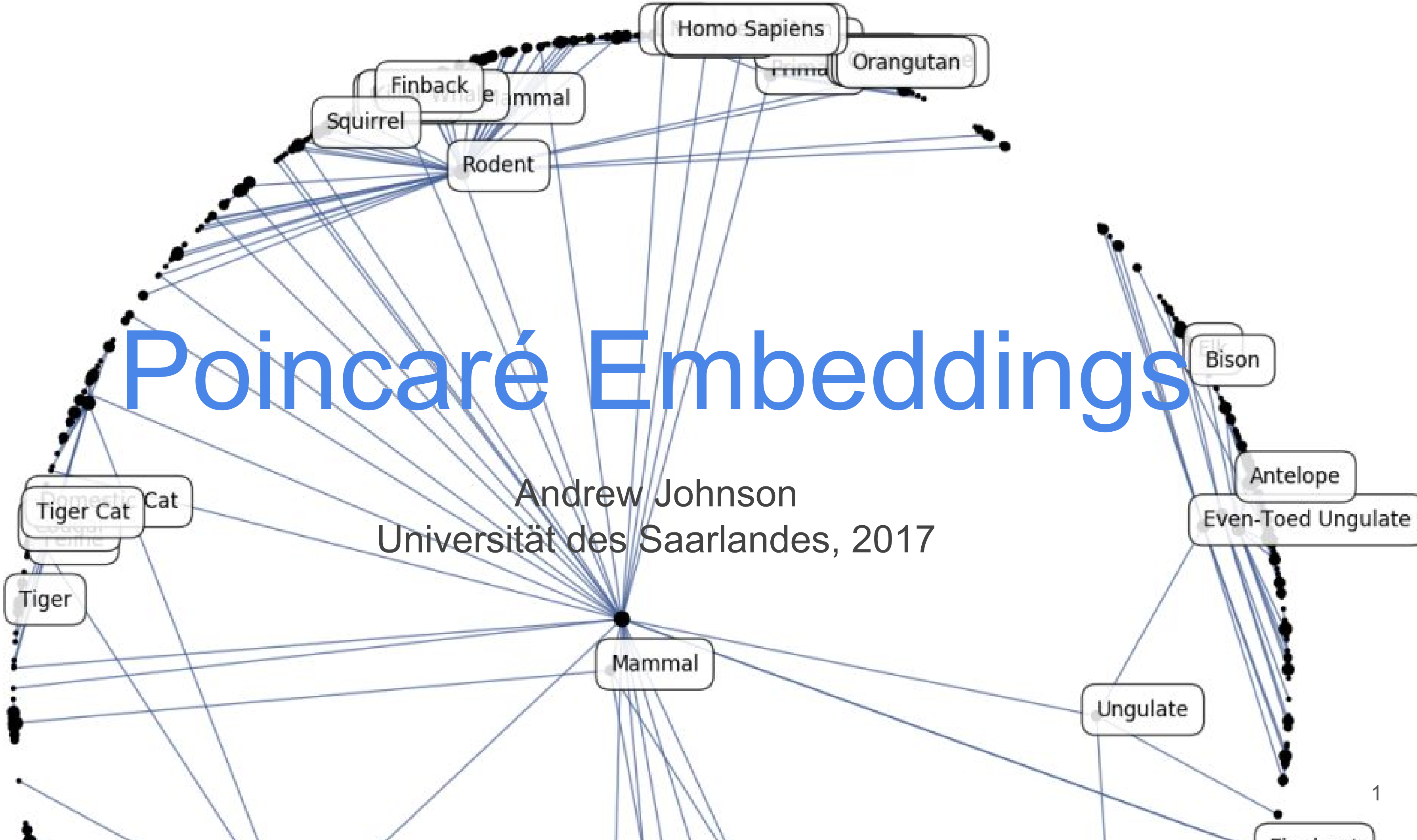
Poincaré Embeddings

Andrew Johnson
Universität des Saarlandes, 2017



From May 2017: https://arxiv.org/pdf/1705.08039.pdf

Poincaré Embeddings for Learning Hierarchical Representations

Maximilian Nickel
Facebook AI Research
maxn@fb.com

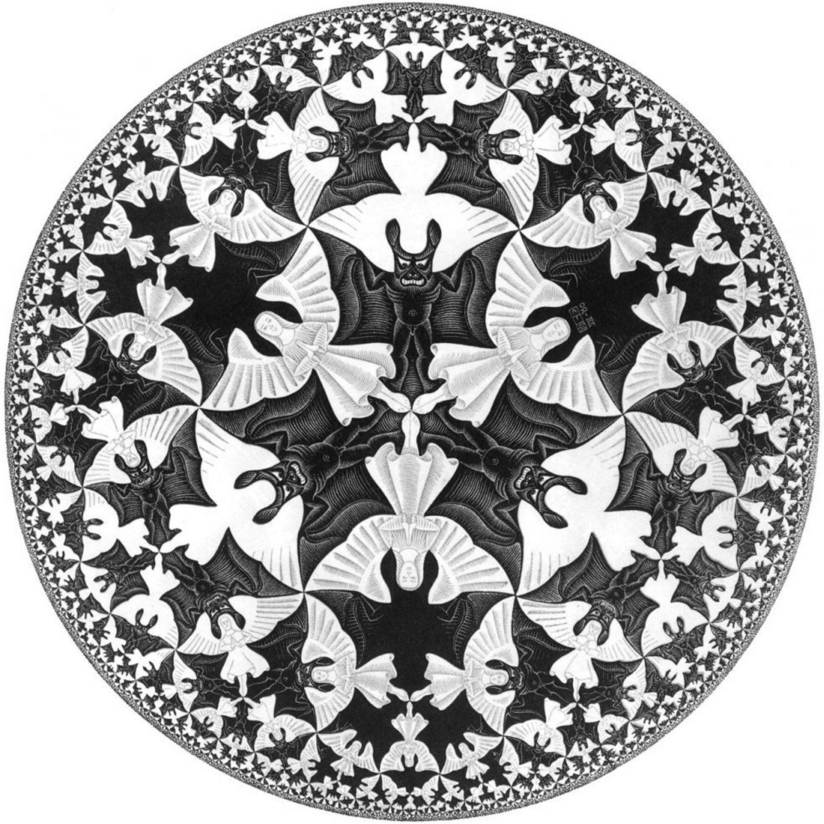
Douwe Kiela
Facebook AI Research
dkiela@fb.com

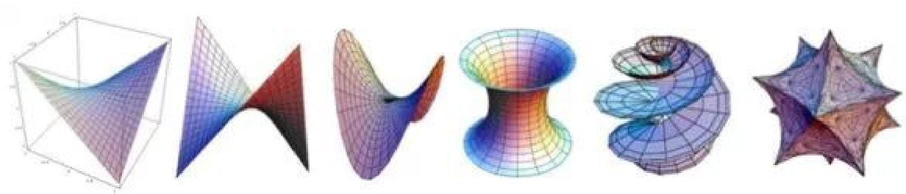
Abstract

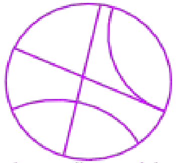
Representation learning has become an invaluable approach for learning from symbolic data such as text and graphs. However, while complex symbolic datasets often exhibit a latent hierarchical structure, state-of-the-art methods typically learn embeddings in Euclidean vector spaces, which do not account for this property. For this purpose, we introduce a new approach for learning hierarchical representations of symbolic data by embedding them into hyperbolic space – or more precisely into an n-dimensional Poincaré ball. Due to the underlying hyperbolic geometry, this allows us to learn parsimonious representations of symbolic data by simultaneously capturing hierarchy and similarity. We introduce an efficient algorithm to learn

(Very)

