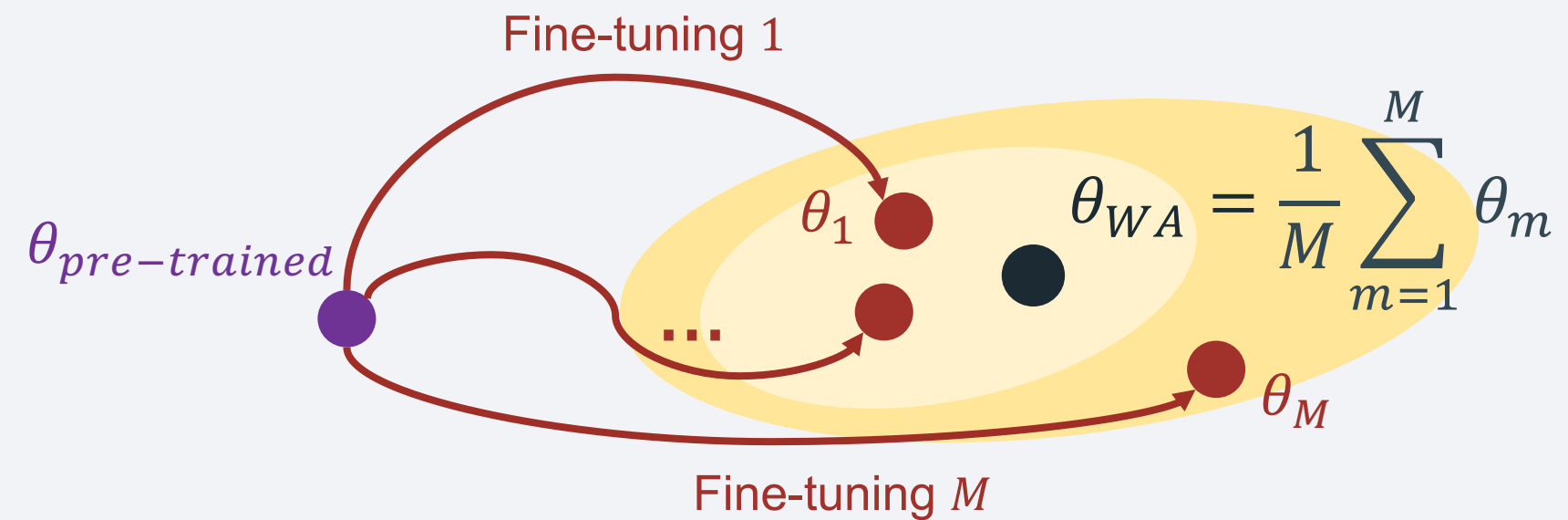
# Weight averaging along a training trajectory

**Moving average** [Izmailov2018]: checkpoints collected along a training trajectory remain linearly connected.



# Weight averaging from multiple trajectories

**Model soups** [Wortsman2022]: when fine-tuned from a shared pre-trained model with different hyperparams, weights remain linearly connected.



[Neyshabur2020] What is being transferred in transfer learning? NeurIPS.

[Wortsman2022] Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. ICML.

[Rame2022] DiWA: diverse weight averaging for out-of-distribution generalization. NeurIPS.

Weight averaging approximates ensembling.

Thus the 3 key criteria to trade-off:

### 1. Averageability

The weights should remain linearly connected.

### 2. Individual accuracies

The weights should be individually accurate.

### 3. Diversity

The weights should be sufficiently diverse to reduce variance.



# Era of open-source datasets and weights

[huggingface.co/datasets](https://huggingface.co/datasets)

