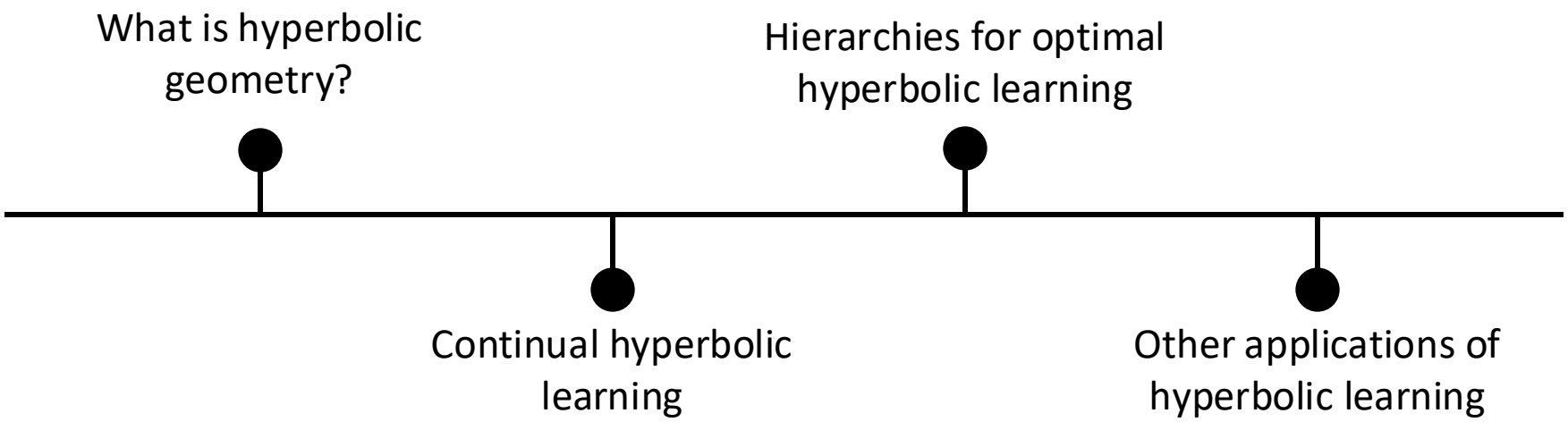
UNIVERSITY OF AMSTERDAM
Faculty of Science

Hyperbolic Deep Learning

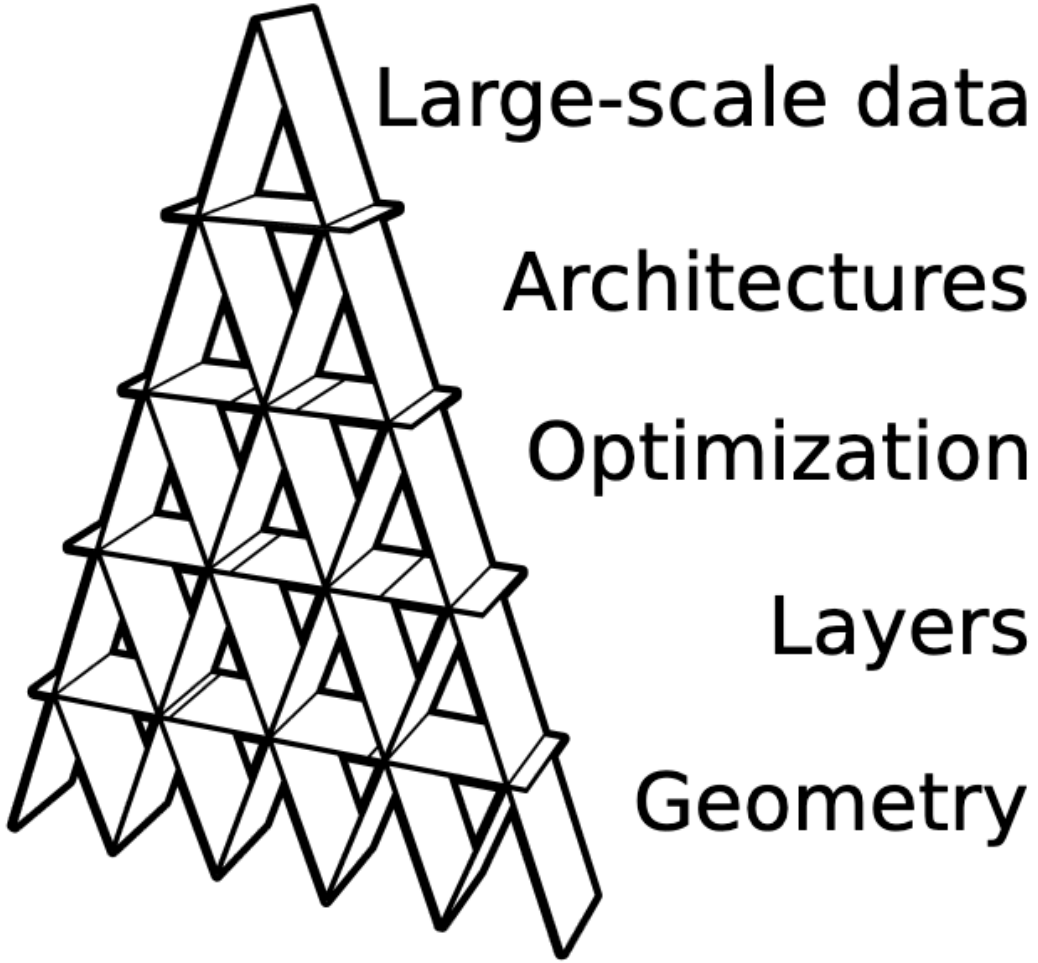
Melika Ayoughi
m.ayoughi@uva.nl

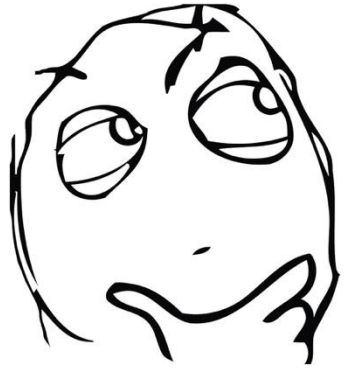
November 19th 2025

Music Cognition Group


What is hyperbolic geometry?
Hierarchies for optimal hyperbolic learning
Continual hyperbolic learning
Other applications of hyperbolic learning

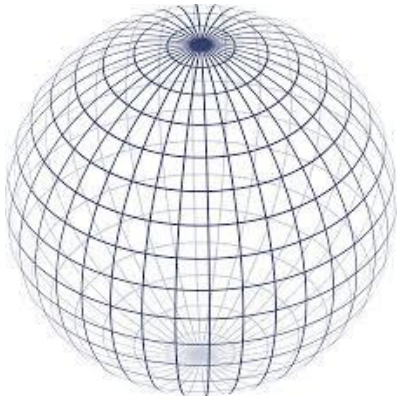
Euclidean = The default geometry for deep learning



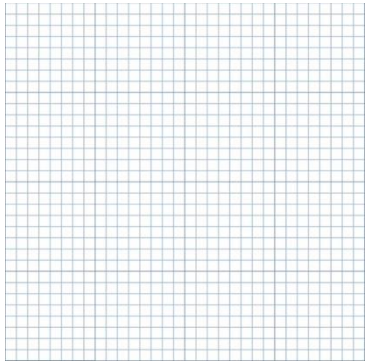


Other geometries exist!