1. Basic Introduction to Hyperbolic Geometry
2. Mathematical Underpinnings
3. Results



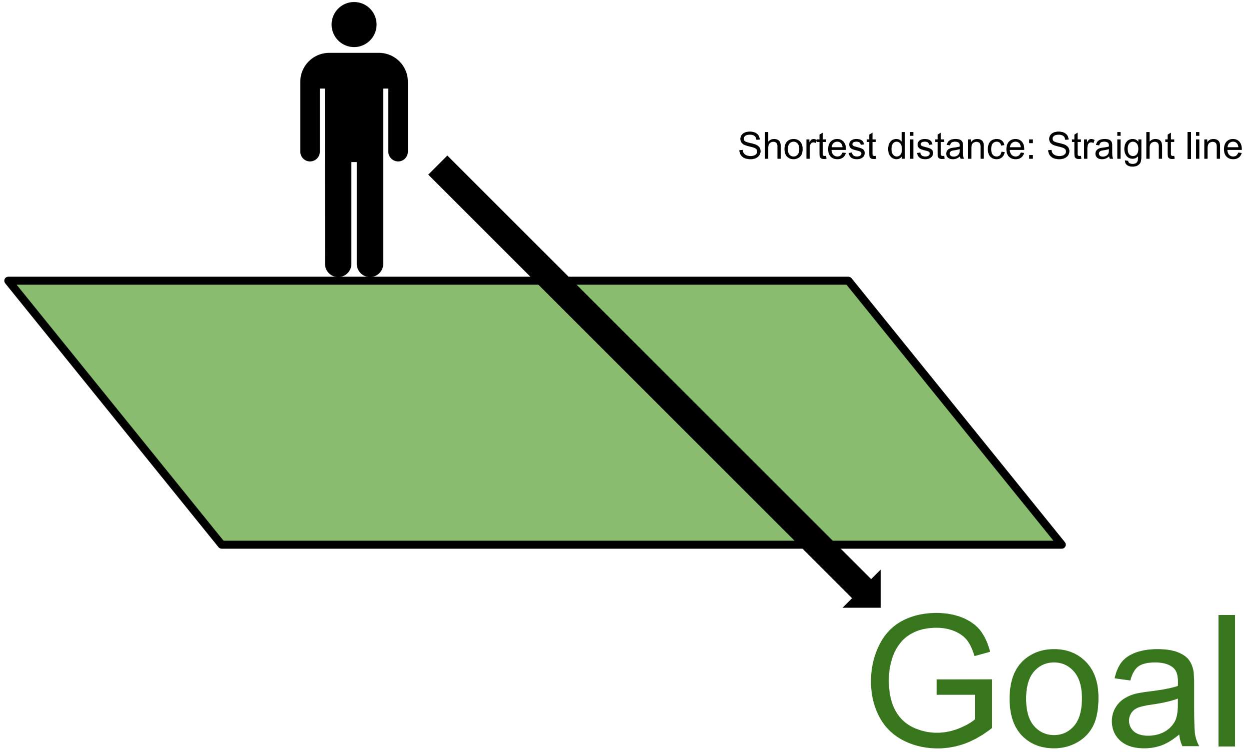


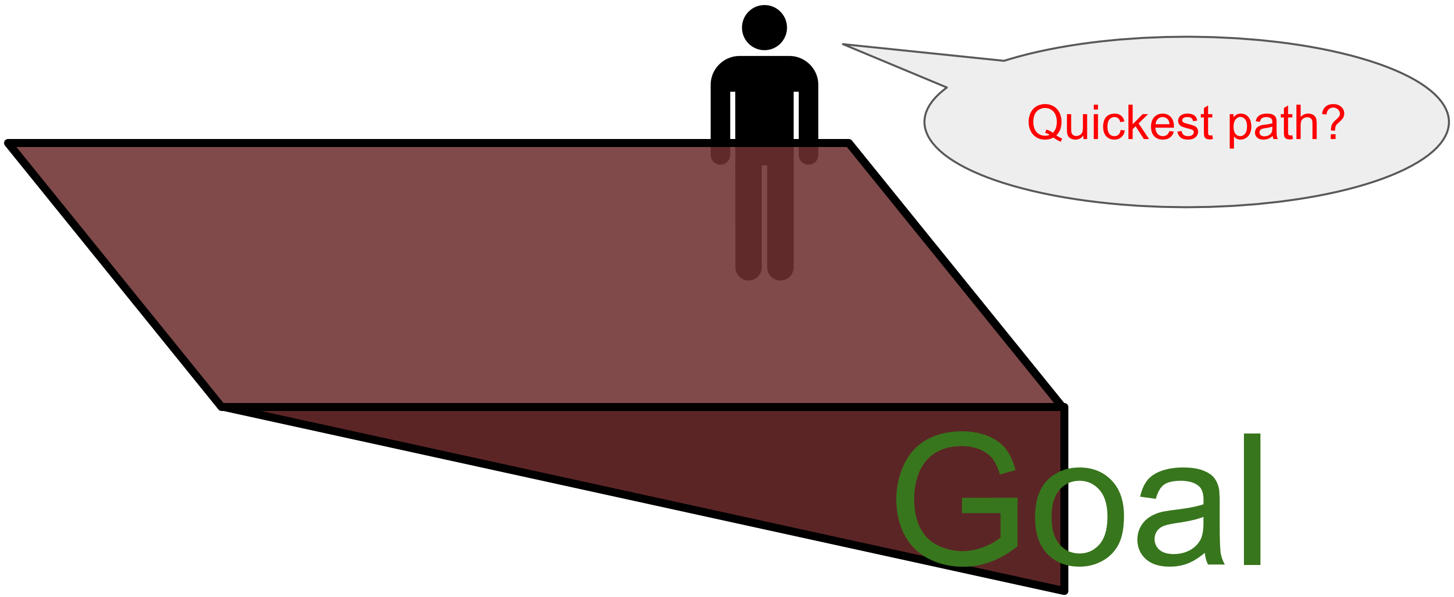


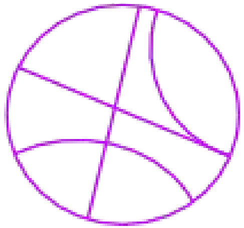
Hyperbolic space: Constant negative curvature

Henri Poincaré: Probably best known for the Poincaré Conjecture (proven in 2003)