|                                                                                            |                                                                                         |
|--------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|
| <b>cifar10</b><br>Preview · Updated 14 days ago · ↓ 18.5k · ♥ 8                            | <b>fashion_mnist</b><br>Preview · Updated 14 days ago · ↓ 17.5k · ♥ 9                   |
| <b>imagenet-1k</b><br>Preview · Updated Nov 3, 2022 · ↓ 9.82k · ♥ 57                       | <b>beans</b><br>Preview · Updated 14 days ago · ↓ 9.72k · ♥ 6                           |
| <b>mnist</b><br>Preview · Updated 14 days ago · ↓ 8k · ♥ 13                                | <b>cifar100</b><br>Preview · Updated 14 days ago · ↓ 6.91k · ♥ 3                        |
| <b>food101</b><br>Preview · Updated 14 days ago · ↓ 5.75k · ♥ 6                            | <b>Maysee/tiny-imagenet</b><br>Preview · Updated Jul 12, 2022 · ↓ 3.48k · ♥ 5           |
| <b>facebook/winoground</b><br>Preview · Updated 20 days ago · ↓ 3.13k · ♥ 27               | <b>rvl_cdip</b><br>Preview · Updated 14 days ago · ↓ 1.83k · ♥ 15                       |
| <b>sasha/dog-food</b><br>Preview · Updated Oct 25, 2022 · ↓ 1.48k · ♥ 1                    | <b>cats_vs_dogs</b><br>Preview · Updated 14 days ago · ↓ 1.1k · ♥ 5                     |
| <b>imagenet_sketch</b><br>Preview · Updated 14 days ago · ↓ 921                            | <b>svhn</b><br>Preview · Updated 14 days ago · ↓ 821 · ♥ 1                              |
| <b>Bingsu/Cat_and_Dog</b><br>Preview · Updated 13 days ago · ↓ 775                         | <b>competitions/aiornot</b><br>Preview · Updated 6 days ago · ↓ 765 · ♥ 17              |
| <b>frgm/imagenette</b><br>Preview · Updated Dec 11, 2022 · ↓ 712 · ♥ 3                     | <b>alkzar90/NIH-Chest-X-ray-dataset</b><br>Preview · Updated Nov 22, 2022 · ↓ 320 · ♥ 8 |
| <b>alkzar90/CC6204-Hackaton-Cub-Dataset</b><br>Preview · Updated 27 days ago · ↓ 221 · ♥ 1 | <b>nelorth/oxford-flowers</b><br>Preview · Updated Dec 11, 2022 · ↓ 221                 |
| <b>alkzar90/rock-glacier-dataset</b><br>Preview · Updated Dec 19, 2022 · ↓ 211 · ♥ 1       | <b>biglam/nls_chapbook_illustrations</b><br>Preview · Updated 25 days ago · ↓ 149 · ♥ 4 |

[huggingface.co/models/resnet-50](https://huggingface.co/models/resnet-50)

|                                                                                                    |                                                                                                     |
|----------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <b>nateraw/resnet50-oxford-iiit-pet</b><br>Preview · Updated Dec 3, 2021 · ↓ 16                    | <b>arize-ai/resnet-50-fashion-mnist-quality-drift</b><br>Preview · Updated Aug 1, 2022 · ↓ 16 · ♥ 2 |
| <b>ImageIN/resnet-50_finetuned</b><br>Preview · Updated Sep 23, 2022 · ↓ 16                        | <b>jayanta/resnet-50-finetuned-memes-v2</b><br>Preview · Updated Oct 19, 2022 · ↓ 16                |
| <b>jayanta/microsoft-resnet-50-cartoon-face-recogni...</b><br>Preview · Updated 16 days ago · ↓ 12 | <b>keithanpai/resnet-50-finetuned-eurosat</b><br>Preview · Updated Aug 1, 2022 · ↓ 11               |
| <b>morganchen1007/resnet-50-finetuned-resnet50</b><br>Preview · Updated Aug 24, 2022 · ↓ 11        | <b>arize-ai/resnet-50-cifar10-quality-drift</b><br>Preview · Updated Jul 22, 2022 · ↓ 10            |
| <b>morganchen1007/resnet-50-finetuned-resnet50_0831</b><br>Preview · Updated Sep 1, 2022 · ↓ 10    | <b>Francesco/resnet50-224-1k</b><br>Preview · Updated Feb 23, 2022 · ↓ 9                            |
| <b>torchxrayvision/resnet50-res512-all</b><br>Preview · Updated Jun 21, 2022 · ↓ 9                 | <b>Celal11/resnet-50-finetuned-FER2013CKPlus-0.003</b><br>Preview · Updated 9 days ago · ↓ 9        |
| <b>Celal11/resnet-50-finetuned-FER2013-0.003-CKPlus</b><br>Preview · Updated 9 days ago · ↓ 9      | <b>jayanta/resnet50-finetuned-memes</b><br>Preview · Updated Sep 17, 2022 · ↓ 8                     |
| <b>jayanta/resnet-50-FV2-finetuned-memes</b><br>Preview · Updated Oct 21, 2022 · ↓ 8               | <b>Francesco/resnet50</b><br>Preview · Updated Mar 1, 2022 · ↓ 7                                    |
| <b>YKXBCi/resnet-50-ucSat</b><br>Preview · Updated Jul 3, 2022 · ↓ 7                               | <b>jayanta/microsoft-resnet-50-cartoon-emotion-dete...</b><br>Preview · Updated 6 days ago · ↓ 7    |
| <b>nateraw/resnet50-beans-dummy-sagemaker</b><br>Preview · Updated Sep 22, 2021 · ↓ 6              | <b>YKXBCi/resnet-50-euroSat</b><br>Preview · Updated Jul 3, 2022 · ↓ 6                              |



