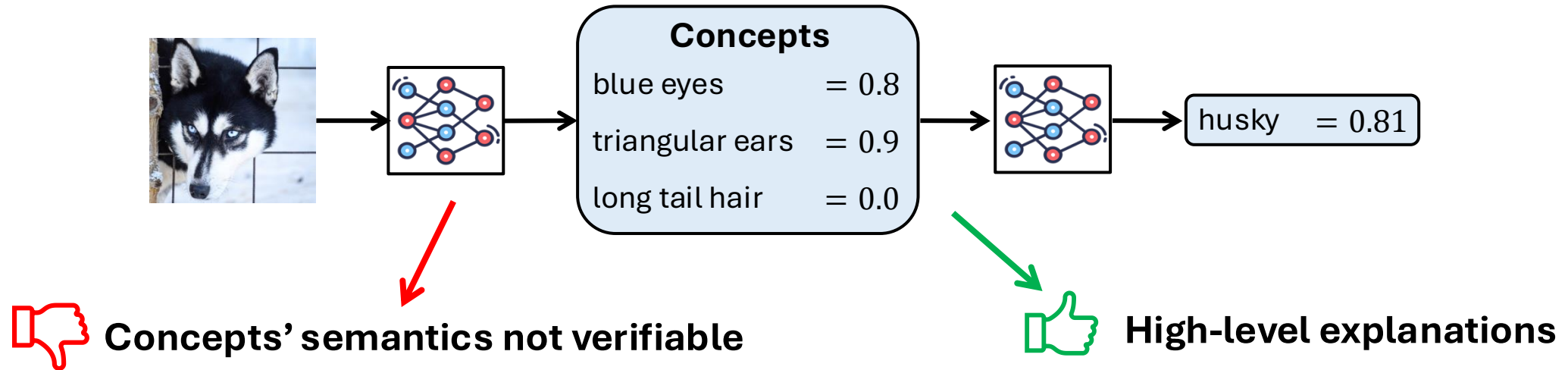


Prototype-Grounded Concept Models for Verifiable Concept Alignment

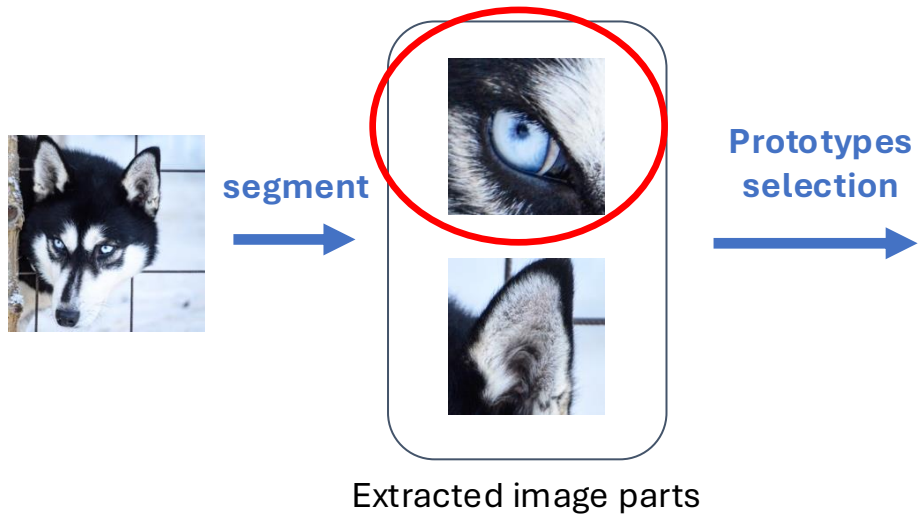
Stefano Colamonaco, David Debot, Pietro Barbiero, Giuseppe Marra

Concept Bottleneck Models (CBMs)

CBMs = intrinsically explainable models that predict **concepts**, using them to predict a downstream task
= high-level, human-understandable features

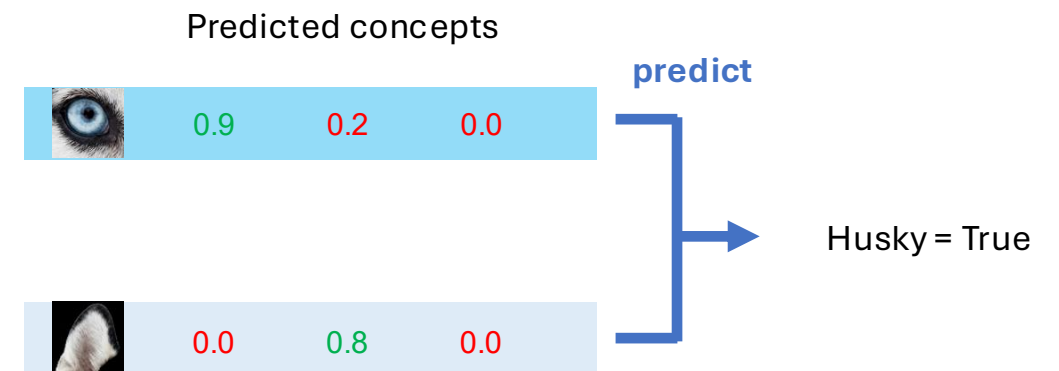


Prototype-Grounded Concept Models (PGCMs)