Shortest distance: Straight line

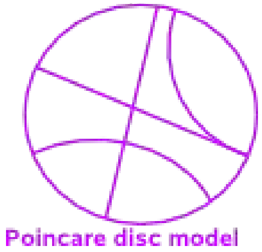


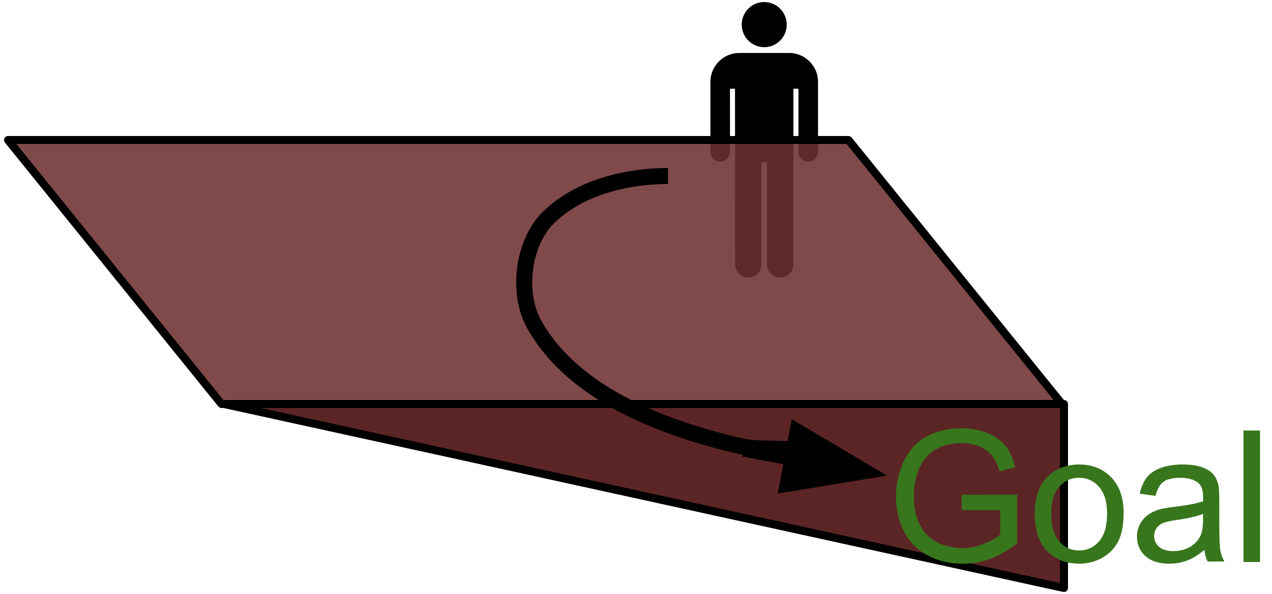




Poincare disc model

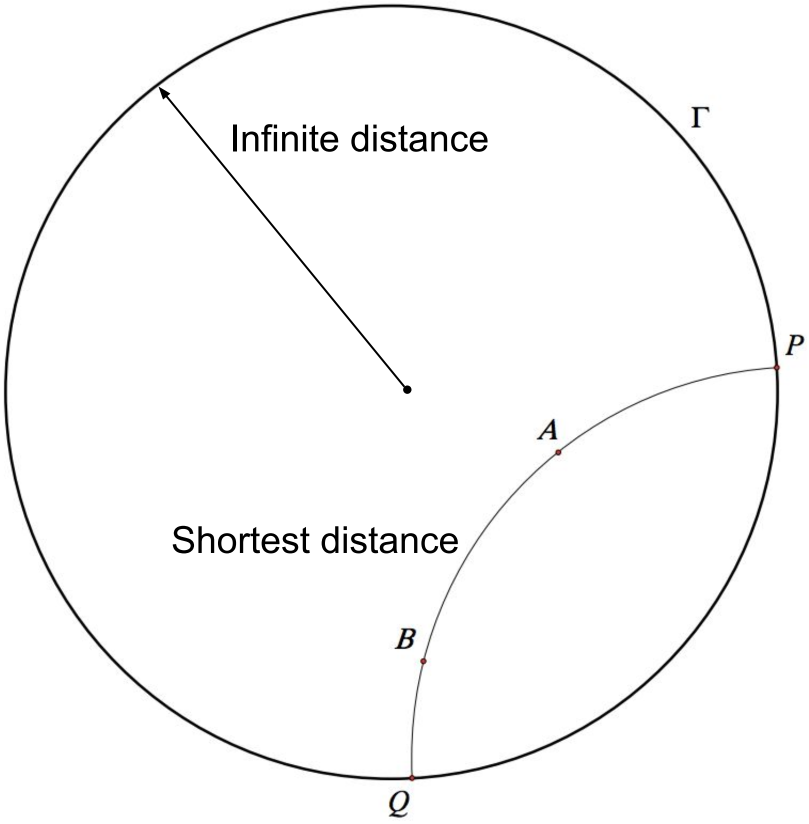
Analogy: Pool of molasses, deeper at the edges.
This slows you down.