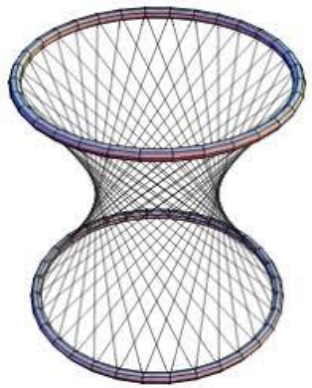
Elliptical



Euclidean



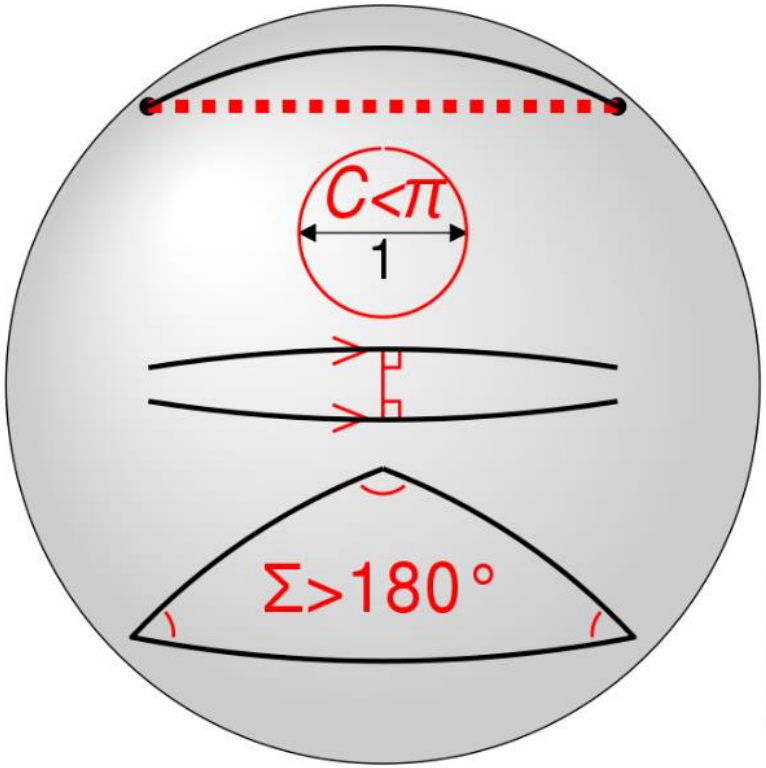
Hyperbolic



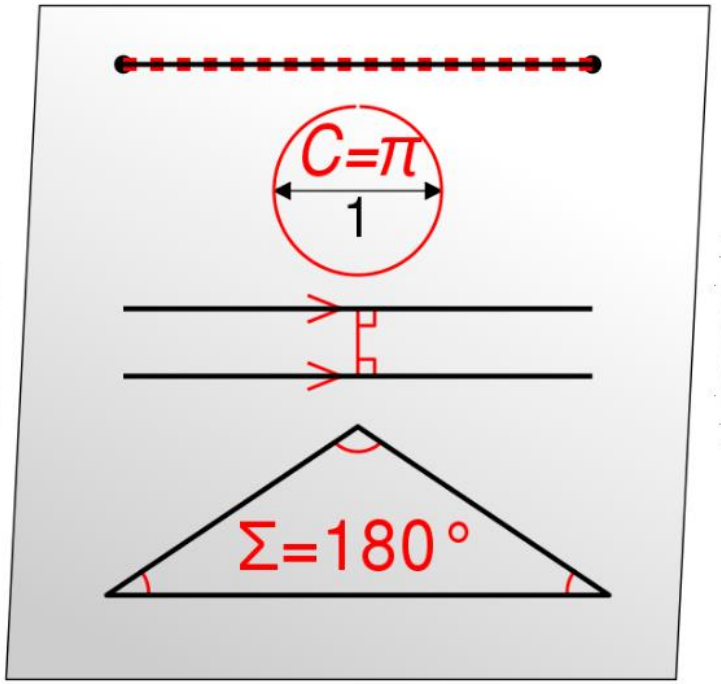
positive 0 negative

Properties of non-Euclidean geometries

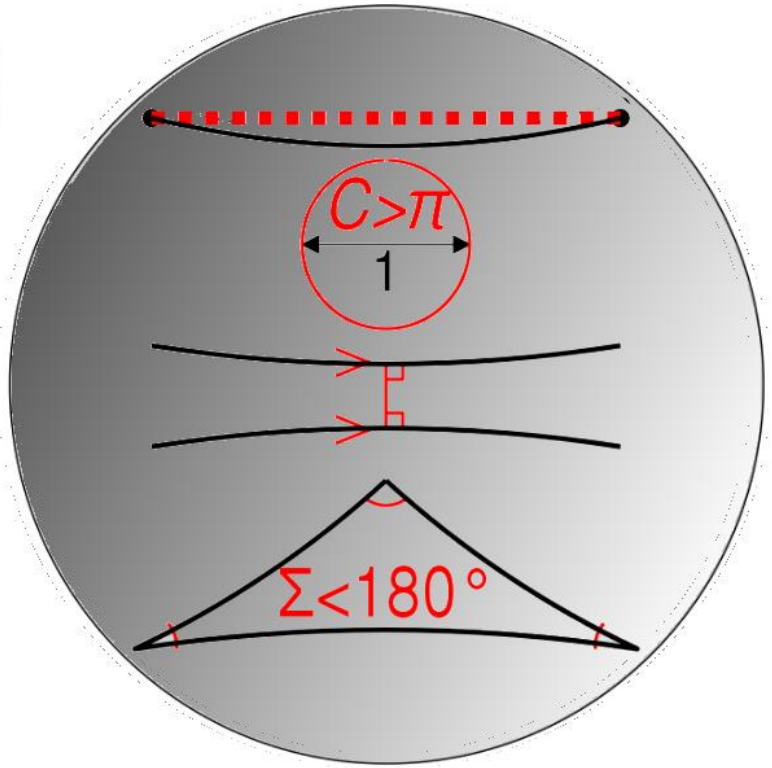
Elliptical



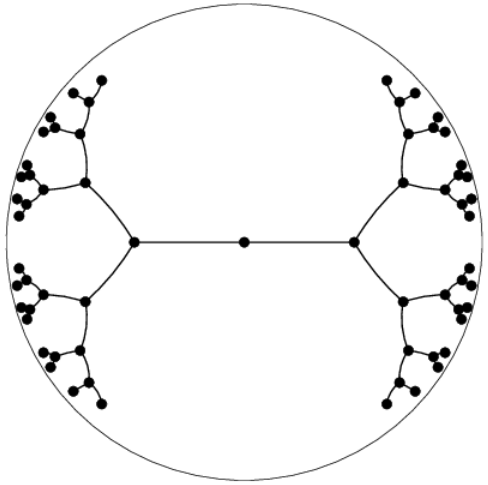
Euclidean

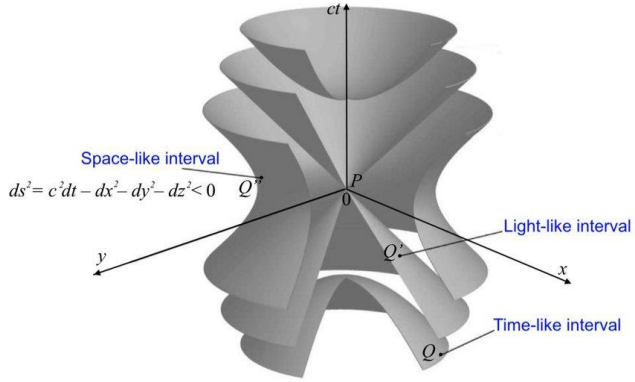


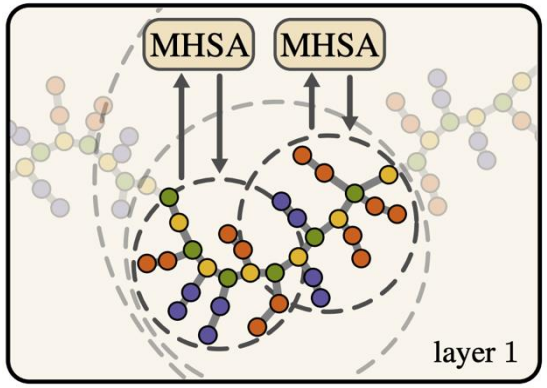
Hyperbolic

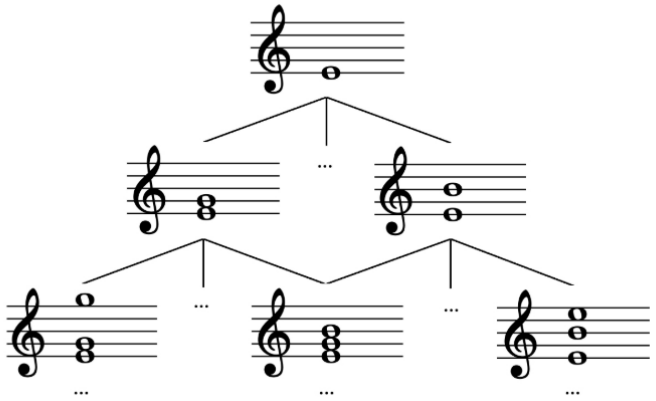


Why should we care about other geometries?



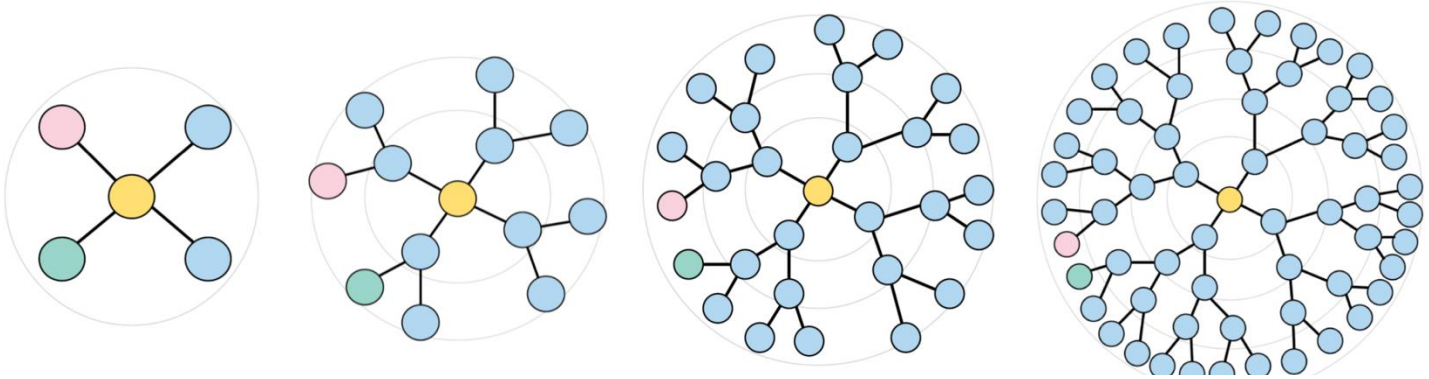




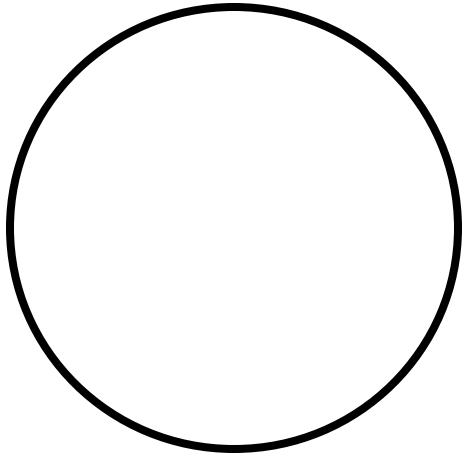


Polynomial vs exponential growth

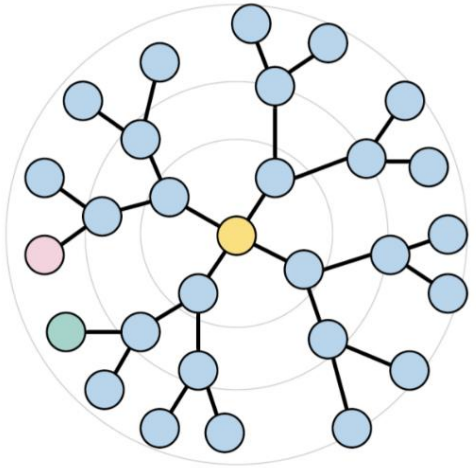
Euclidean



[Bachmann et al. ICML 2020]



$v = \pi r^2$

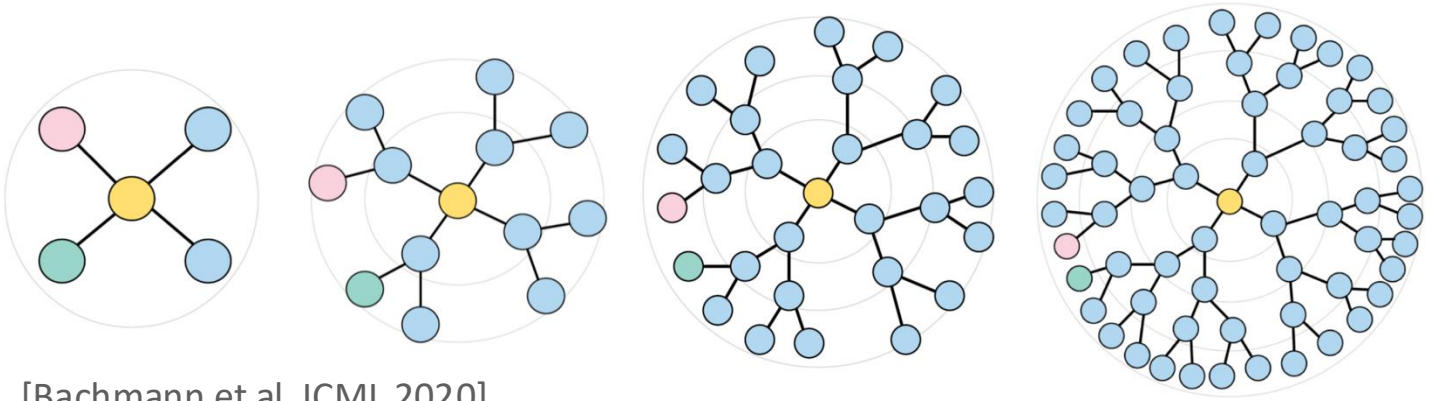


$n = b^d$

Euclidean space and hierarchies are a mismatch: linear vs. exponential growth.

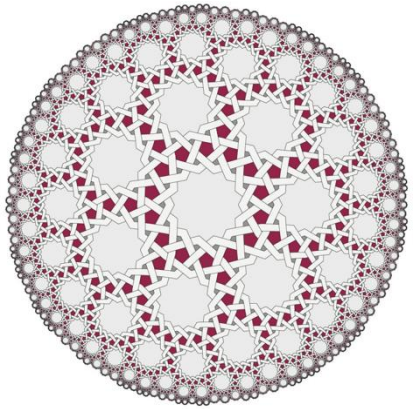
Optimal geometry for hierarchical representation

Euclidean



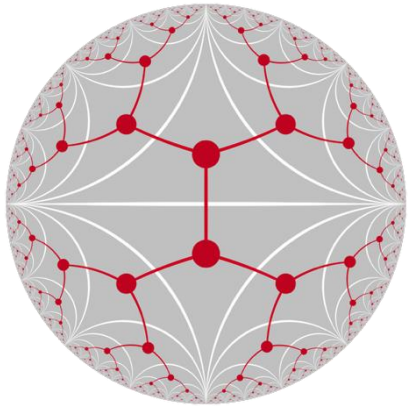
[Bachmann et al. ICML 2020]

Hyperbolic



[KAPLAN et al. TOG 2004]

[M.C.Escher's "Circle Limit III"]



"Hyperbolic space can be thought of as a **continuous** analogue to **discrete trees**"

Where do **you** think hyperbolic learning could be useful?

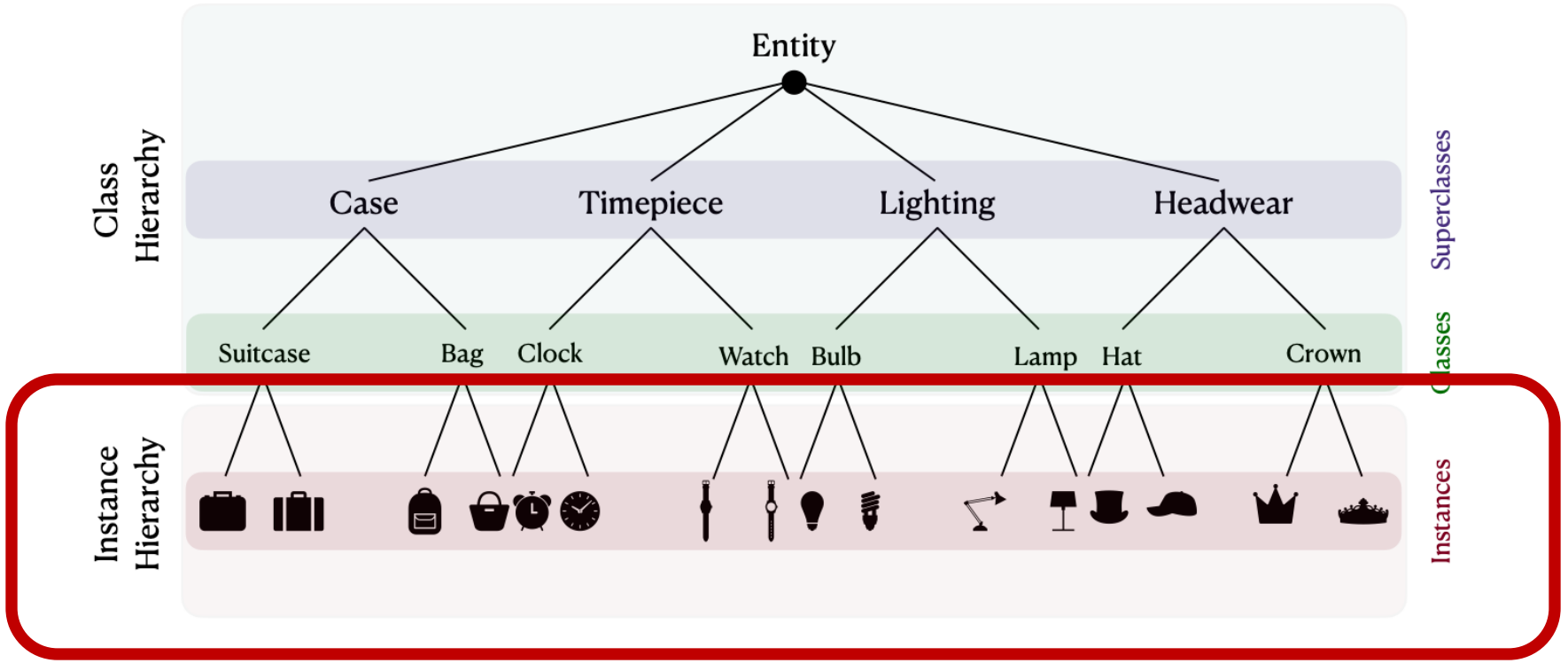
UNIVERSITY OF AMSTERDAM
Faculty of Science

