

PUTTING A FACE TO FORGETTING

CONTINUAL LEARNING MEETS MECHANISTIC INTERPRETABILITY

Sergi Masip, Gido van de Ven, Javier Ferrando, Tinne Tuytelaars



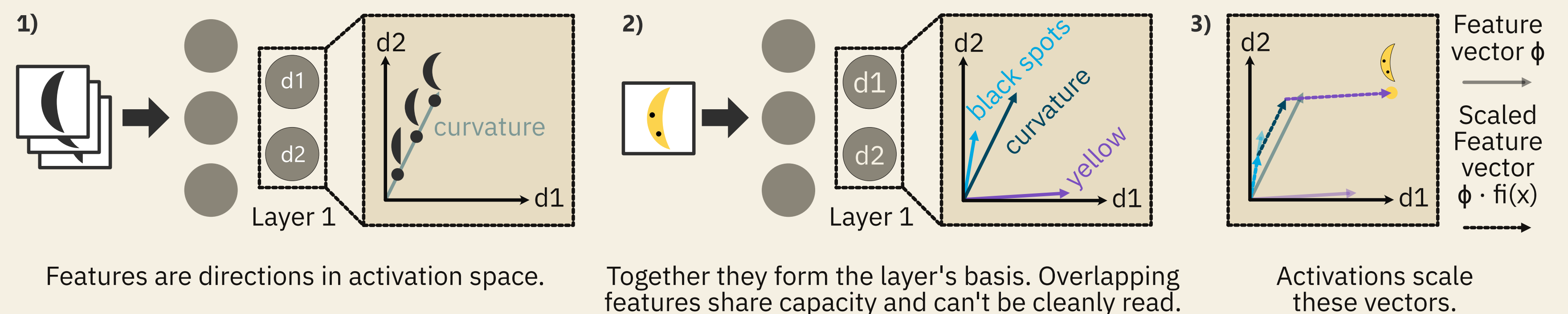
SCAN FOR
PROJECT WEBSITE

MOTIVATION

Forgetting is diagnosed by accuracy drops or last-layer representation similarity. That tells us that it happened, **not how**.

Missing: a description of what happens to individual features.

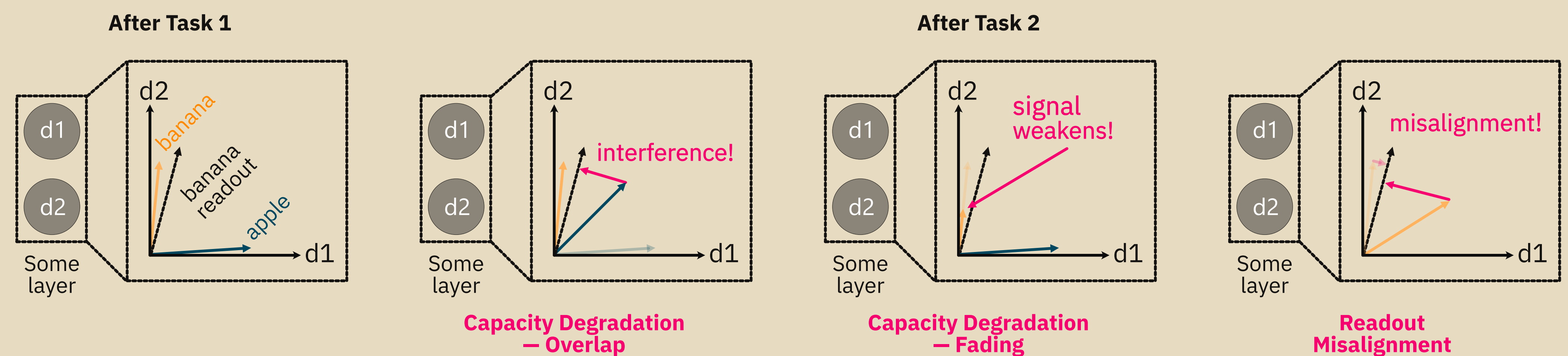
ANATOMY OF A FEATURE



THE SUSPECTS

[NOT RECOVERABLE]
Capacity degradation — Overlap
Capacity degradation — Fading

[RECOVERABLE]
Readout misalignment



A MODEL FOR THE CASE

A layer encodes features Φ ; a probe w reads them out. Train on a new task and both move toward each other.