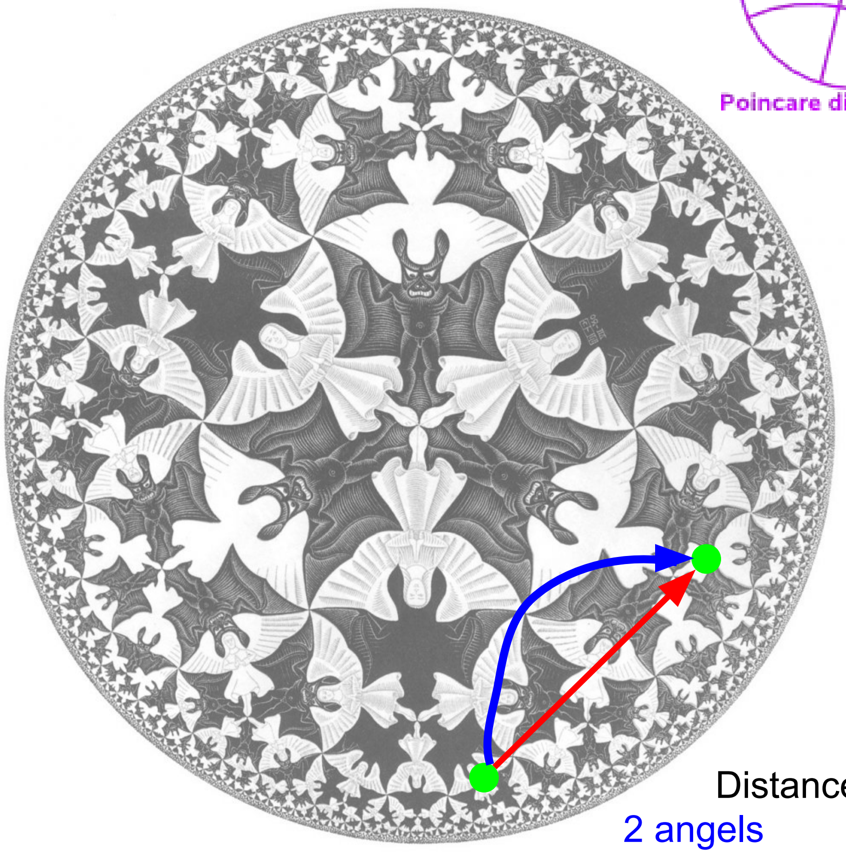


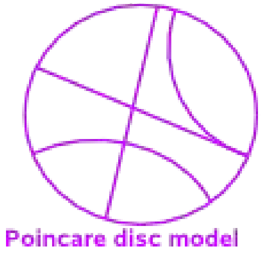


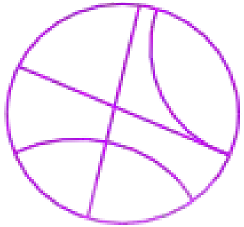
Imperfect analogy: “deeper molasses” not just slower, but a *longer distance*

Analogy: Pool of molasses, deeper at the edges.
This slows you down.