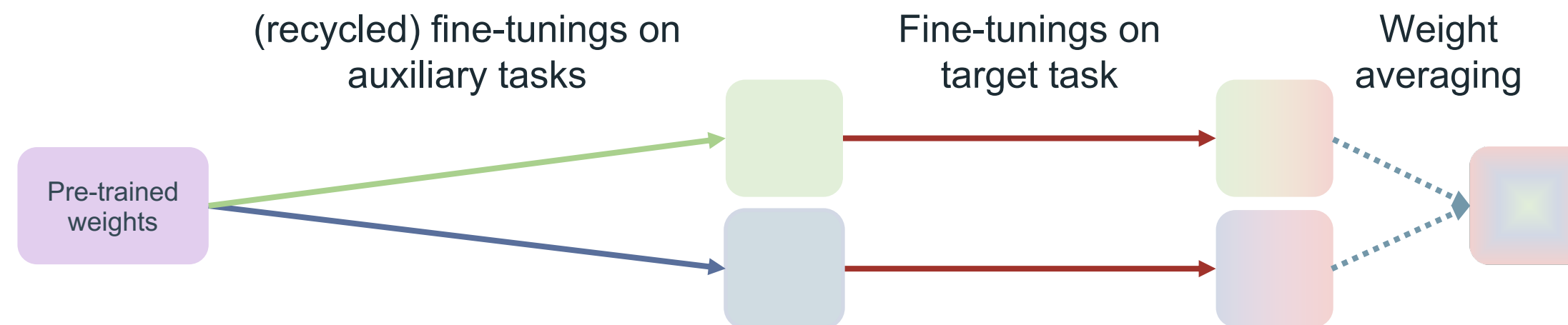
*Key idea:* recycle these weights as initializations for the target task.



# Model Ratatouille: Recycling Diverse Models for Out-of-Distribution Generalization

A simple strategy:

1. From a shared pre-trained network.
2. Recycle multiple fine-tunings on auxiliary tasks.
3. From these weights, launch multiple fine-tunings on the target task.
4. Average all the fine-tuned weights.



# Does Ratatouille meet the 3 key criteria for successful weight averaging ?

## 1. Averageability

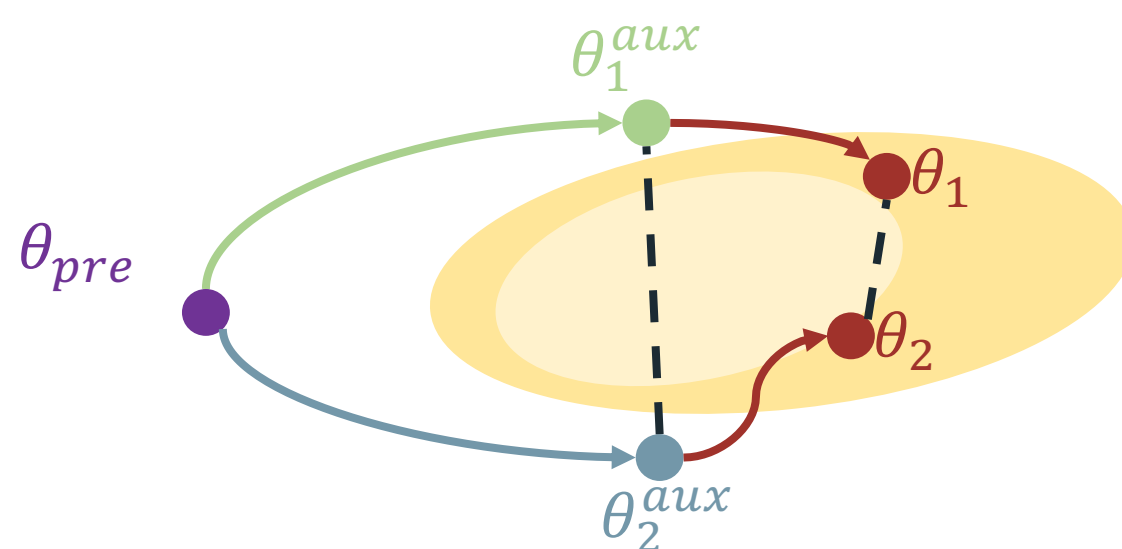
Yes, the weights remain linearly connected when auxiliary tasks are sufficiently similar.