INDE lab
INtelligent Data Engineering

VIS LAB
VIDEO & IMAGE SENSE LAB

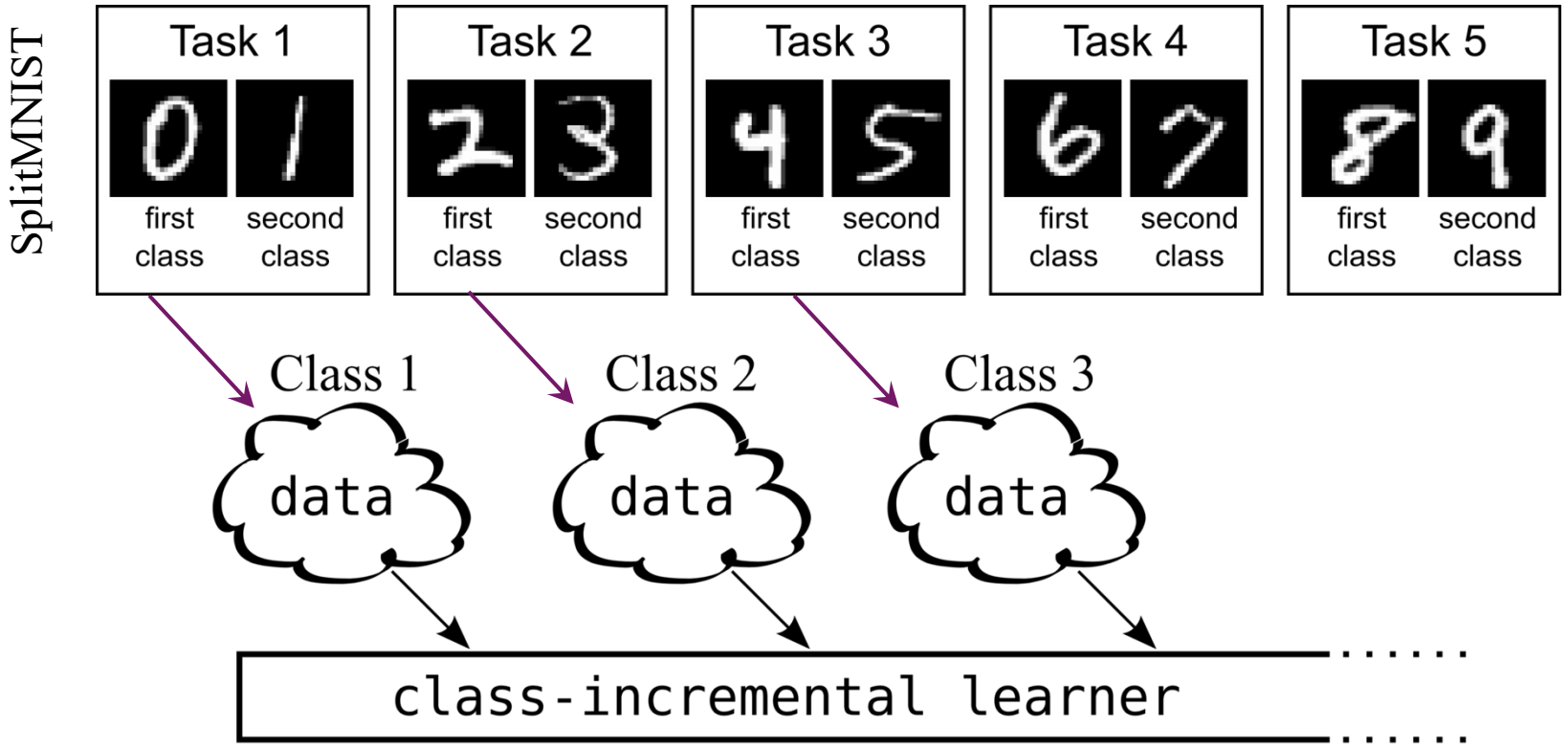
Continual hyperbolic learning of instances and classes

Instance-level classification

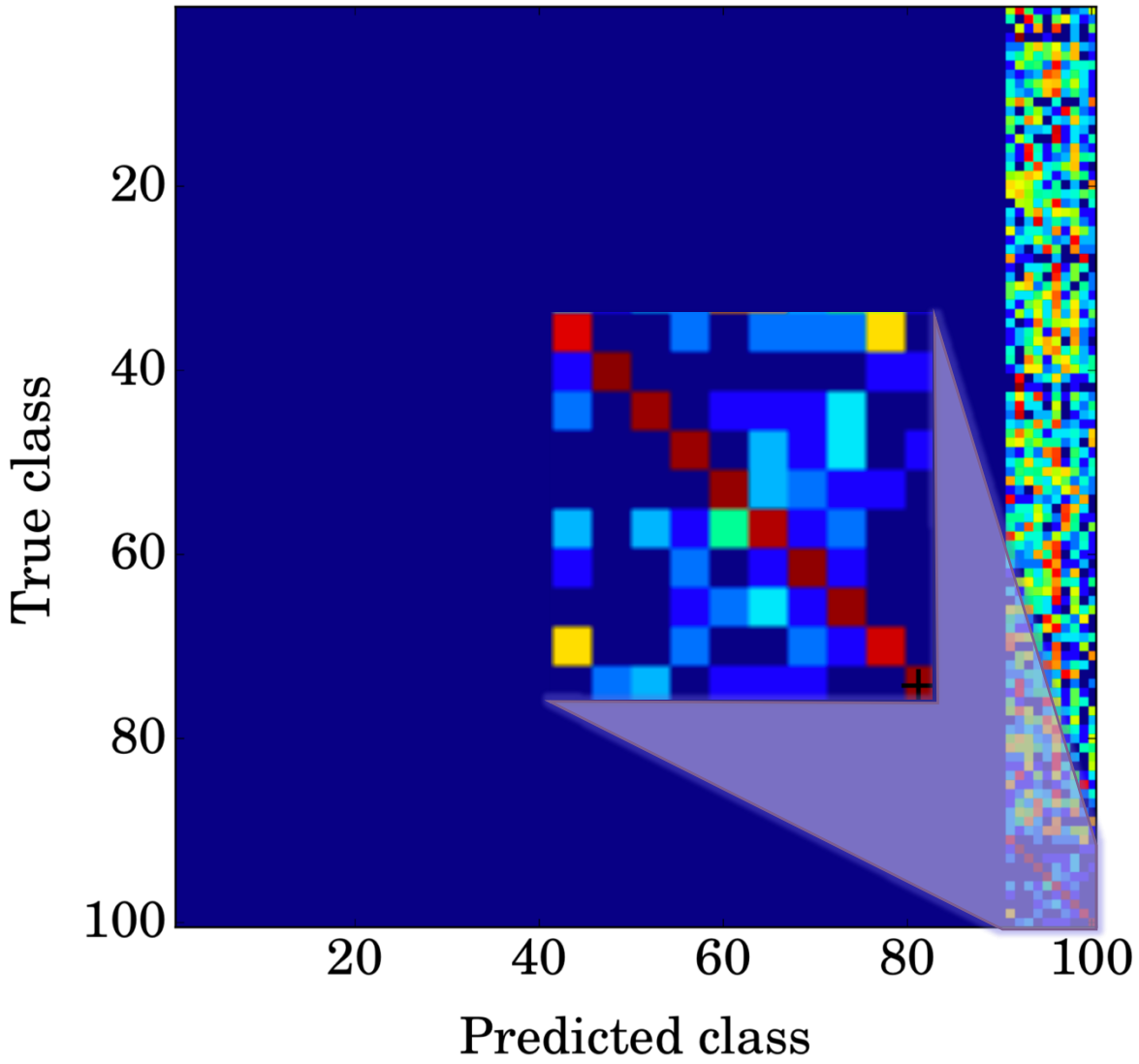


Is one level of granularity sufficient in real-world applications?

Class-incremental continual learning



Catastrophic forgetting

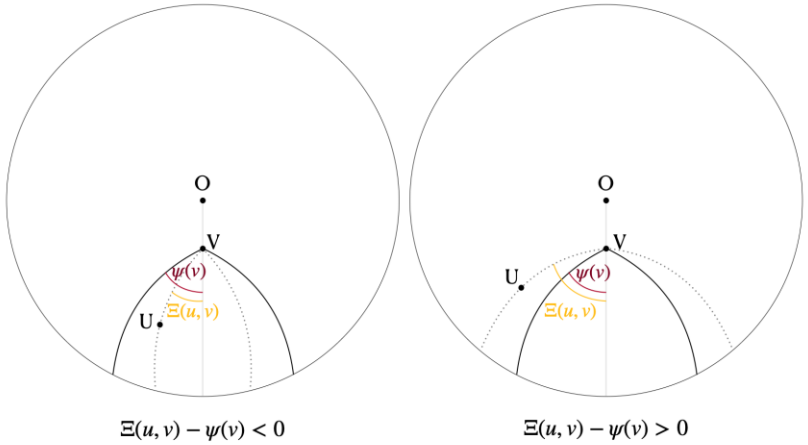


Embedding in hyperbolic space

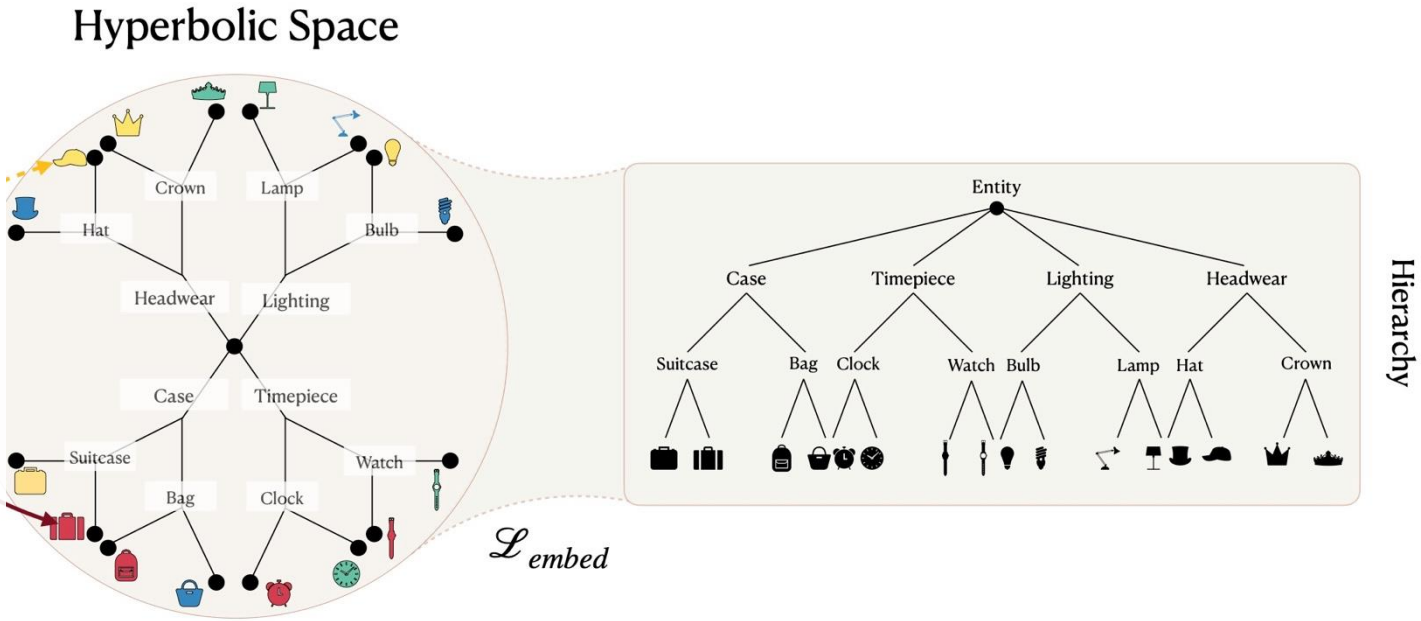
$$\mathcal{L}_{Poincaré} = \sum_{(u,v)\in\mathcal{R}} \log \frac{e^{-d_B(u,v)}}{\sum_{v'\in\mathcal{N}(u)} e^{-d_B(u,v')}}.$$

$$\mathcal{L}_{Entailment} = \sum_{(u,v)\in\mathcal{R}} E(u,v) + \sum_{(u',v')\in\mathcal{N}} max(0, \gamma - E(u',v')).$$

$$\mathcal{L}_S(P) = \vec{1}^T(\bar{P}\bar{P}^T - I)\vec{1},$$



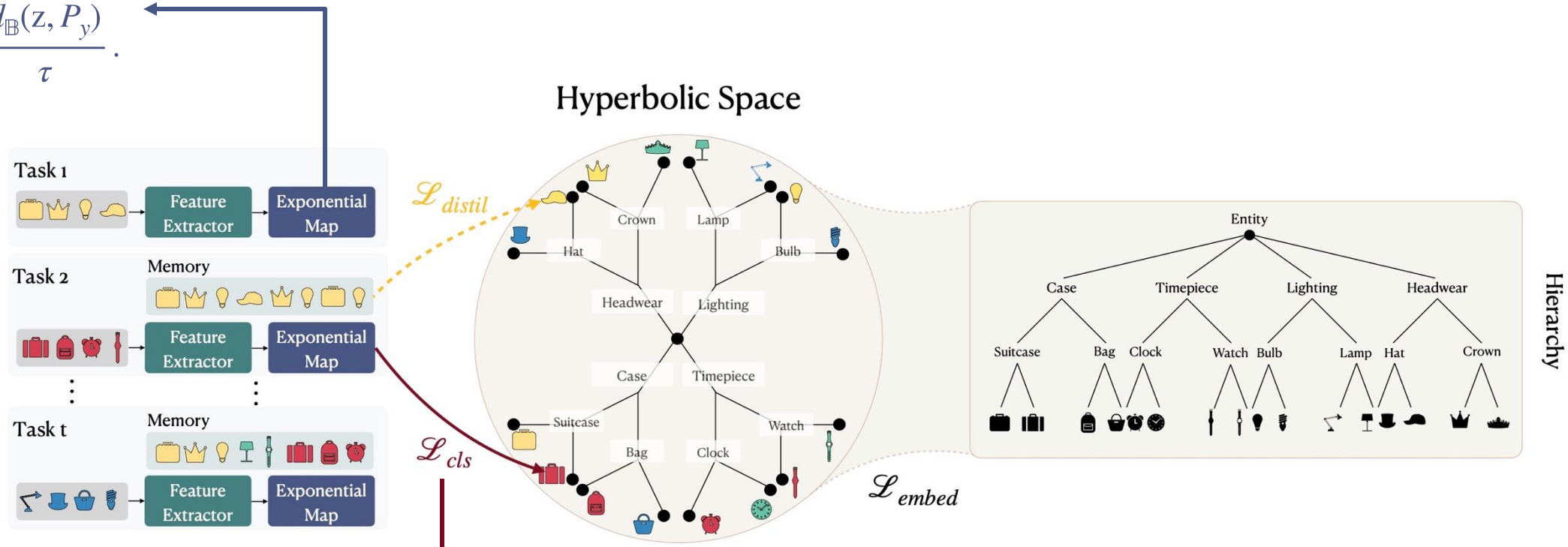
$\Xi(u,v) - \psi(v) < 0$ $\quad$ $\Xi(u,v) - \psi(v) > 0$