$$\hat{y}(x) = w^\top \Phi f(x) \quad \mathbb{E}_{x \sim \mathcal{D}_B} [\Delta \phi_i] = -\eta \left(\sum_j \gamma_j^{(B)} \Sigma_{ij}^{(B)} - \beta_i^{(B)} \right) w_{(B)}$$

β — how predictive the feature is for the task
 γ — how much the probe leans on it
 Σ — how often features fire together

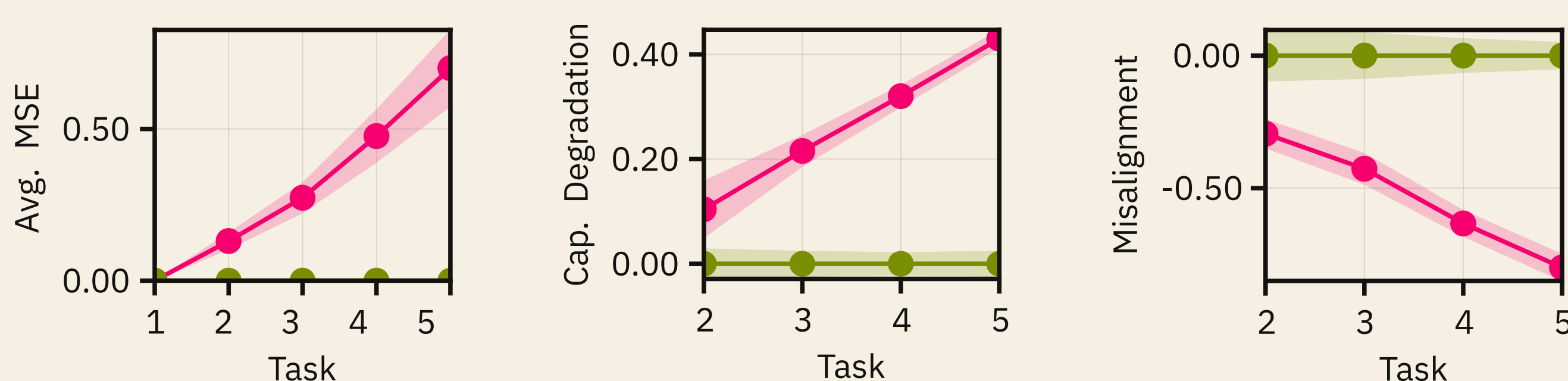
WHO GETS HURT

[SAFE]
Disjoint features. The old feature never fires for the new task, so Σ and β are both zero. No gradient reaches it.

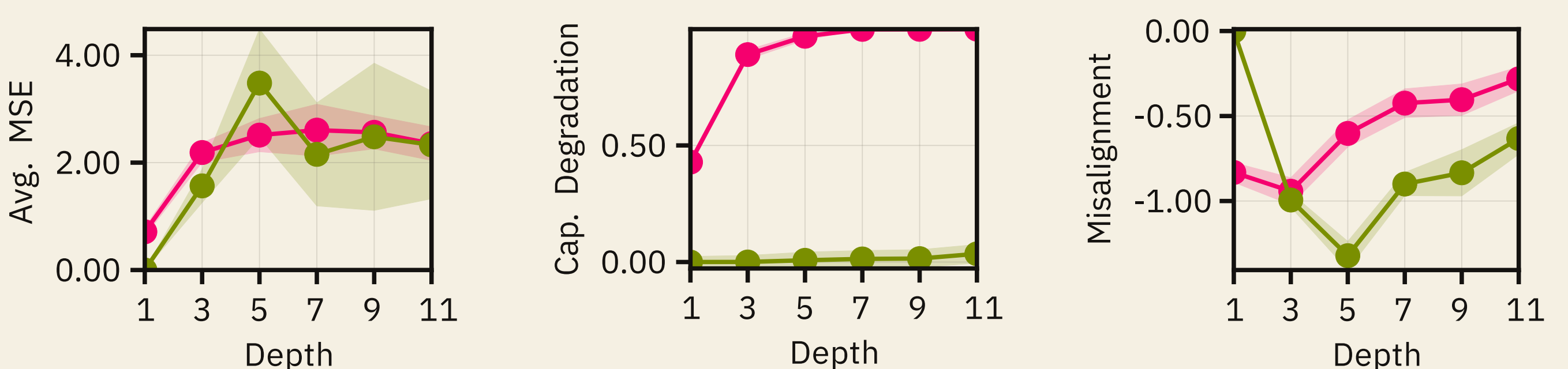
[WORST CASE]
Active but uninformative (intermediate similarity). The feature fires constantly yet predicts nothing. It gets pushed around meaninglessly.