Infinite distance
Γ
P
A
Shortest distance
B
Q

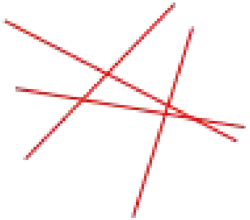

Distance:
2 angels
2 demons + 1 angel


Poincare disc model

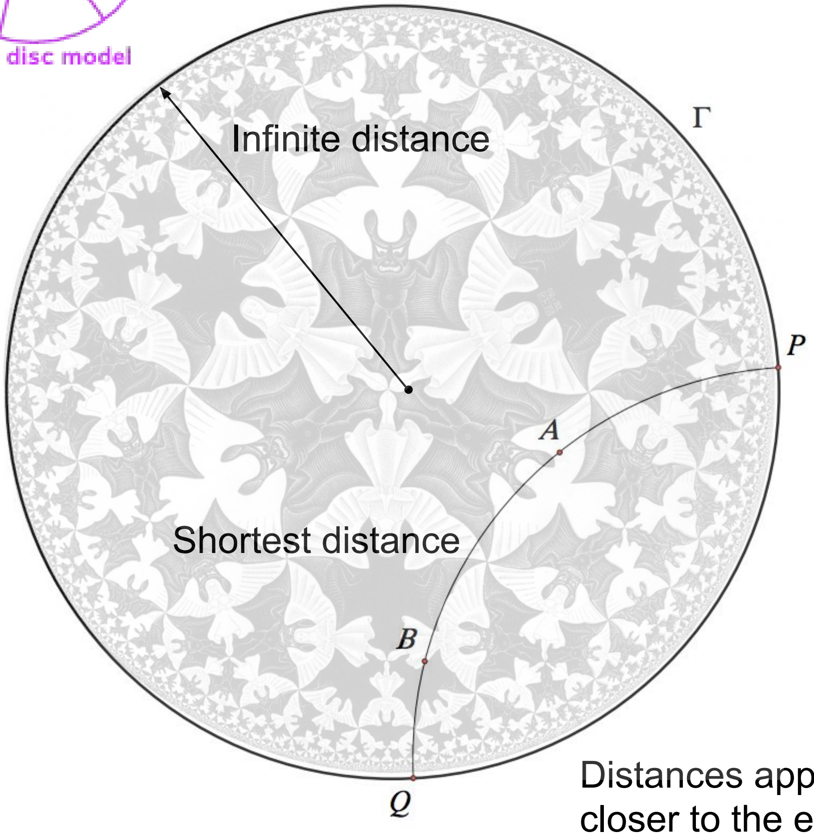


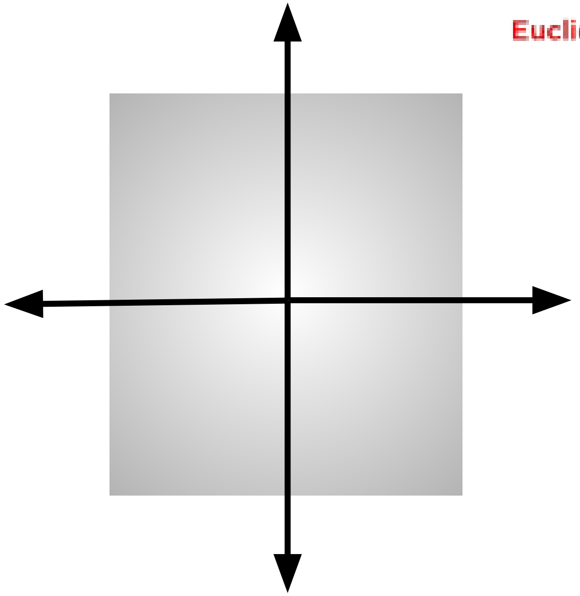
Poincare disc model

$\mathcal{B}^2$ vs. R^2