Classification loss

$$z_i^t = \exp_0(\varphi(x_i; \theta^t)) .$$

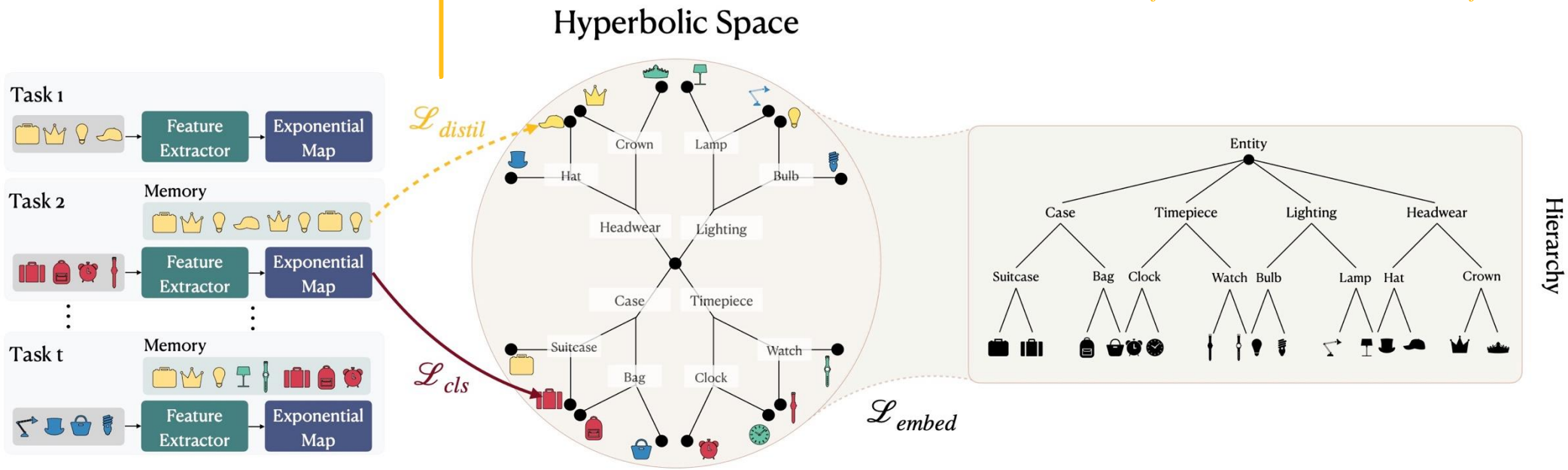
$$h(z, y) = -\frac{d_{\mathbb{B}}(z, P_y)}{\tau} .$$



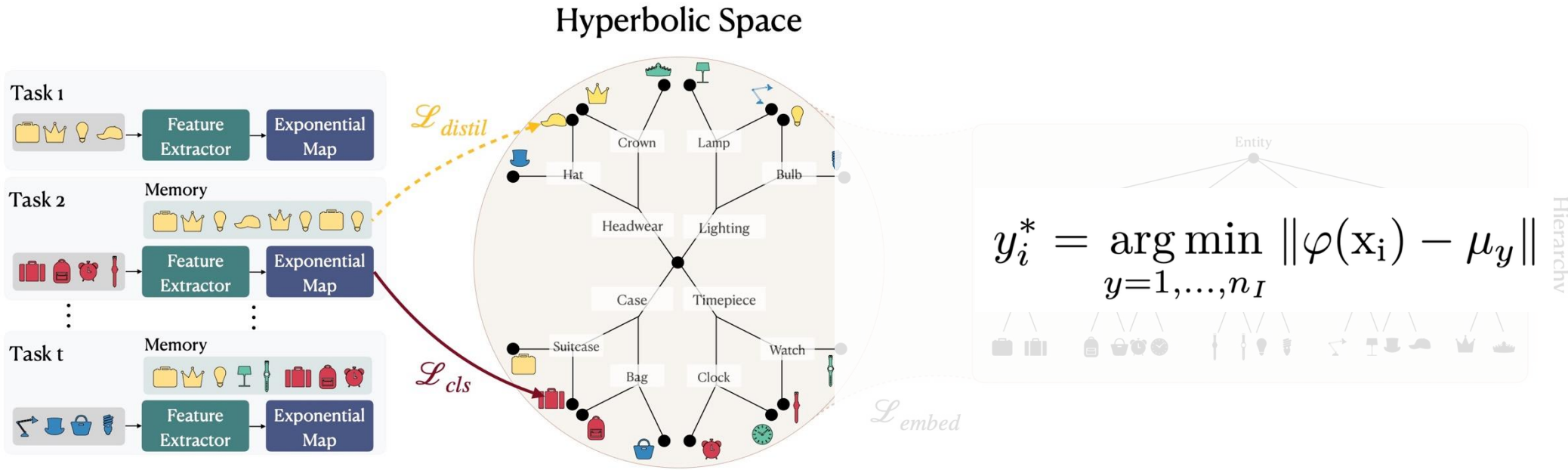
$$\mathcal{L}_{cls} = -\frac{1}{|\mathcal{D}^t|} \sum_{(x_i, y_i) \in \mathcal{D}^t} \sum_{i=1}^{|C|} y_i \cdot \log\left(\frac{e^{h(z_i^t, y_i)}}{\sum_{j=1}^{|C|} e^{h(z_i^t, y_j)}}\right) .$$

Distillation loss

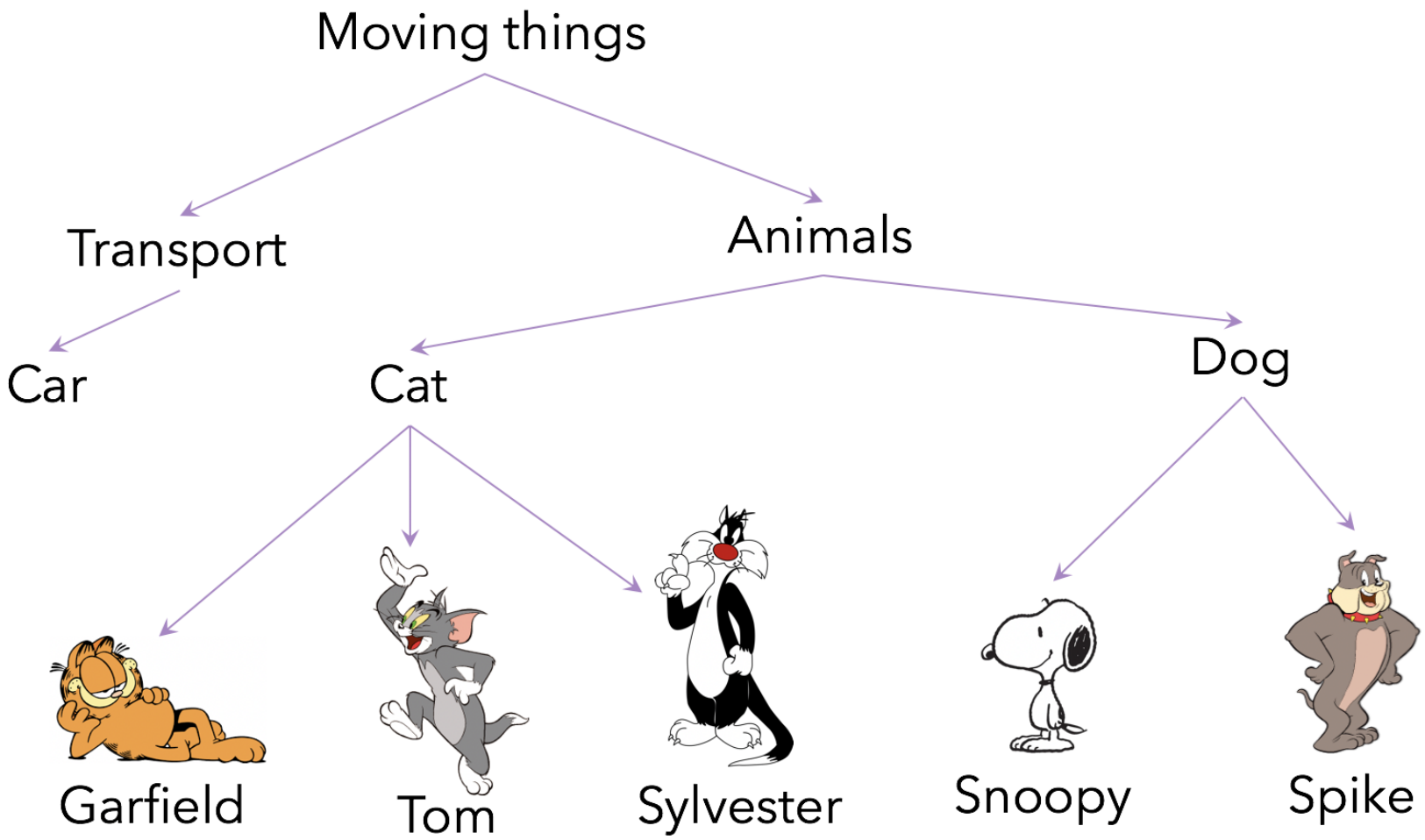
$$\mathcal{L}_{distil} = -\frac{1}{|\mathcal{D}^{b<t}|} \sum_{(x_i,y_i)\in\mathcal{D}^{b<t}} \sum_{i=1}^{|C^{b<t}|} \left(\frac{e^{h(z_i^{t-1},y_i)}}{\sum_{j=1}^{|C^{b<t}|} e^{h(z_i^{t-1},y_j)}}\right) \cdot \log\left(\frac{e^{h(z_i^{t},y_i)}}{\sum_{j=1}^{|C^{b<t}|} e^{h(z_i^{t},y_j)}}\right).$$