THE EMPIRICAL EVIDENCE

SHARED FEATURES HURT MOST



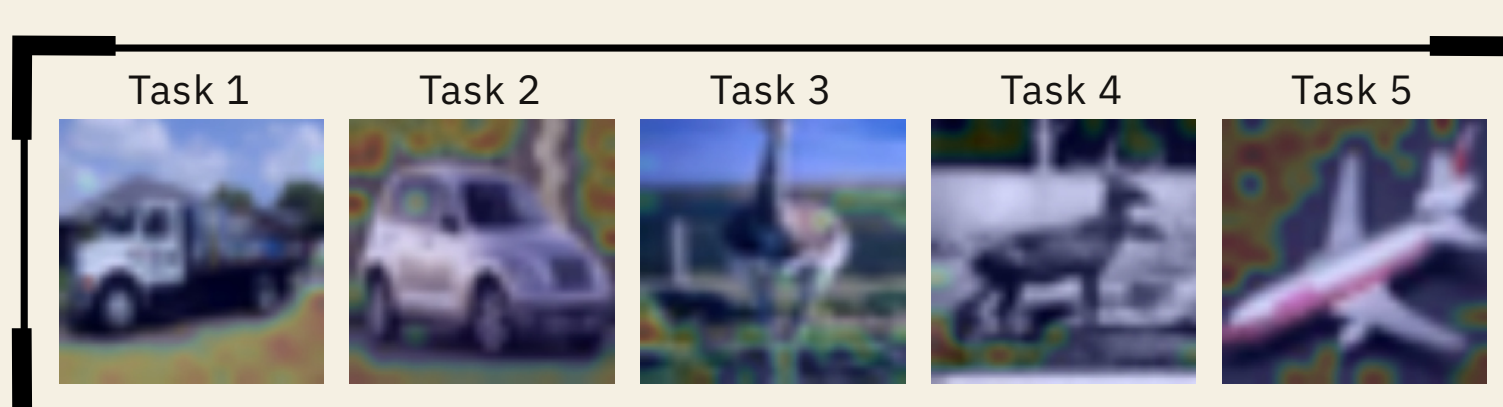
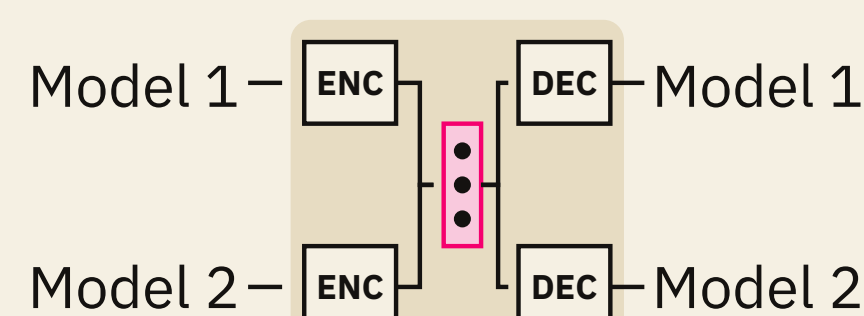
DEPTH MAKES IT WORSE



● Active but uninformative
● Disjoint features

TAKING IT TO THE FIELD

Crosscoders let us track the same feature across a sequence of models.



[PRELIMINARY]
On a ViT trained on task-incremental Split-CIFAR-10, fading dominates.

TAKEAWAY

Forgetting has a geometry: features overlap, fade, and drift from their readouts.

Mitigating it means **preserving features across layers**, not just constraining weights or last-layer representations.