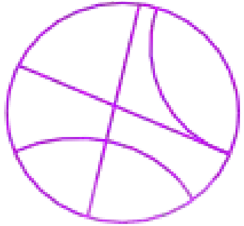
Euclidean Space



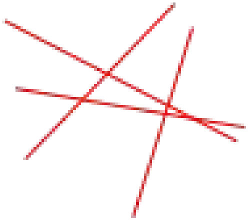


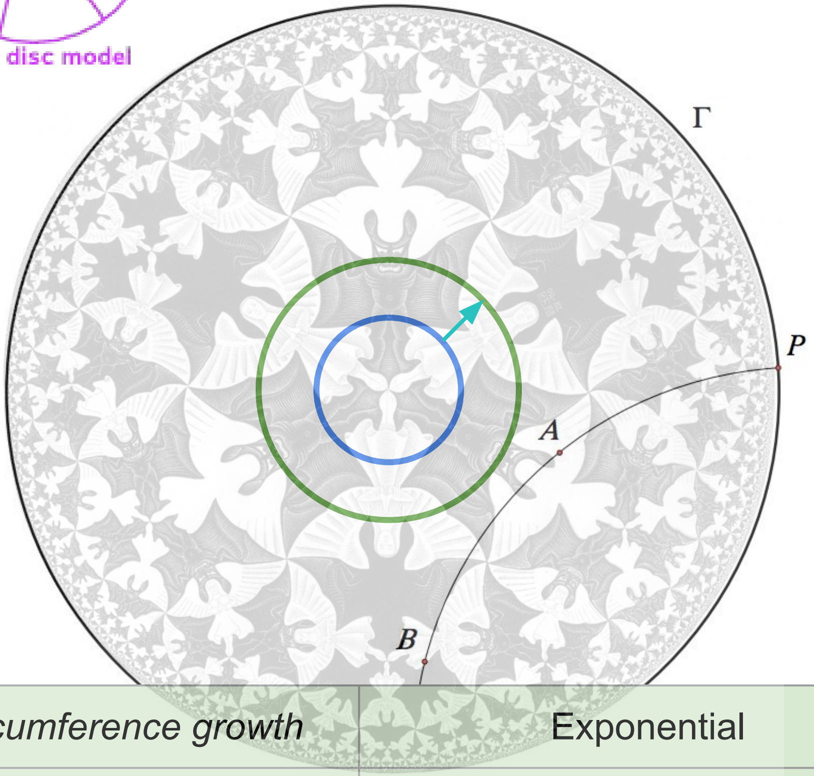
Distances approach inifinity closer to the edges of the disk

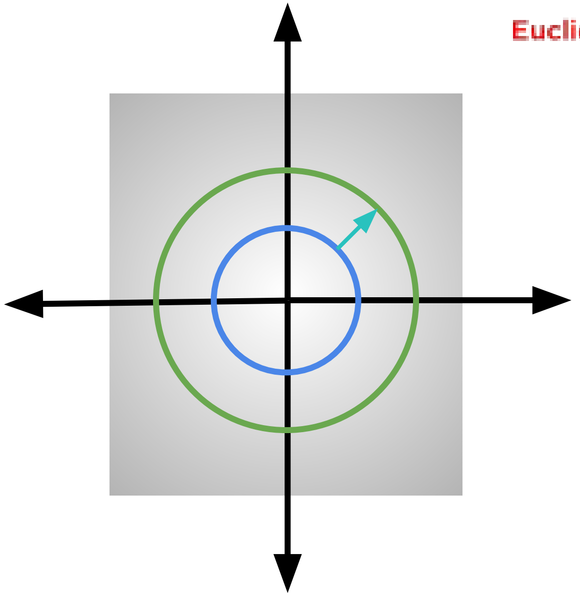
Infinity not shown here, space continues outward