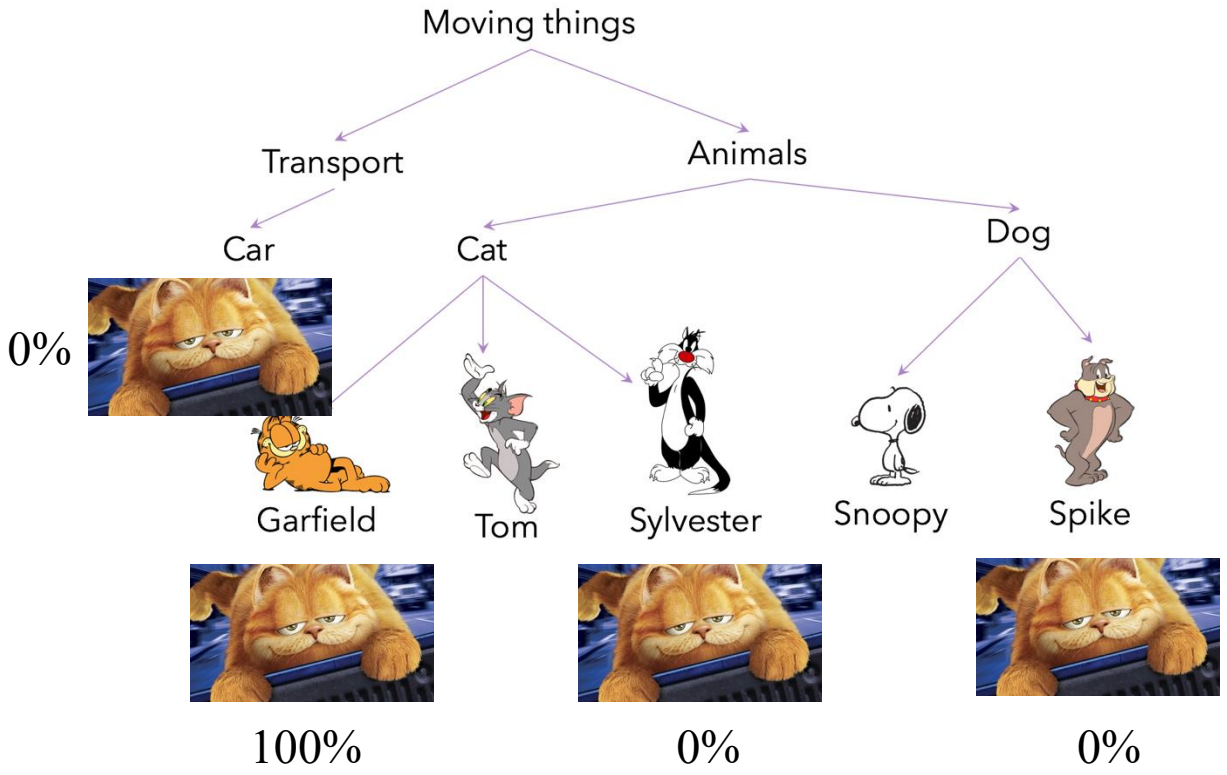
Inference: nearest mean of exemplars



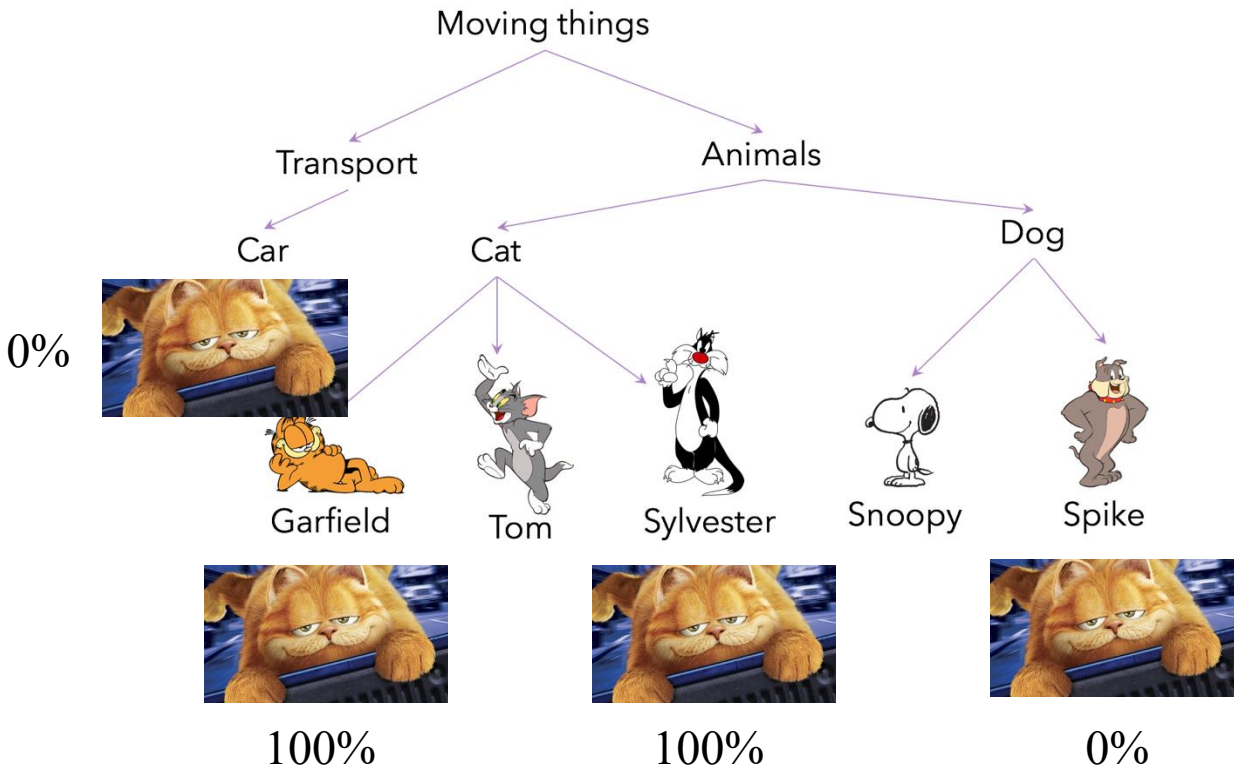
Evaluation metrics



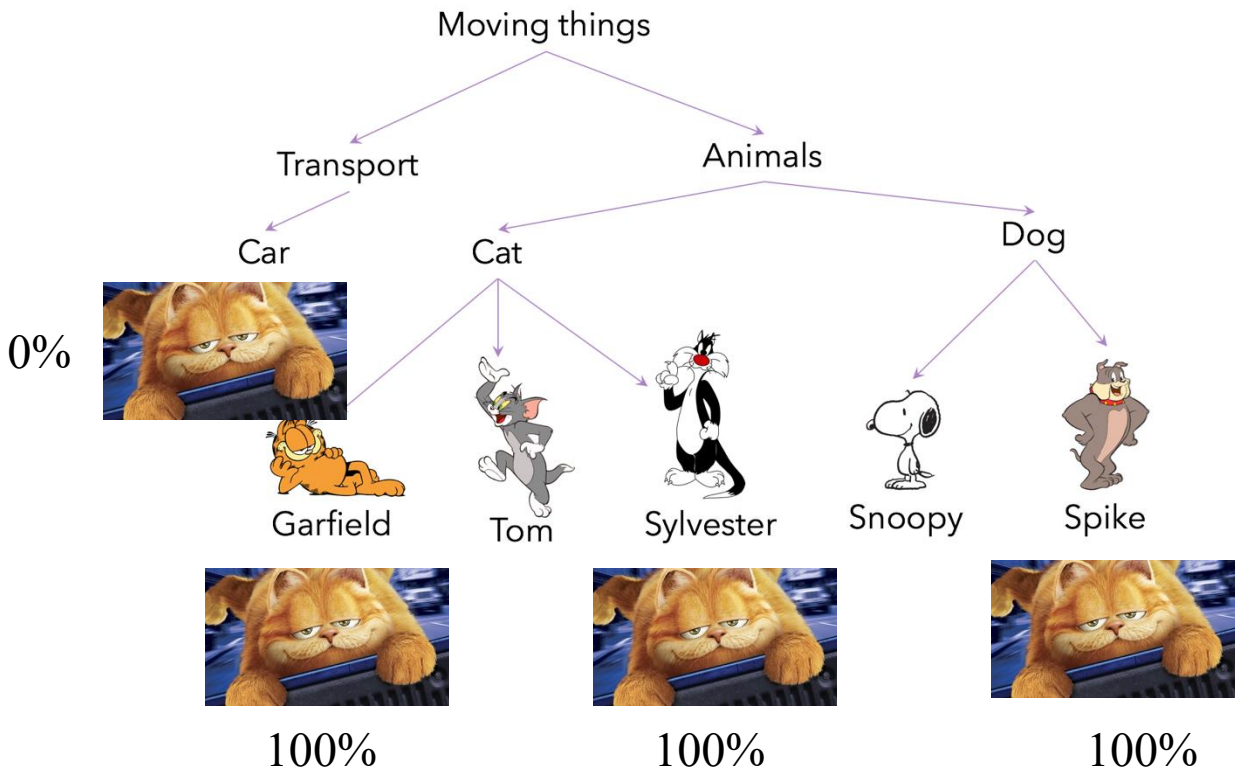
Accuracy: Cost of misclassification?



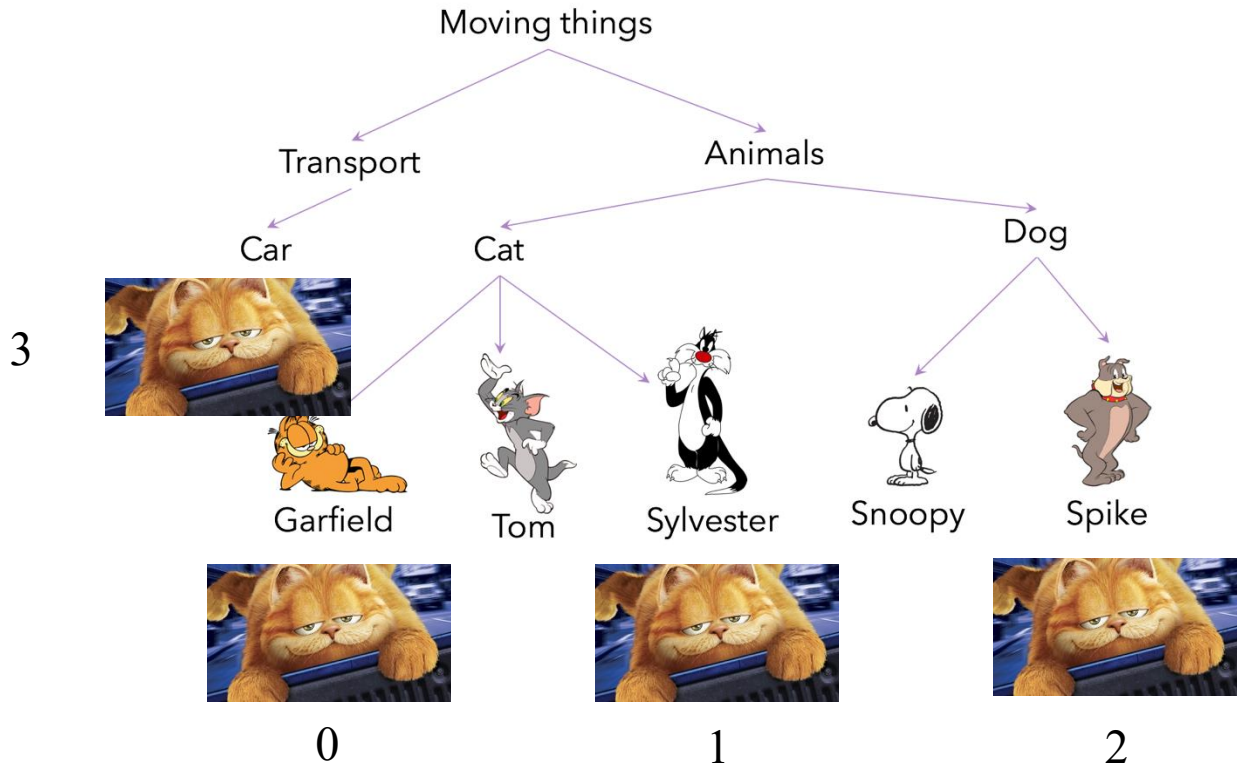
1-hop or sibling accuracy



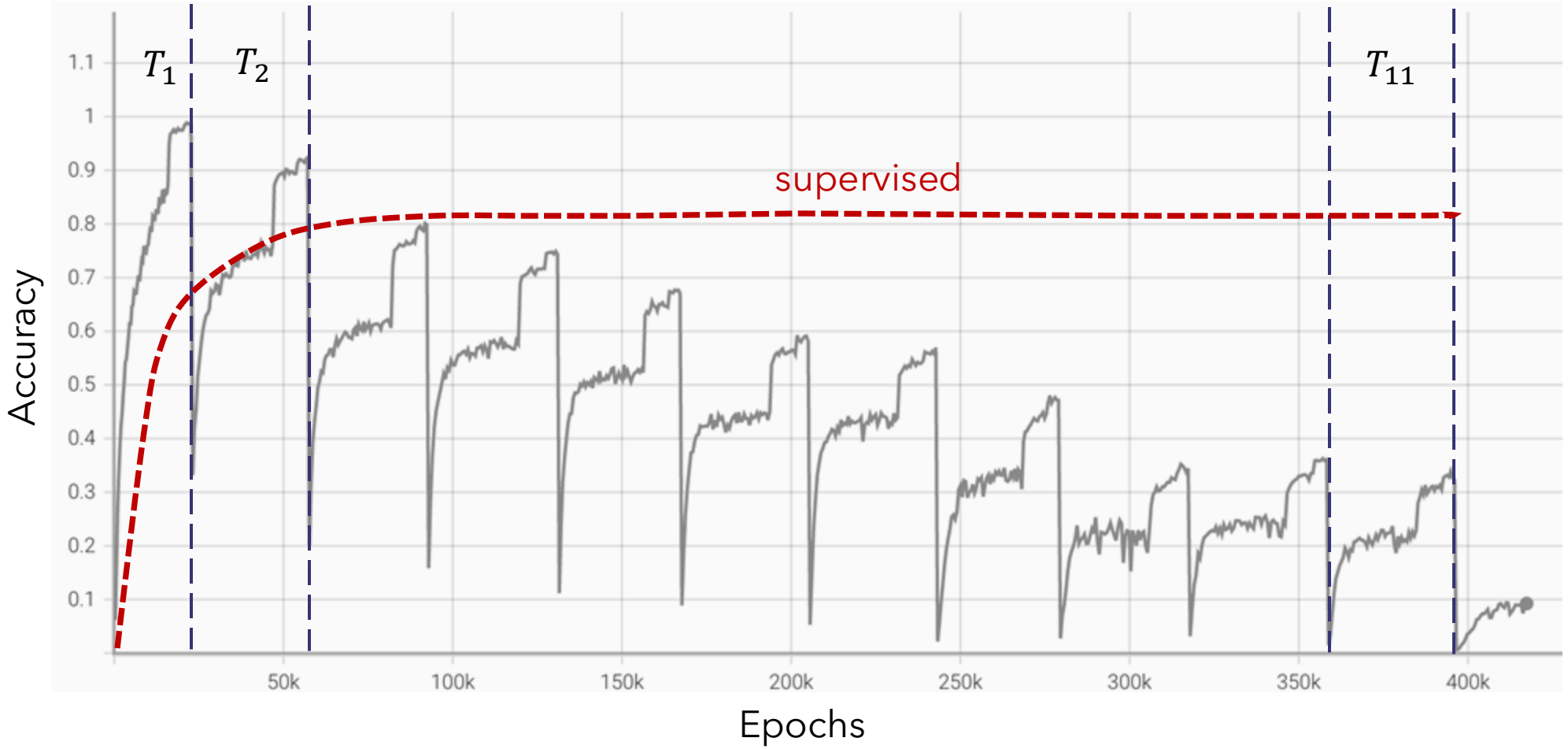
2-hop or cousin accuracy



LCA: lowest common ancestor



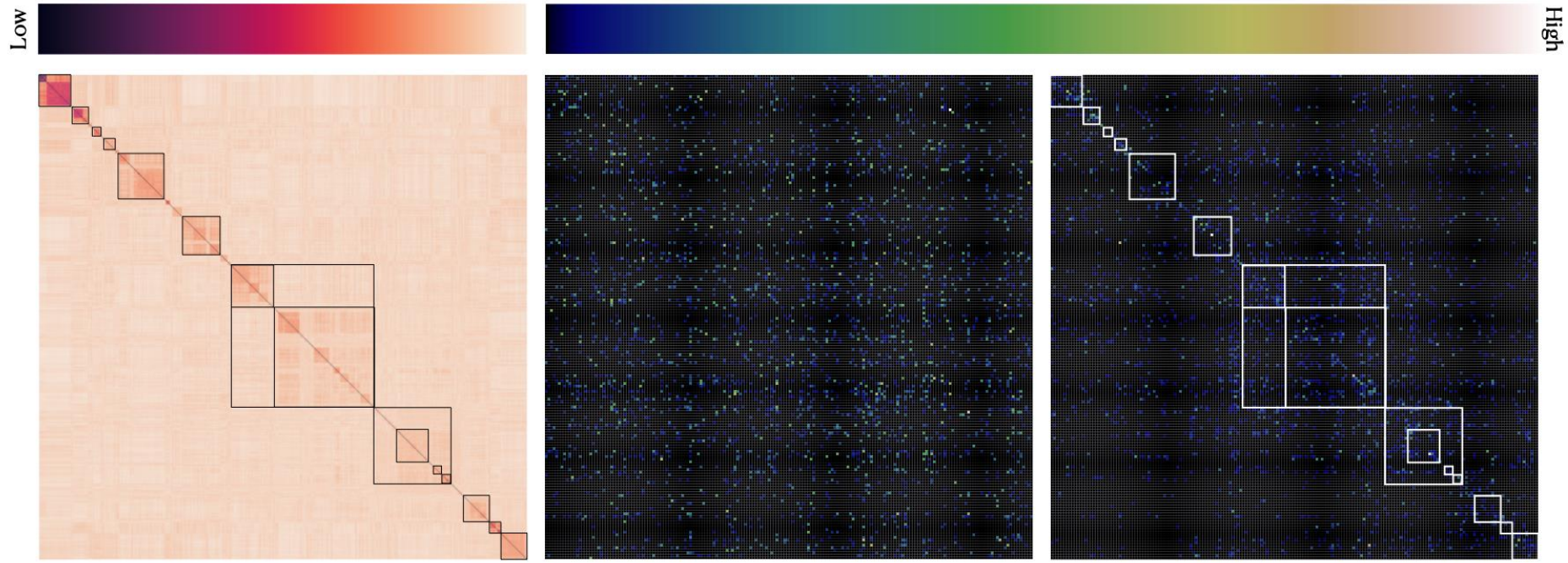
Forgetting


T_1
T_2
T_{11}
supervised
Accuracy
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
50k
100k
150k
200k
250k
300k
350k
400k
Epochs

