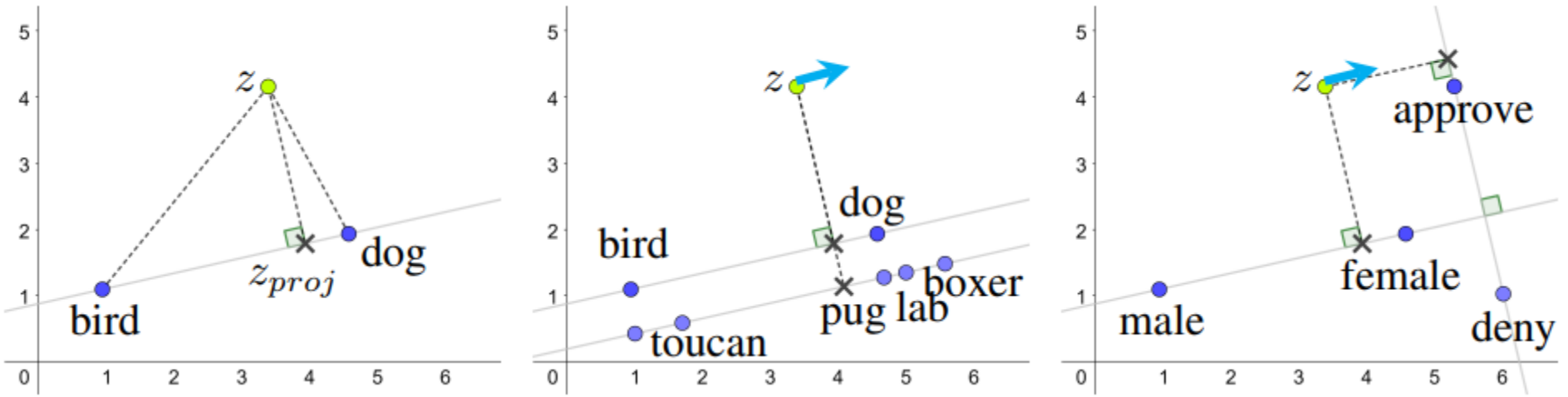
Prototype Based Classification from Hierarchy to Fairness

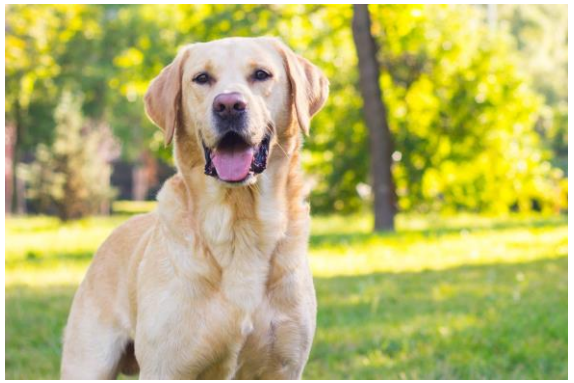


Mycal Tucker mycal@mit.edu

Julie Shah julie_a_shah@csail.mit.edu

Hierarchical classification

Given x, classify it at many levels



L0: Animal
L1: Dog
L2: Labrador
…

Hierarchical classification

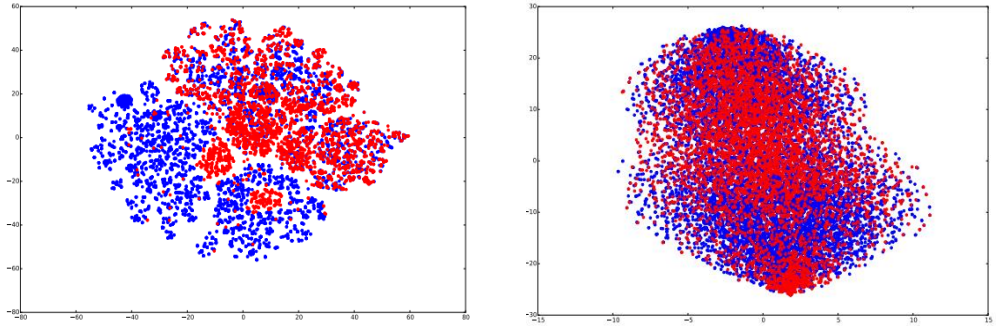
Given x, classify it at many levels

L0: Animal
L1: Dog
L2: Labrador
…

Fair classification

Given x, predict y regardless of s

Should person A (x) get a loan (y) regardless of their sex (s)?