CONCEPT ALIGNMENT TABLE

	Almond eyes	Triangular ear	Snowman
	0.9	0.2	0.0
	0.0	0.8	0.0
	0.0	0.0	1.0



PGCMs' concept alignment table allows verifying concept alignment

Each row in the table can be
verified by the human

Image	Active Concepts
	Bangs, Blond Hair, Wavy Hair
	Heavy Makeup, No Beard, Pale Skin
	Digit is 4
	Small, Cyan, Sphere, Metal
	Small, Yellow, Cube, Rubber

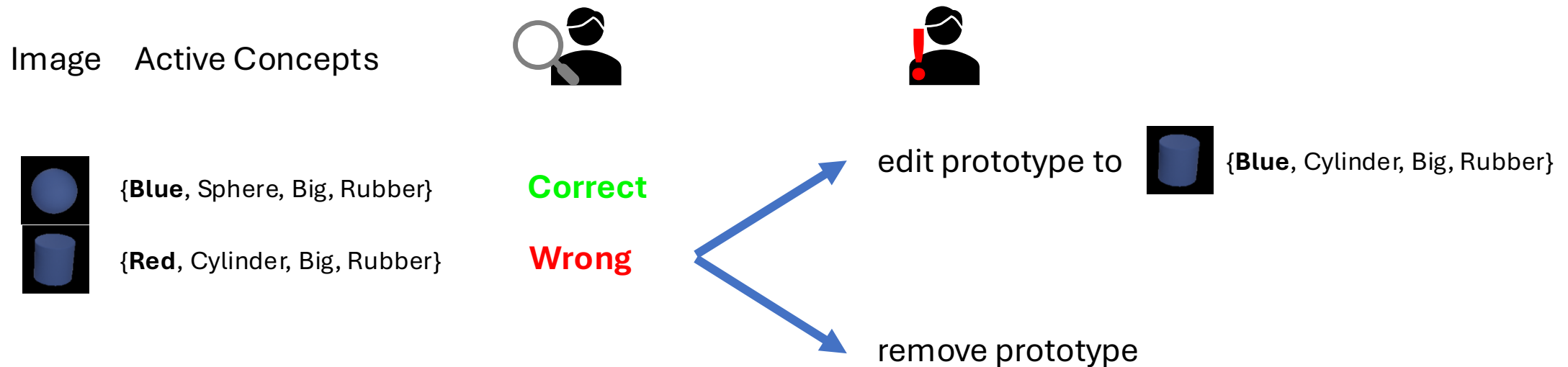
Each concept's semantics is given by
the prototypes for which it is active

Concept	Semantics
Brown Hair	 ∨ 
Pointy Nose	 ∨ 
Digit is 3	 ∨ 
Blue	 ∨  ∨ ...
Cube	 ∨  ∨ ...

PGCMs enable model interventions to re-align learned concepts

Misalignment detected by the human?

We can **intervene** on the table, removing or editing incorrect rows!



PGCMs are more responsive to concept interventions than CBMs

Standard CBMs: concepts are modelled independently of each other

Intervention
 C_1 is True



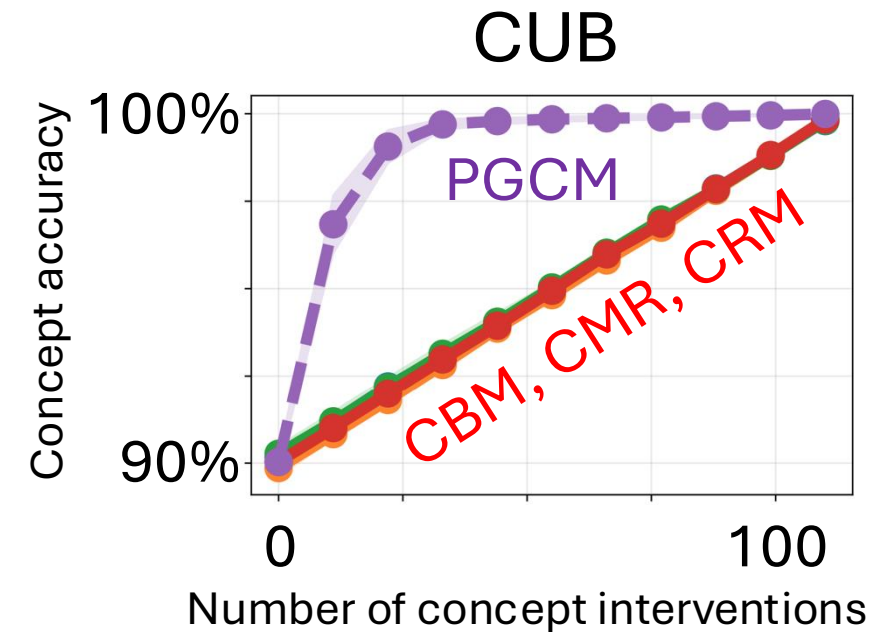
Set **only** C_1 to True

PGCMs: concepts are inter-dependent through the prototype selection

Intervention
 C_1 is True



Ignore **all** prototypes where C_1 is False, potentially affecting all concepts



PGCMs achieve similar accuracy as standard CBMs

Model	ColorMNIST+		CelebA		CUB	
	Concept	Task	Concept	Task	Concept	Task
CBM	99.2	99.6	81.3	84.0	89.9	63.6
PGCM	99.2	99.7	78.5	83.0	90.7	63.4

Conclusion

Use PGCMs to create interpretable models

with verifiable concept alignment

through visually-grounded concept prediction