Results

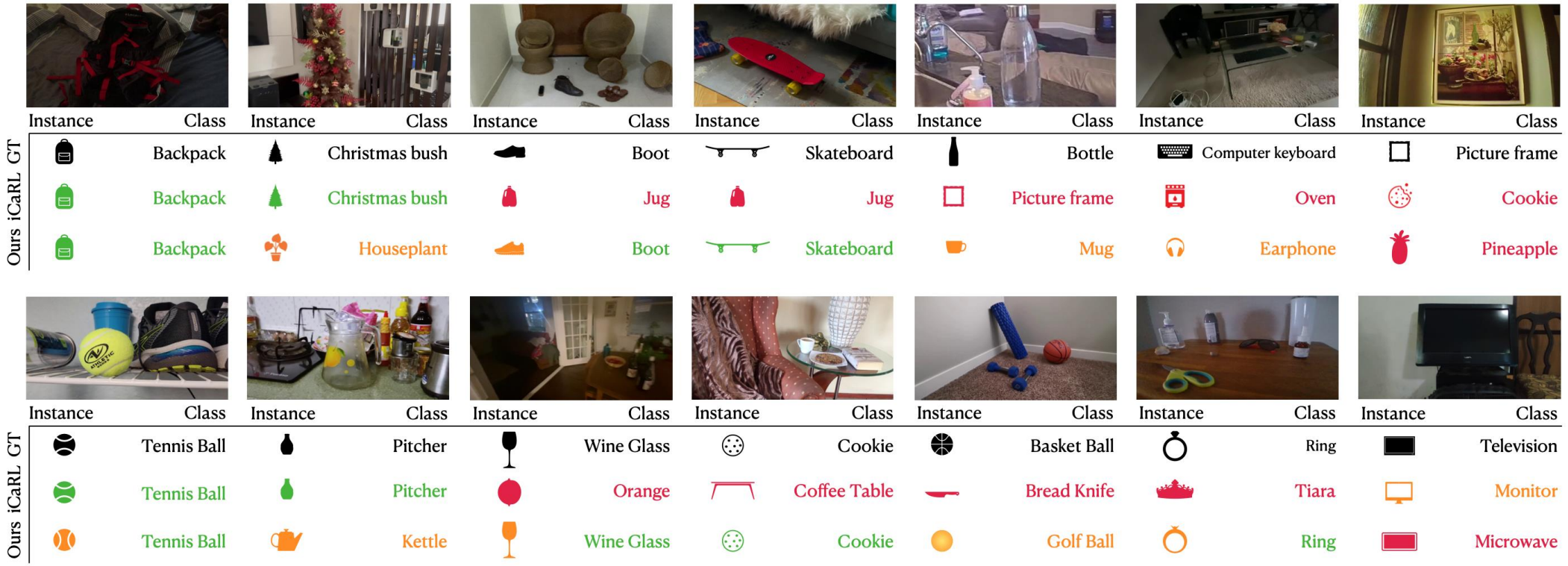
	Accuracy ↑				
	Instance	Class	Super-class	**LCA ↓**	**Forgetting ↓**
Naive	0.44 (6.06)	1.07 (15.55)	1.58 (18.24)	5.74 (4.70)	10.38 (93.03)
EWC [14]	0.22 (6.07)	0.66 (15.77)	1.42 (18.37)	5.26 (4.74)	10.38 (92.83)
iCaRL [20]	20.05 (81.63)	21.39 (87.02)	22.24 (87.81)	5.45 (3.89)	10.57 (3.77)
CoPE [4]	3.14 (37.15)	4.45 (49.89)	4.91 (52.51)	5.63 (4.22)	25.38 (32.40)
GDumb [19]	0.50 (61.51)	1.27 (71.00)	1.88 (72.82)	5.50 (4.02)	5.87 (18.22)
DER [25]	19.59 (80.00)	20.59 (84.40)	21.14 (85.20)	5.61 (4.28)	35.39 (10.59)
HyperCLIC	**41.76** (84.81)	**45.91** (91.94)	**48.04** (92.67)	**4.93** (2.99)	**4.17** (7.08)

Severity of model's mistakes



Class-level predictions still follow the hierarchical structure even when instances are misclassified

Severity of model's mistakes

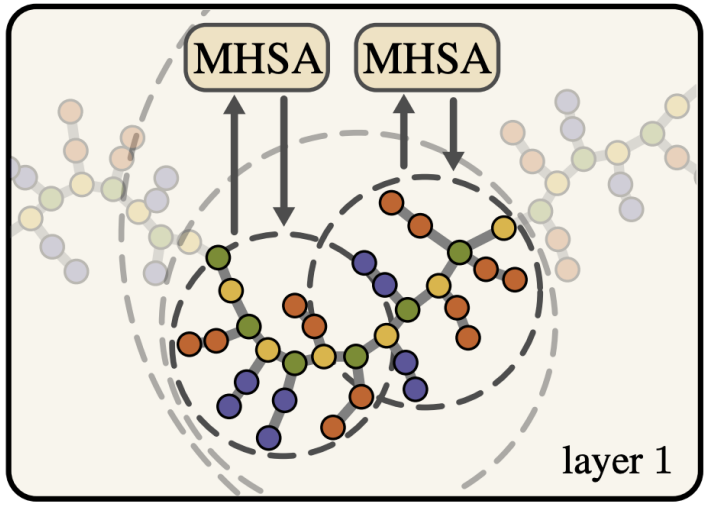


The severity of hierarchical errors reduces by incorporating hierarchical information.

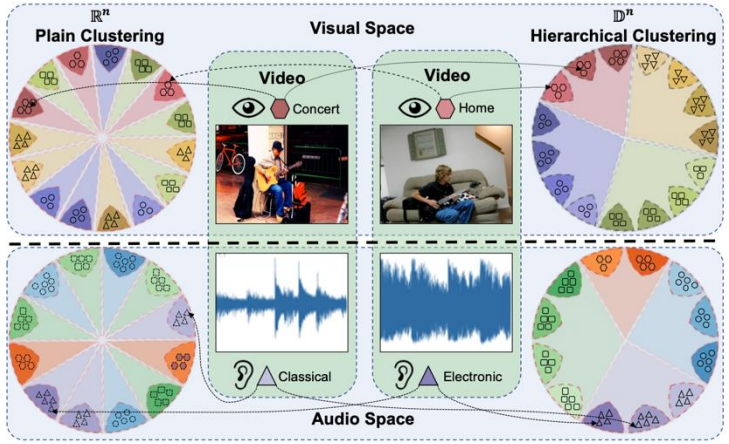
University of Amsterdam
Faculty of Science

Designing Hierarchies for Optimal Hyperbolic Embedding

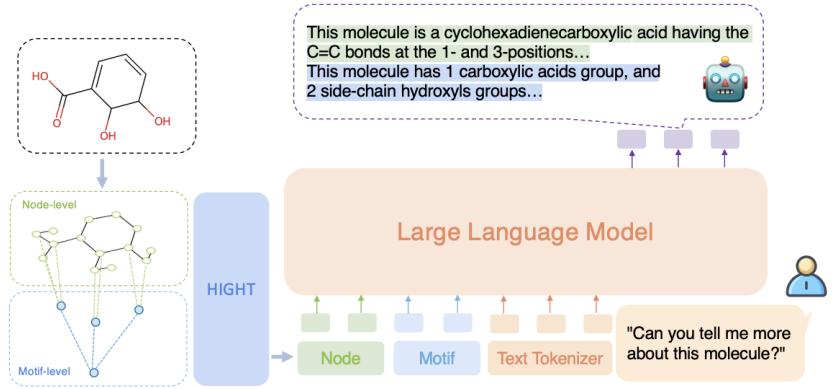
Embedded Hierarchies



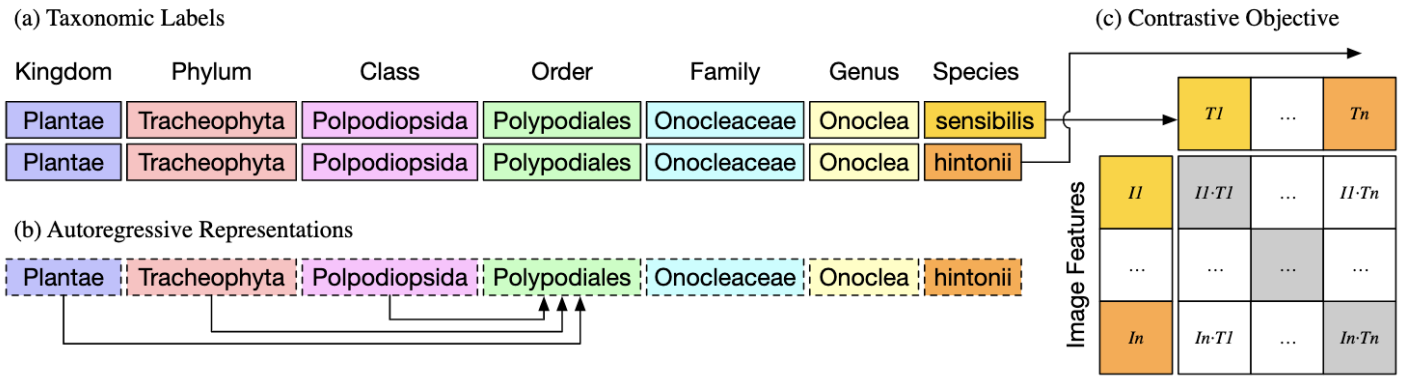
Molecular dynamics



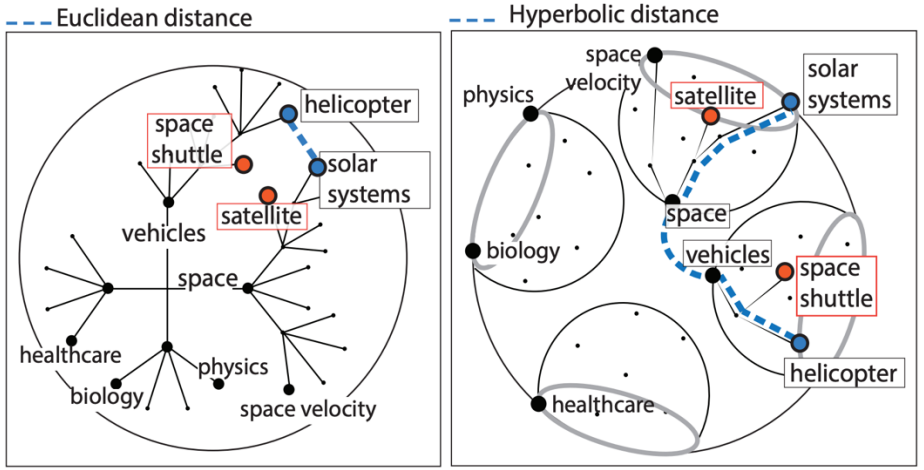
Multimodal audio-visual Clustering



Molecular Description Generation



Fine-grained biology classification tasks

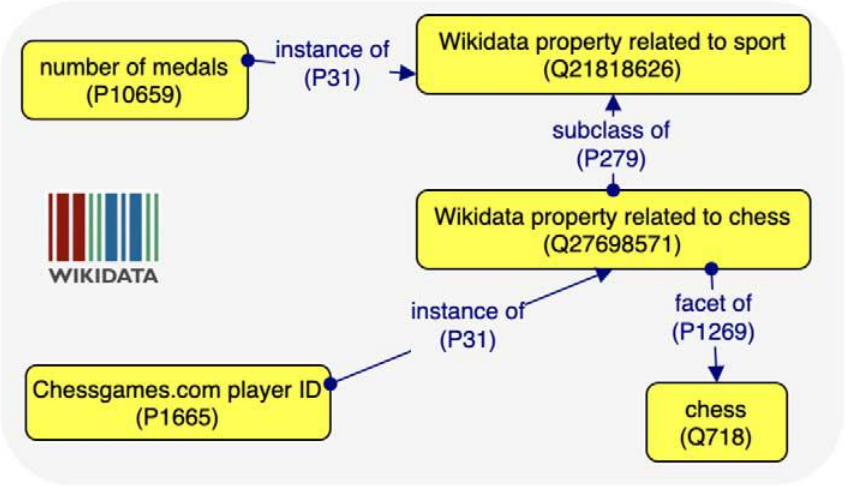


Hierarchical Topic Models

Knowledge Graphs and Ontologies

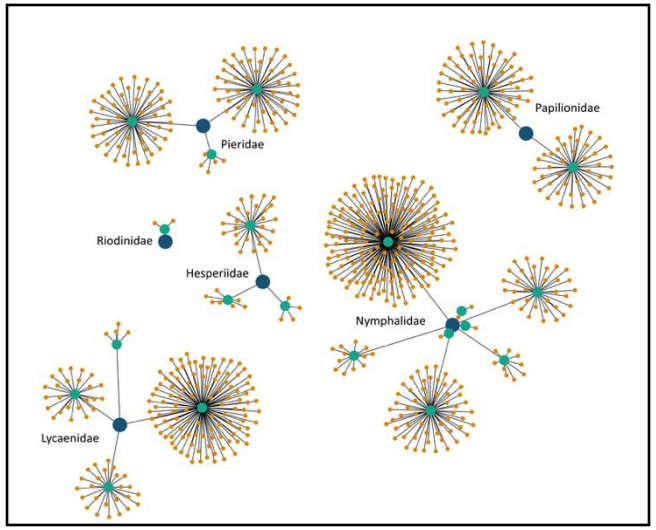
- Thing -
  - Action +
  - BioChemEntity +
  - CreativeWork +
  - Event +
  - Intangible +
  - MedicalEntity +
  - Organization +
  - Person +
  - Place -
    - Accommodation +
    - AdministrativeArea +
    - CivicStructure +
    - Landform +
    - LandmarksOrHistoricalBuildings
    - LocalBusiness
    - Residence -
      - ApartmentComplex
      - GatedResidenceCommunity
    - TouristAttraction
    - TouristDestination
  - Product +

[schema.org]



[Carriero et al. Semantic Web 2024]

WordNet
A Lexical Database for English



[The ETHEC entomological hierarchy]

Most methods use the hierarchies **as is**, however, we have **control** over the hierarchy.