From *The Variational Fair Autoencoder.* Louizos et al. ICLR 2016.

Hierarchical classification

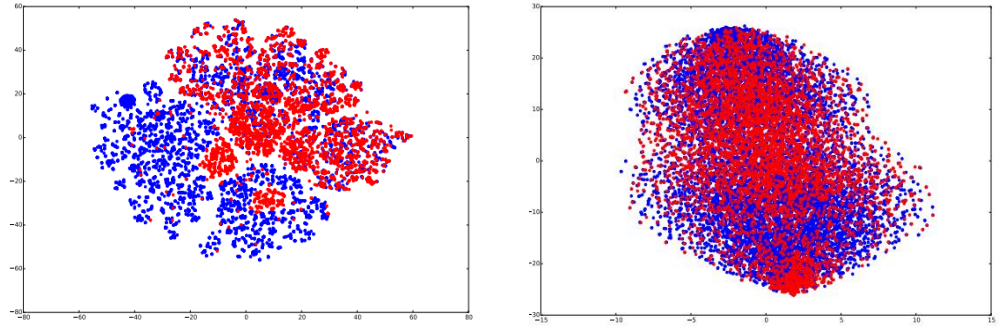
Given x, classify it at many levels

L0: Animal
L1: Dog
L2: Labrador
…

Fair classification

Given x, predict y regardless of s

Should person A (x) get a loan (y) regardless of their sex (s)?



Separate problems solved by separate architectures

From *The Variational Fair Autoencoder.* Louizos et al. ICLR 2016.

Can we build a classifier that

Uses interpretable representations?

Supports multiple concept relationships?

Can we build a classifier that

Uses interpretable representations?

Prototype-based classification

Supports multiple concept relationships?

Training losses for concept “alignment”

Can we build a classifier that

Uses interpretable representations?

Prototype-based classification

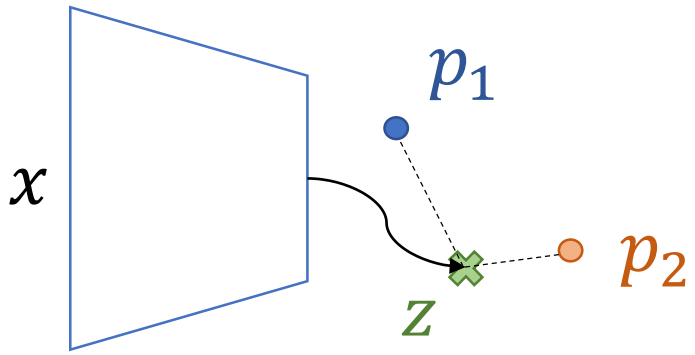
Supports multiple concept relationships?

Training losses for concept “alignment”

Concept Subspace Network

Prototype Based Classification

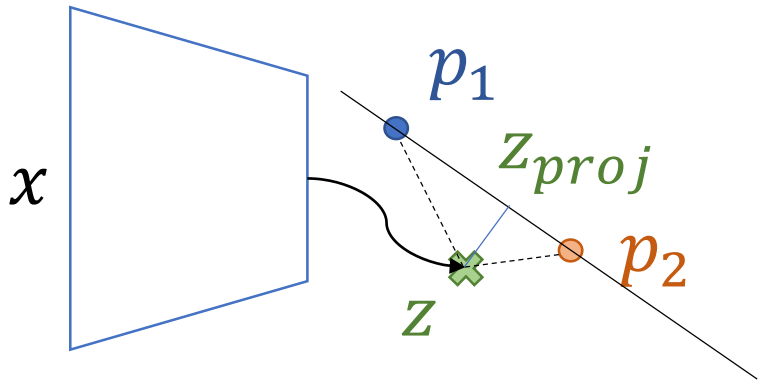
- Encode x into a latent representation z
- Measure Euclidean distance to each (learnable) prototype



$$\mathbb{P}(p_i|z) \propto e^{-(p_i - z)^2}$$

Prototype Based Classification

- Encode x into a latent representation z
- Measure Euclidean distance to each (learnable) prototype