## 2. Individual accuracies

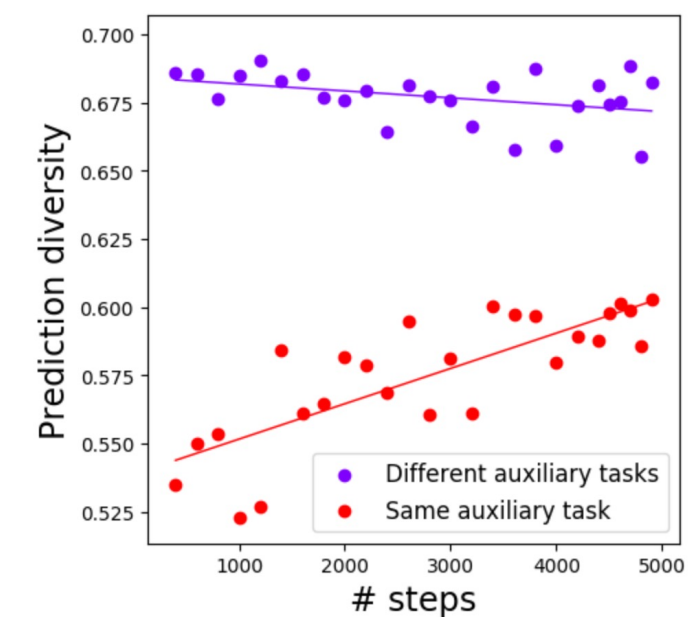
Yes, when the auxiliary tasks learn rich features, that help for the target task.

## 3. Diversity

Yes !! huge gain in diversity caused by different initialization and remains along fine-tuning on the target task.



[Phang2018] Sentence encoders on stilts: Supplementary training on intermediate labeled data tasks.  
[Choshen2022] Where to start? analyzing the potential value of intermediate models.



# New SoTA on DomainBed

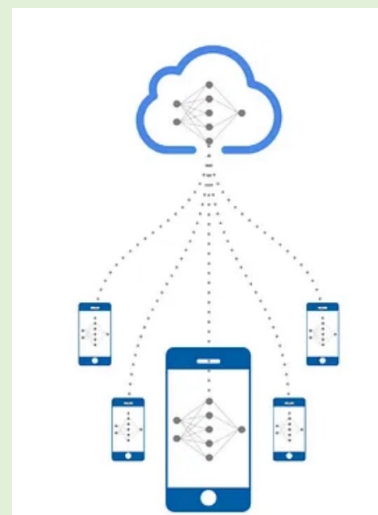
- Use the other datasets from DomainBed as the auxiliary datasets.
- No inference overhead.
- No training overhead if auxiliary weights are recycled.

| Algo               | Strategy                              | VLCS        | PACS        | OH          | Terra       | DomainNet   | Avg         |
|--------------------|---------------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| ERM                | Standard ft                           | 78.1        | 85.9        | 69.4        | 50.4        | 44.3        | 65.6        |
| Soups              | WA of networks with same inits        | 78.4        | 88.7        | 72.1        | 51.4        | 47.4        | 67.6        |
| Inter-training     | Auxiliary task                        | 77.7        | 89.0        | 69.9        | 46.7        | 44.5        | 65.6        |
| <b>Ratatouille</b> | <b>WA of networks with intertrain</b> | <b>78.5</b> | <b>89.5</b> | <b>73.1</b> | <b>51.8</b> | <b>47.5</b> | <b>68.1</b> |

# Updatable machine learning [Raffel2023]

## Data privacy concerns

- Only share weights,
  - data are kept private
- ⇒ scalable federated strategy.



## Collaboration and open-source

Foundation models come with risk of:

- Centralization.
- Lack of reproducibility.
- Two-speed research.

⇒ new collaborative solutions.

[github.com/r-three/git-theta](https://github.com/r-three/git-theta)



## Embarrassingly simple parallelization

Compute parallelism [Wortsman2022]!

- Simple engineering.
- Efficiency and training time.
- No waste: leverage all runs.
- Better compute scaling laws ?

|             | Average updates per second, normalized (↑) |                                   |                                                 |
|-------------|--------------------------------------------|-----------------------------------|-------------------------------------------------|
|             | fully synchronized<br>(TRANSFORMER-LM)     | partially synchronized<br>(DEMIX) | BTM: embarrassingly parallel<br>(branched ELMs) |
| <b>125M</b> | 1.00                                       | 1.01                              | 1.05                                            |
| <b>350M</b> | 1.00                                       | 1.11                              | 1.23                                            |
| <b>750M</b> | 1.00                                       | 1.01                              | 1.27                                            |
| <b>1.3B</b> | 1.00                                       | 0.97                              | 1.33                                            |

## Conclusion

- Linear mode connectivity across weights fine-tuned on different tasks
- New ratatouille strategy for out-of-distribution generalization
- Code is available: [github.com/facebookresearch/ModelRatatouille](https://github.com/facebookresearch/ModelRatatouille)

Thank you for your attention