Flipped perspective:
How can hierarchies be designed to suite ML tasks?

[Tree of life by Gustav Klimt]

Controlled Experiment Setup

Hyperbolic Embedding Methods, Hierarchies, Evaluation Metrics

Hyperbolic Embedding Methods

Gradient-based

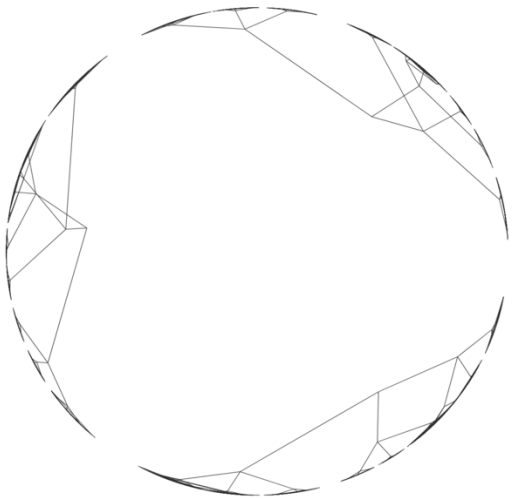
Construction-based

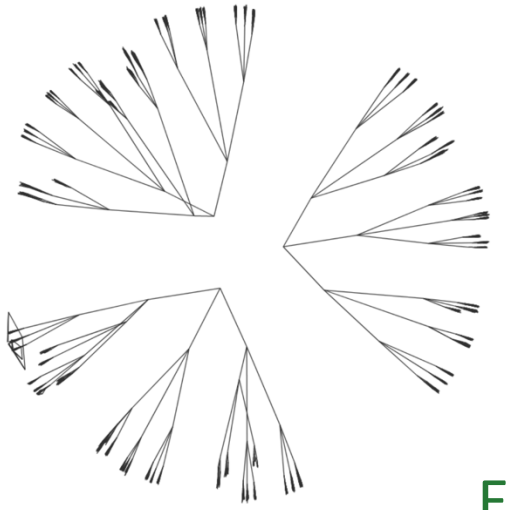
Poincaré

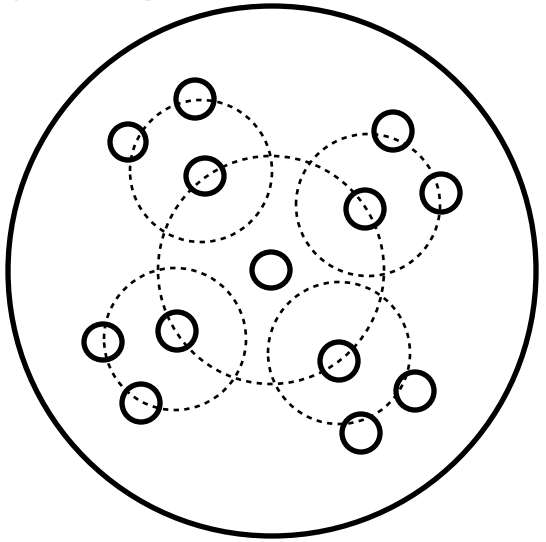
Entailment

Precomputing

Hadamard







Enforce structure

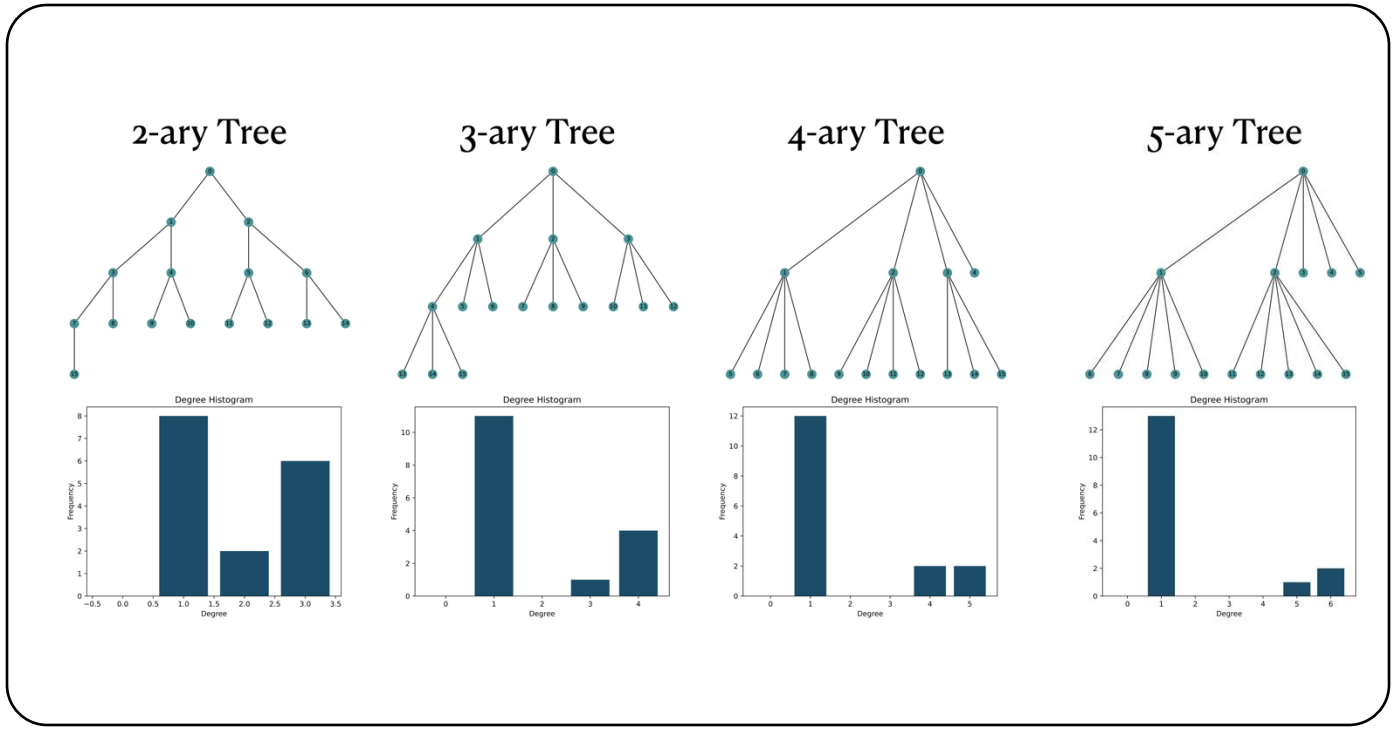
[Ganea et al. ICML 2018]

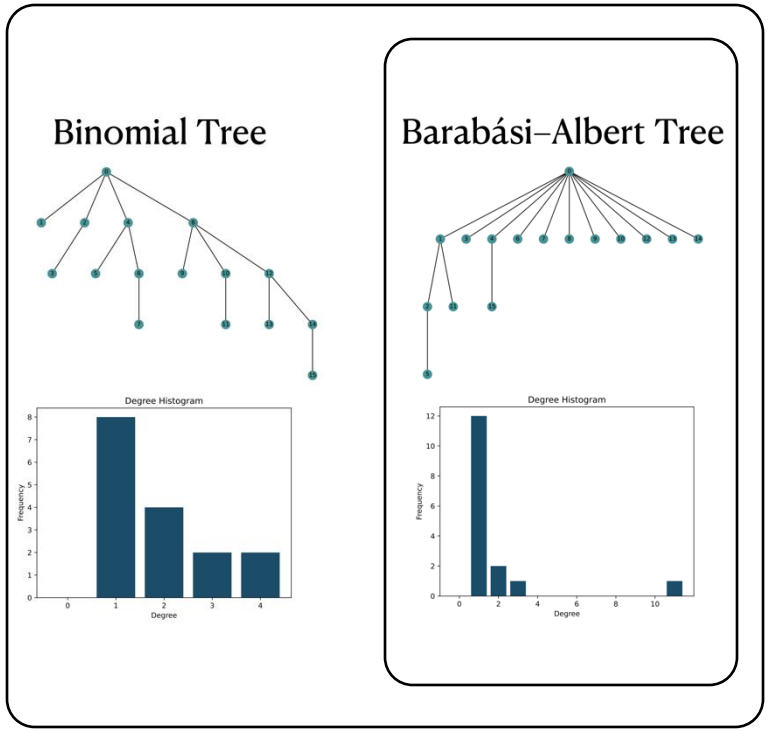
- ⊕ General-purpose
- ⊖ Quality not guarantied

- ⊕ High-quality embeddings
- ⊕ Fast
- ⊕ Preserve original hierarchy structure
- ⊖ Arbitrary-precision arithmetic

Diverse Hierarchies

Long-tailed

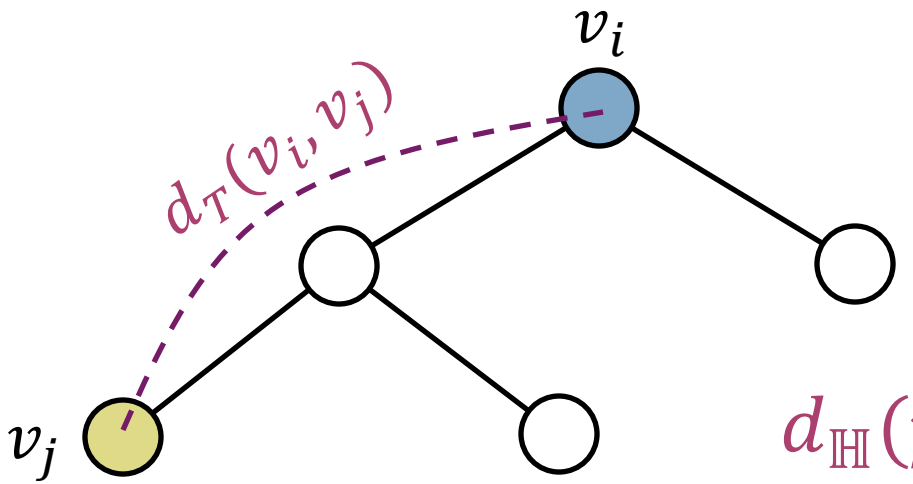




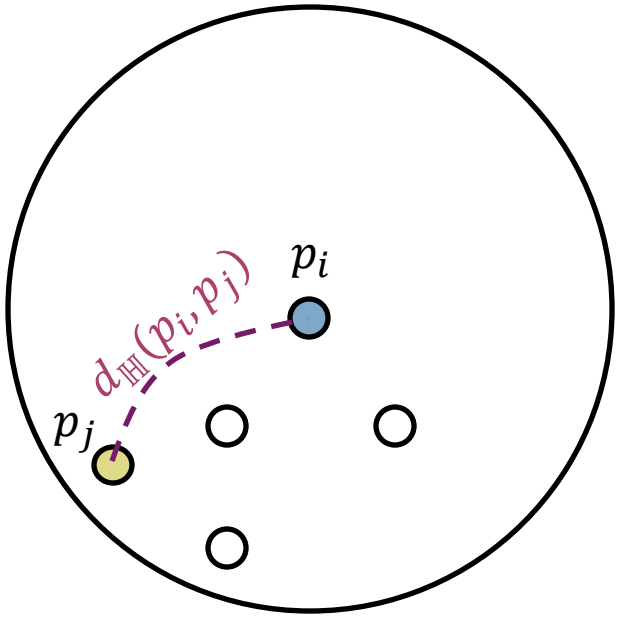
Balanced

Imbalanced

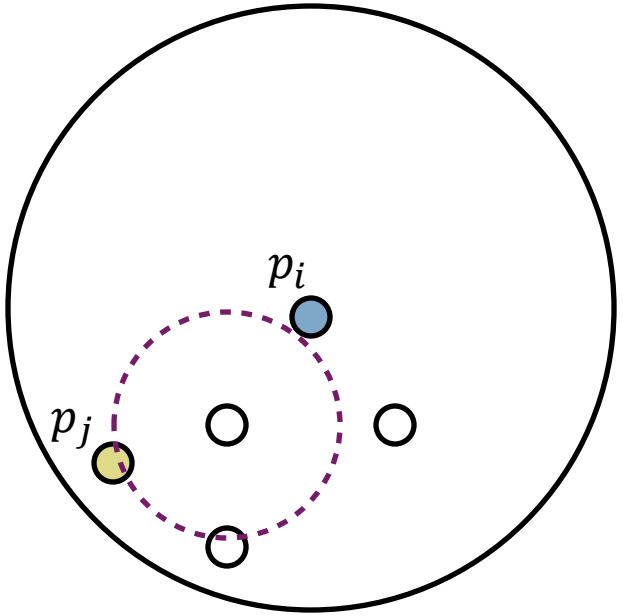
Evaluation Metrics



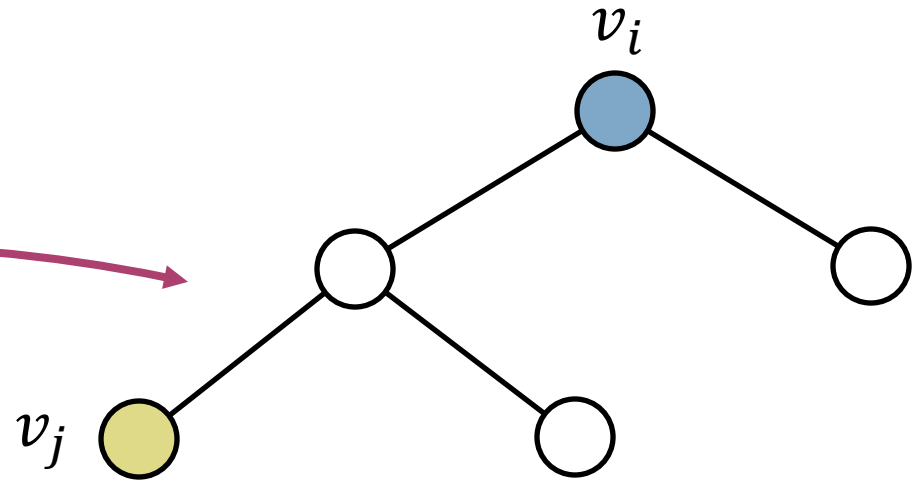
$$d_{\mathbb{H}}(p_i, p_j) \propto d_T(v_i, v_j)$$