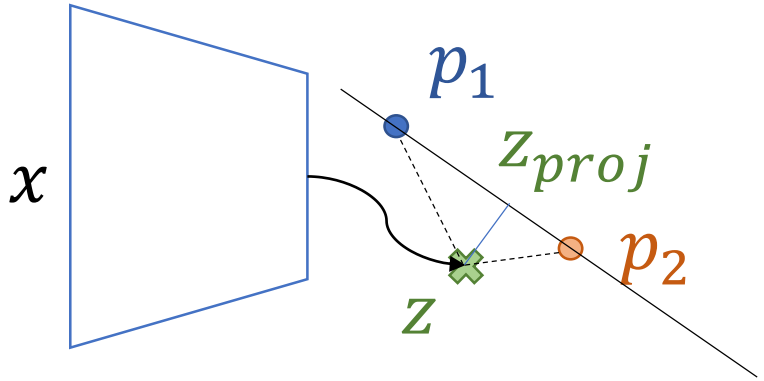


$$\mathbb{P}(p_i|z) \propto e^{-(p_i - z)^2}$$

Prototypes define a “Concept Subspace” for classification.

Prototype Based Classification

- Encode x into a latent representation z
- Measure Euclidean distance to each (learnable) prototype



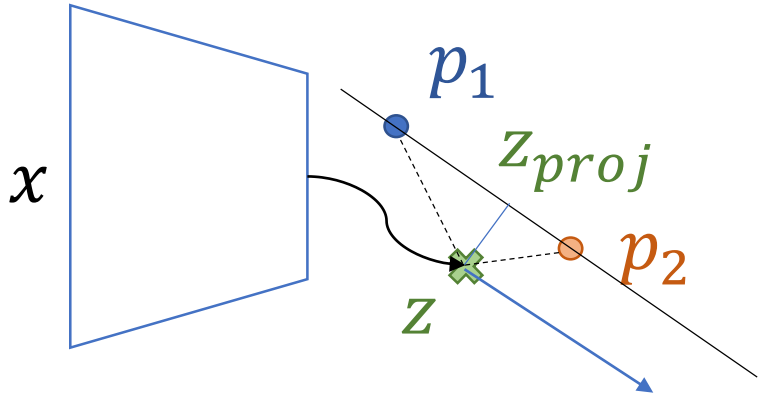
$$\mathbb{P}(p_i|z) \propto e^{-(p_i - z)^2}$$

$$\propto e^{-\left(p_i - z_{proj}\right)^2}$$

Prototypes define a “Concept Subspace” for classification.

Prototype Based Classification

- Encode x into a latent representation z
- Measure Euclidean distance to each (learnable) prototype



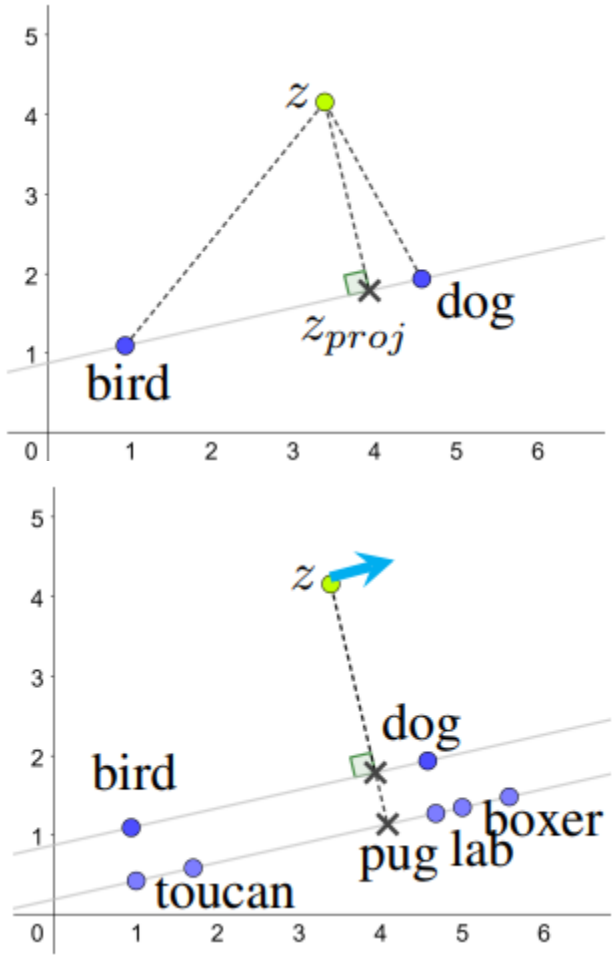
$$\mathbb{P}(p_i|z) \propto e^{-(p_i - z)^2}$$

$$\propto e^{-\left(p_i - z_{proj}\right)^2}$$

Prototypes define a “Concept Subspace” for classification.
Moving relative to that space changes classifications.

Enabling multi-concept classification

- Introduce a new set of prototypes for each classification task:
  - Bird vs. dog
  - Toucan vs. Sparrow vs. Pug vs. Lab etc.

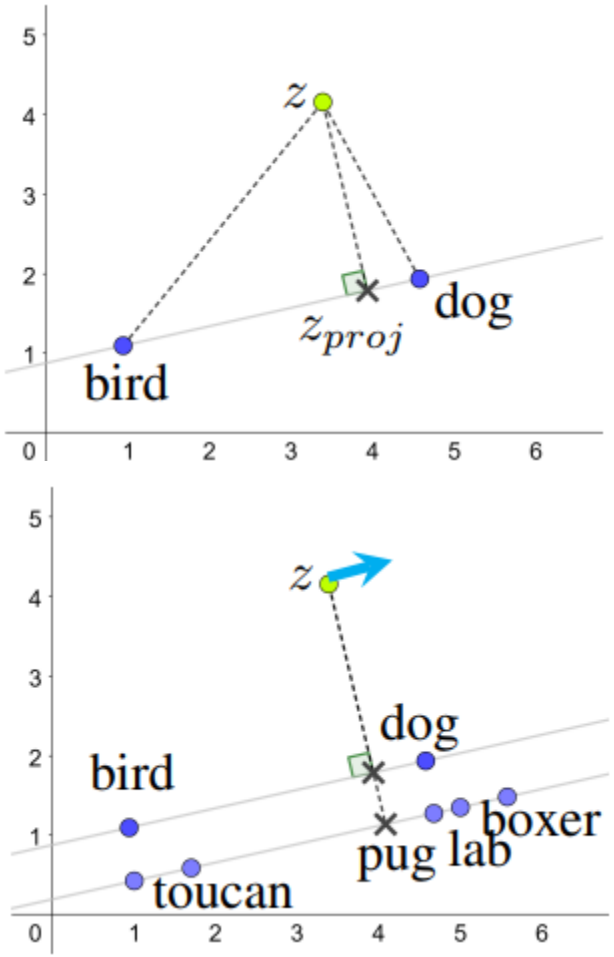


Enabling multi-concept classification

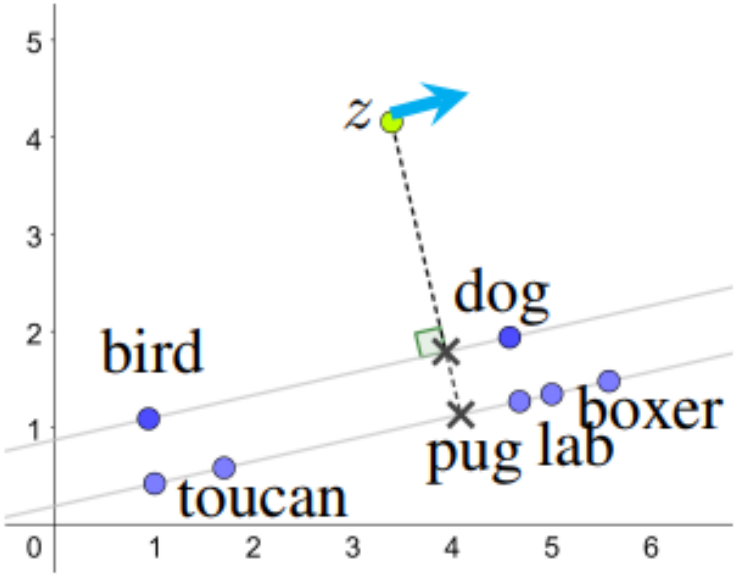
- Introduce a new set of prototypes for each classification task:
  - Bird vs. dog
  - Toucan vs. Sparrow vs. Pug vs. Lab etc.