$\downarrow D_{avg} \downarrow D_{wc}$
Distortion



Reconstruct

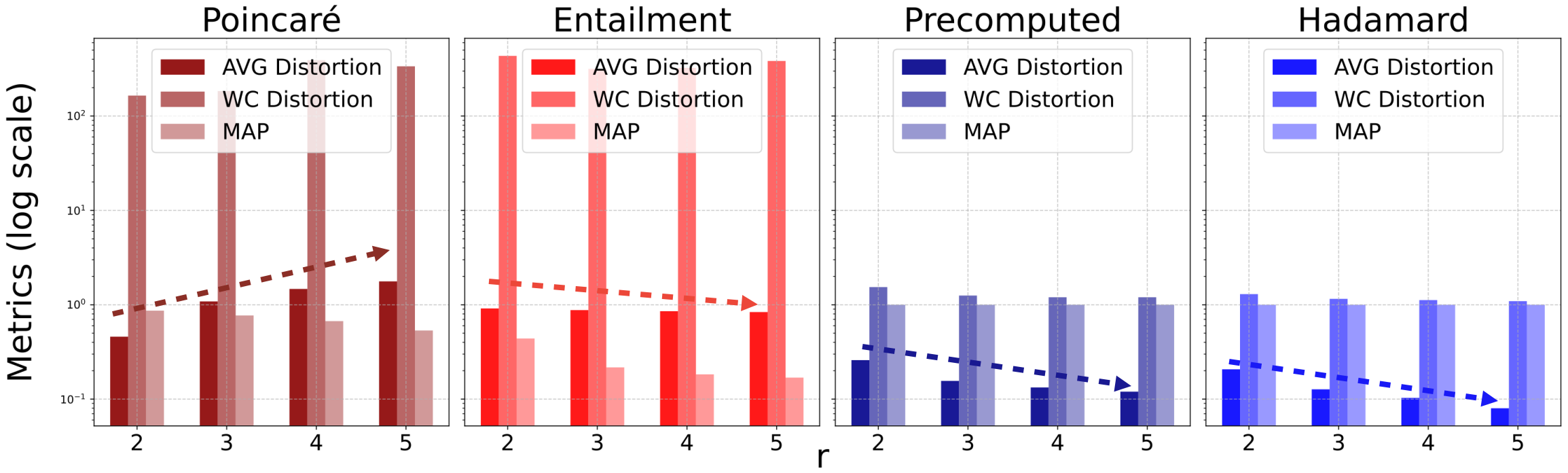


$\uparrow MAP$
Reconstruction

Results

Depth or Width?

Cheatsheet:
$\downarrow D_{avg} \downarrow D_{wc} \uparrow MAP$



I. All methods except for Poincaré: **wide and shallow** hierarchies have lower distortion.
II. **Construction-based** methods paired with **wide** hierarchies achieve **optimal** embeddings.

What is the impact of more nodes on embedding quality?

Cheatsheet:
$\downarrow D_{avg} \downarrow D_{wc} \uparrow MAP$

	Gradient-based						Construction-based					
	Poincaré			Entailment			Precomputed			Hadamard		
	256	512	1024	256	512	1024	256	512	1024	256	512	1024
Balanced												
2-ary	0.880	0.459	0.229	0.816	0.914	0.960	0.220	0.259	0.300	0.176	0.207	0.240
3-ary	1.439	1.085	0.752	0.742	0.878	0.940	0.124	0.156	0.160	0.102	0.127	0.130
4-ary	2.129	1.471	1.092	0.695	0.855	0.928	0.102	0.133	0.137	0.079	0.103	0.105
5-ary	2.472	1.770	1.385	0.657	0.837	0.919	0.115	0.120	0.156	0.078	0.080	0.103
Imbalanced												
Binomial	1.736	1.439	0.988	0.717	0.863	0.932	0.207	0.249	0.298	0.161	0.186	0.211
BA	3.444	2.791	2.206	0.595	0.802	0.903	0.108	0.140	0.178	-	-	-

A strong **increase in semantic complexity** has **minimal impact** on embedding quality.

Balanced or Imbalanced?

Cheatsheet:
$\downarrow D_{avg} \downarrow D_{wc} \uparrow MAP$

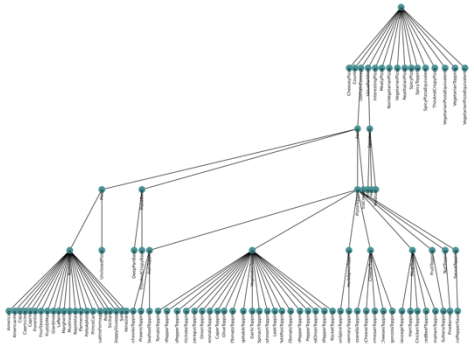
best and worst

	Gradient-based						Construction-based					
	Poincaré			Entailment			Precomputed			Hadamard		
	D_{avg}	D_{wc}	MAP	D_{avg}	D_{wc}	MAP	D_{avg}	D_{wc}	MAP	D_{avg}	D_{wc}	MAP
Balanced												
2-ary	0.459	164.777	0.866	0.914	434.177	0.439	0.259	1.539	1	0.207	1.297	1
3-ary	1.085	183.974	0.770	0.878	316.338	0.217	0.156	1.252	1	0.127	1.155	1
4-ary	1.471	390.397	0.671	0.855	323.967	0.183	0.133	1.201	1	0.103	1.121	1
5-ary	1.770	336.711	0.534	0.837	383.626	0.169	0.120	1.201	1	0.080	1.092	1
Imbalanced												
Binomial	1.439	69.530	0.171	0.863	224.731	0.304	0.249	1.542	1	0.186	1.257	1
BA	2.791	3607.95	0.020	0.802	731.914	0.231	0.140	1.329	1	-	-	-

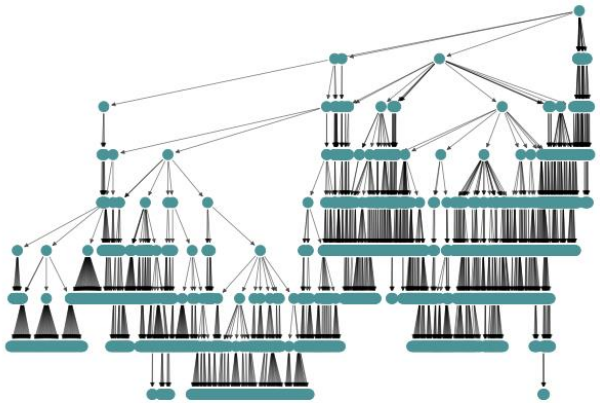
Better to have a wide imbalanced tree than a deep balanced one!

Case study: The Pizza and ImageNet ontologies

Pizza reorganized



ImageNet reorganized



Cheatsheet:
$\downarrow D_{avg} \downarrow D_{wc} \uparrow MAP$

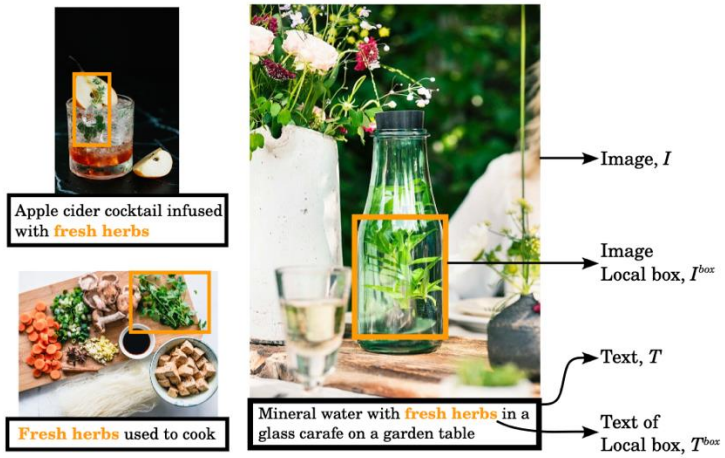
	Gradient-based						Construction-based					
	Poincaré			Entailment			Precomputed			Hadamard		
	D_{avg}	D_{wc}	MAP	D_{avg}	D_{wc}	MAP	D_{avg}	D_{wc}	MAP	D_{avg}	D_{wc}	MAP
Pizza												
Original	3.321	7066.671	0.059	-	-	-	-	-	-	-	-	-
+ single inheritance	3.387	10509.346	0.051	0.499	511.594	0.195	0.234	1.538	1	0.126	1.180	1
+ reorganized	3.422	9343.566	0.045	0.452	1454.972	0.164	0.167	1.329	1	**0.089**	**1.118**	**1**
ImageNet												
Original	0.809	3983.563	0.087	-	-	-	-	-	-	-	-	-
+ single inheritance	0.722	2745.952	0.220	0.961	2364.827	0.293	0.725	885.622	0.725	0.297	1.647	1
+ reorganized	1.008	12715.625	0.156	0.955	4096.000	0.164	0.507	2.698	1	**0.171**	**1.232**	**1**

I. Poincaré: only method to handle multiple inheritance
II. Hierarchy reorganization leads to better distortion and MAP

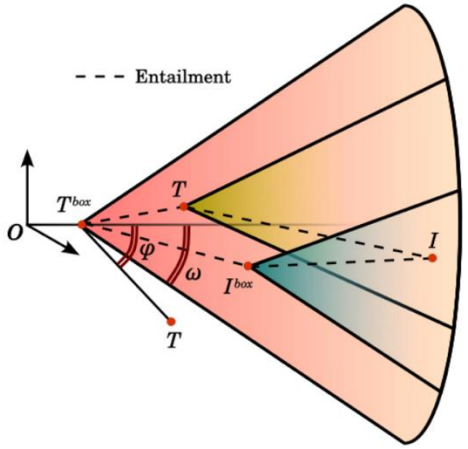
Recommendations

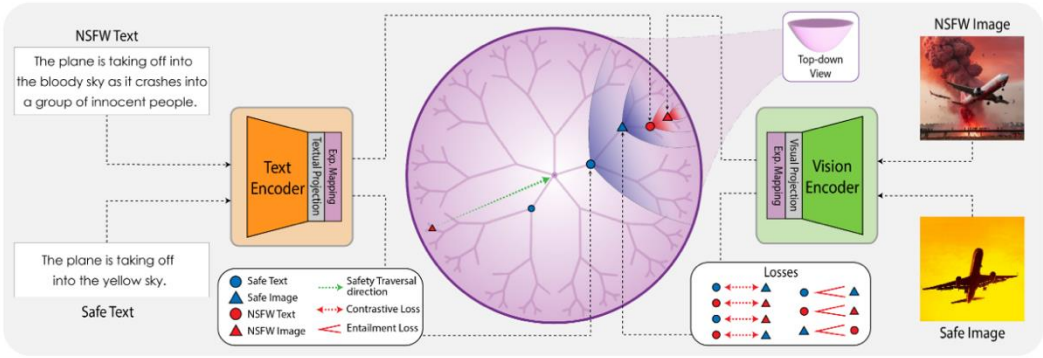
- ✓ Design hierarchies for width
- ✓ Do not worry about balance
- ✓ Hyperbolic embeddings can handle additional node complexity
- ✓ Avoid multiple inheritance; o/w limited to Poincaré method

Other applications of hyperbolic

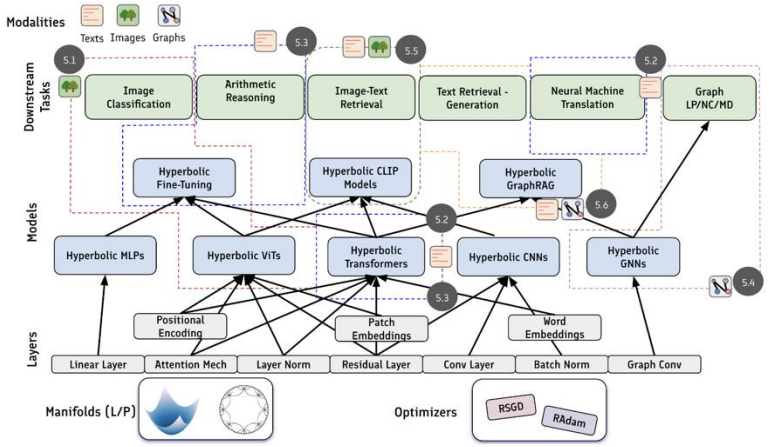


Vision-language models





Hyperbolic Safety-Aware Vision-Language



Hyperbolic ViT and CLIP