- Each set defines a subspace
- Measure *alignment* between subspaces
  - $a = 1$ parallel
  - $a = 0$ orthogonal

$$a(Q_1, Q_2) = \frac{1}{mn} \sum_{i}^{m} \sum_{j}^{n} (Q_1^\top Q_2[i, j])^2$$

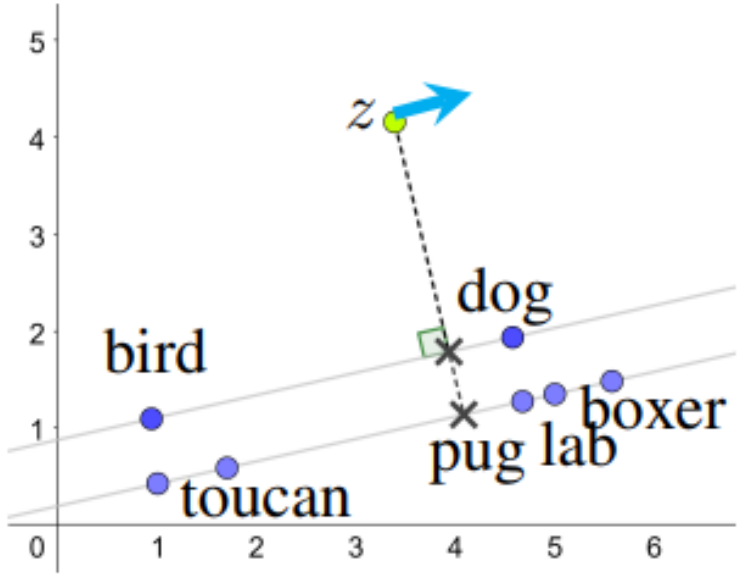


Multi-concept classification

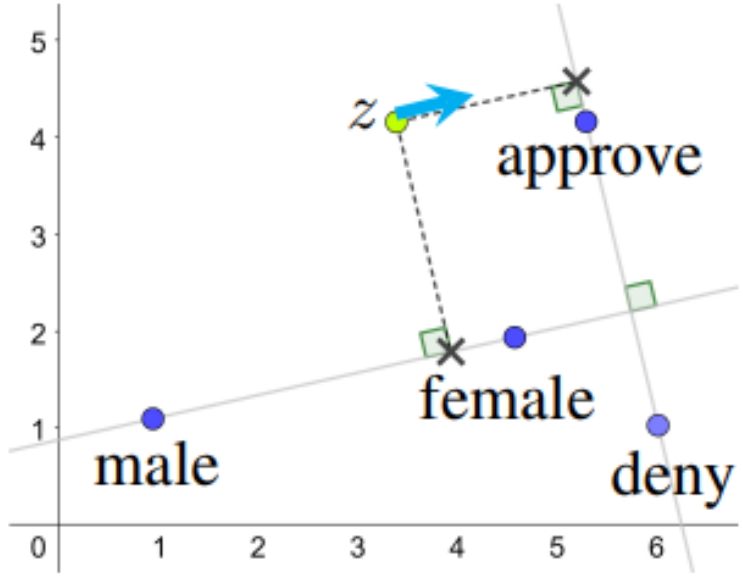


Hierarchical classification → Parallel subspaces

Multi-concept classification



Hierarchical classification → Parallel subspaces



Fair classification → Orthogonal subspaces

Results: Fair Classification

Match SOTA, compared to other fair classifiers

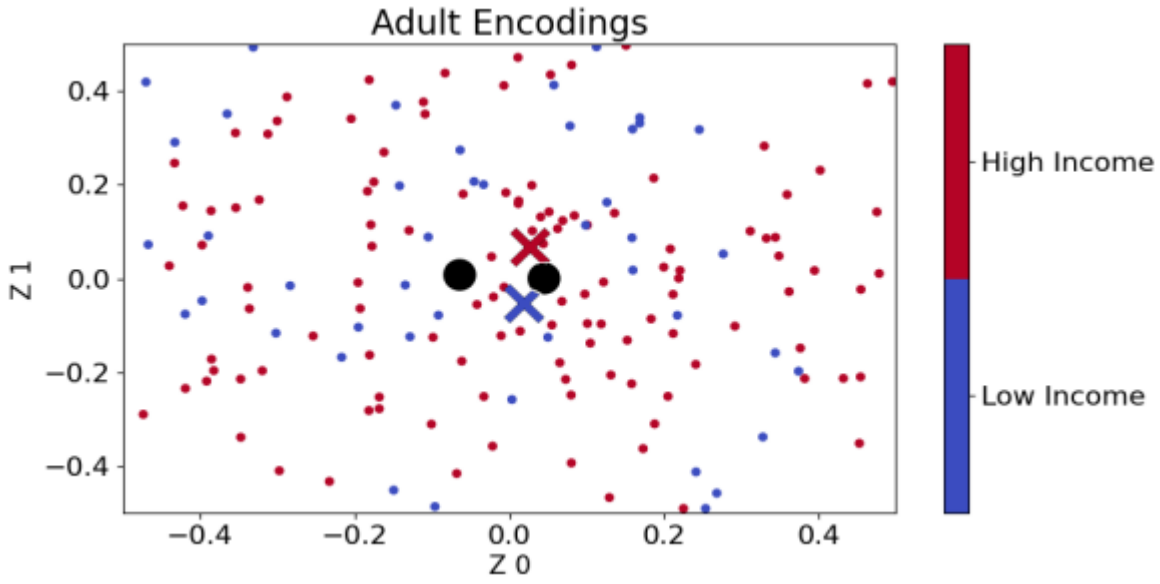


Table 1: Mean Adult dataset fairness results.

Model	y Acc.	s Acc.	D.I.	DD-0.5
CSN	0.85	0.67	0.83	0.16
Adv.	0.85	0.67	0.87	0.16
VFAE	0.85	0.70	0.82	0.17
FR Train	0.85	0.67	0.83	0.16
Wass. DB	0.81	0.67	0.92	0.08
Random	0.76	0.67		

Table 2: Mean German dataset fairness results.

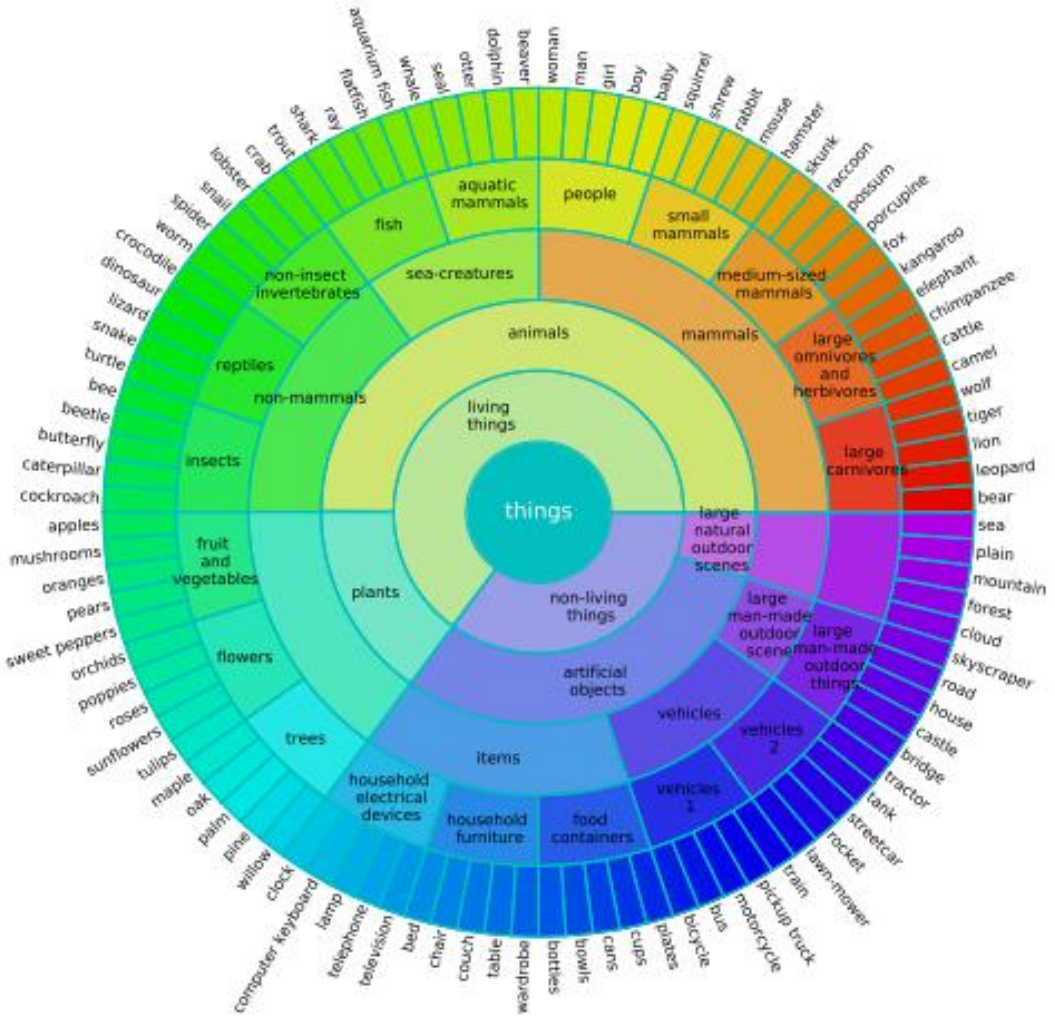
Model	y Acc.	s Acc.	D.I.	DD-0.5
CSN	0.73	0.81	0.70	0.10
Adv.	0.73	0.81	0.63	0.10
VFAE	0.72	0.81	0.47	0.23
FR Train	0.72	0.80	0.55	0.16
Wass DB	0.72	0.81	0.33	0.02
Random	0.70	0.81		

Results: Hierarchical Classification

Achieve SOTA (for given backbone) on CIFAR100

- Greater overall accuracy
- Lower cost of errors

	Y_0%	Y_1%	A.C.
CSN	**0.76** (0.0)	0.85 (0.0)	**0.76** (0.02)
HPN	0.71 (0.0)	0.80 (0.0)	0.97 (0.04)
Init.	0.01	0.05	3.88
CSN	**0.78** (0.0)	0.88 (0.0)	**0.91** (0.0)
MGP	0.76	-	1.05
Init.	0.01	0.05	7.33

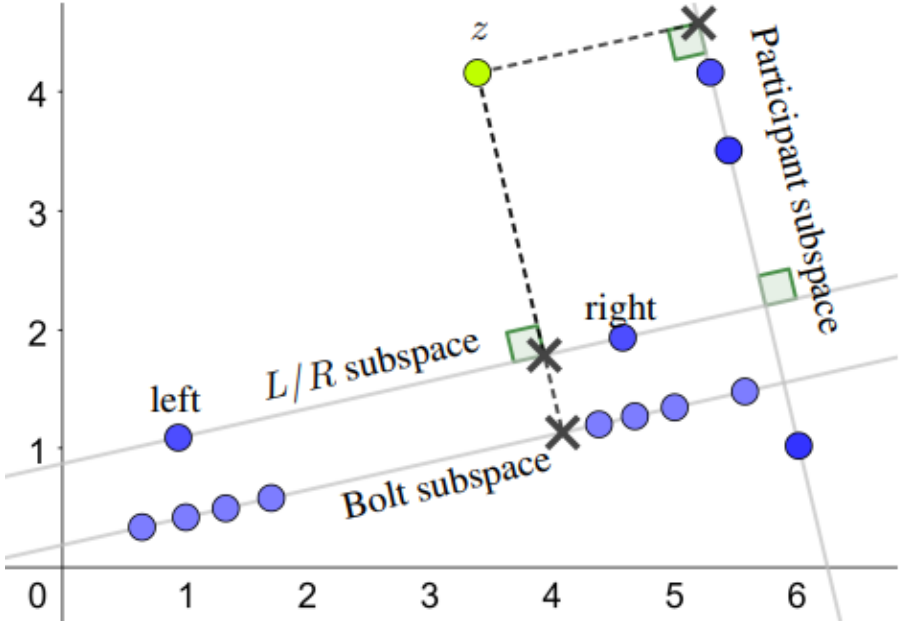


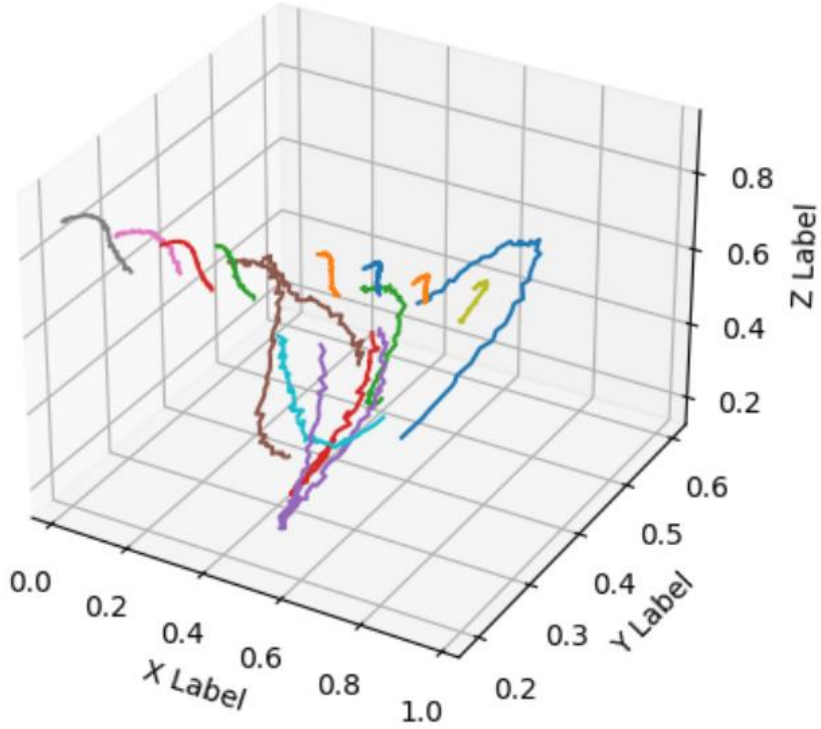
From *Leveraging Class Hierarchies with Metric-Guided Prototype Learning*. Garnot and Landrieu 2021.

Results: Fair and Hierarchical Classification

In human study:

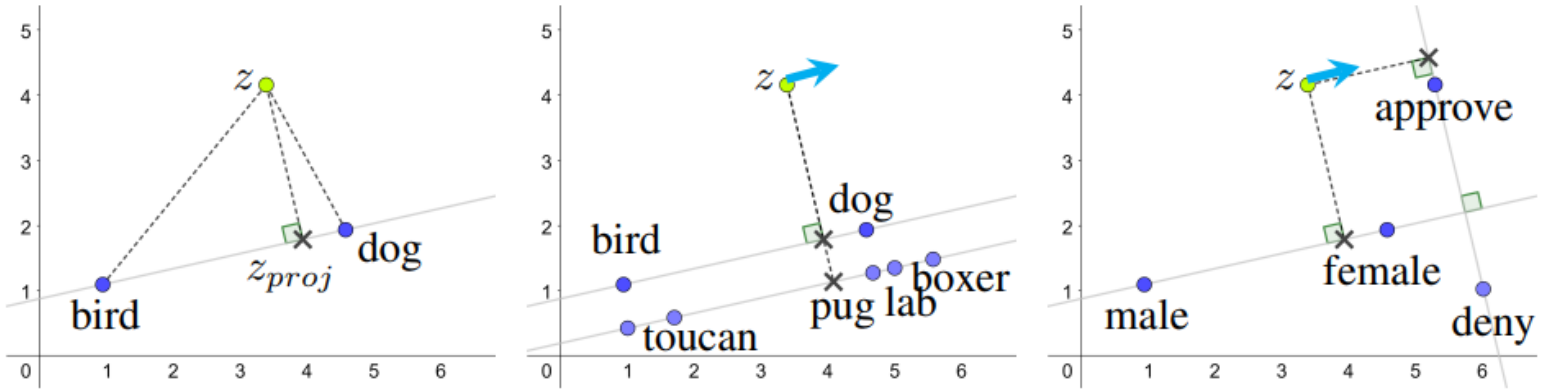
1) Preserve participant privacy
2) Exploit hierarchical structure





Prototype Based Classification from Hierarchy to Fairness

1) Interpretable representations with controllable relations
2) Unified framework for diverse classification tasks



Mycal Tucker mycal@mit.edu

Julie Shah julie_a_shah@csail.mit.edu