$\mathcal{B}^2$ vs. R^2

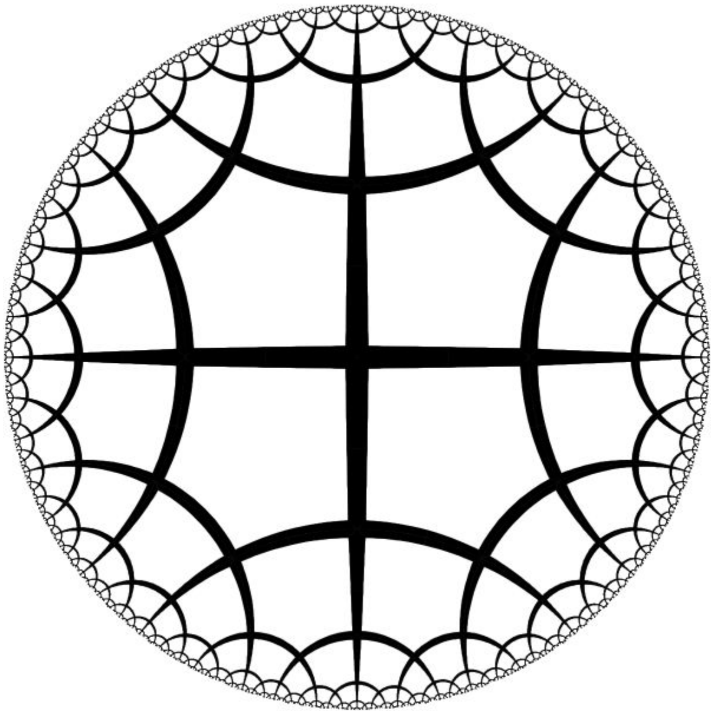




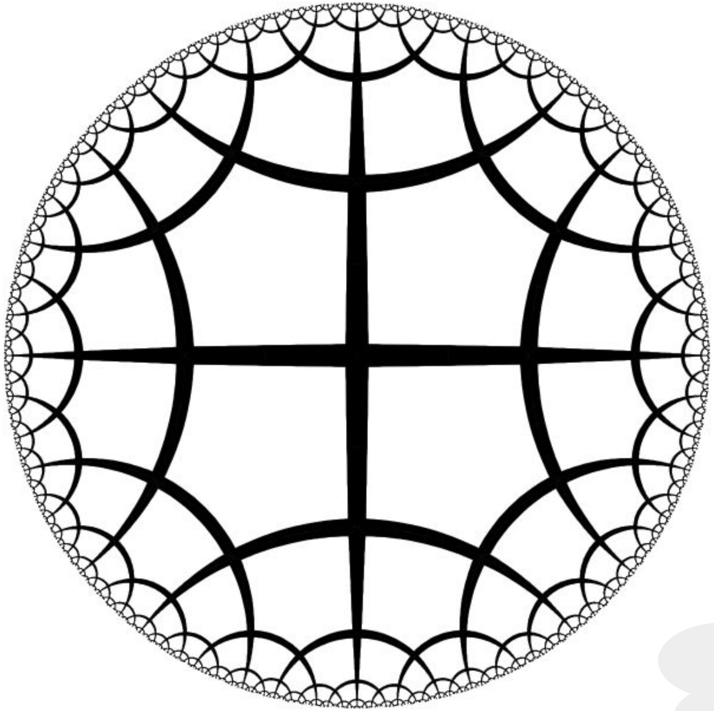


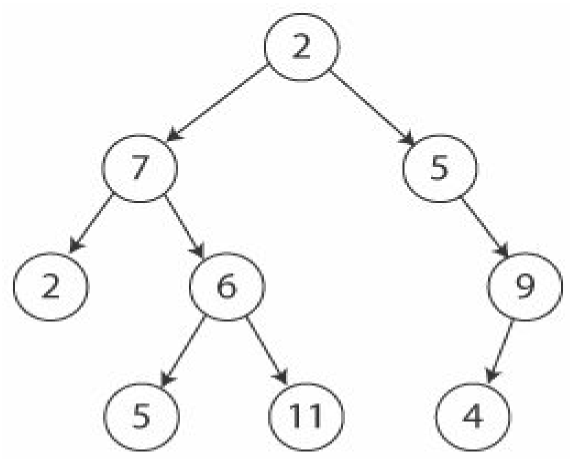
Circumference growth	Exponential	Linear
Area growth	Exponential	Quadratic

What does this look like to you?



Trees!





Actually, the model is a **continuous** version of trees

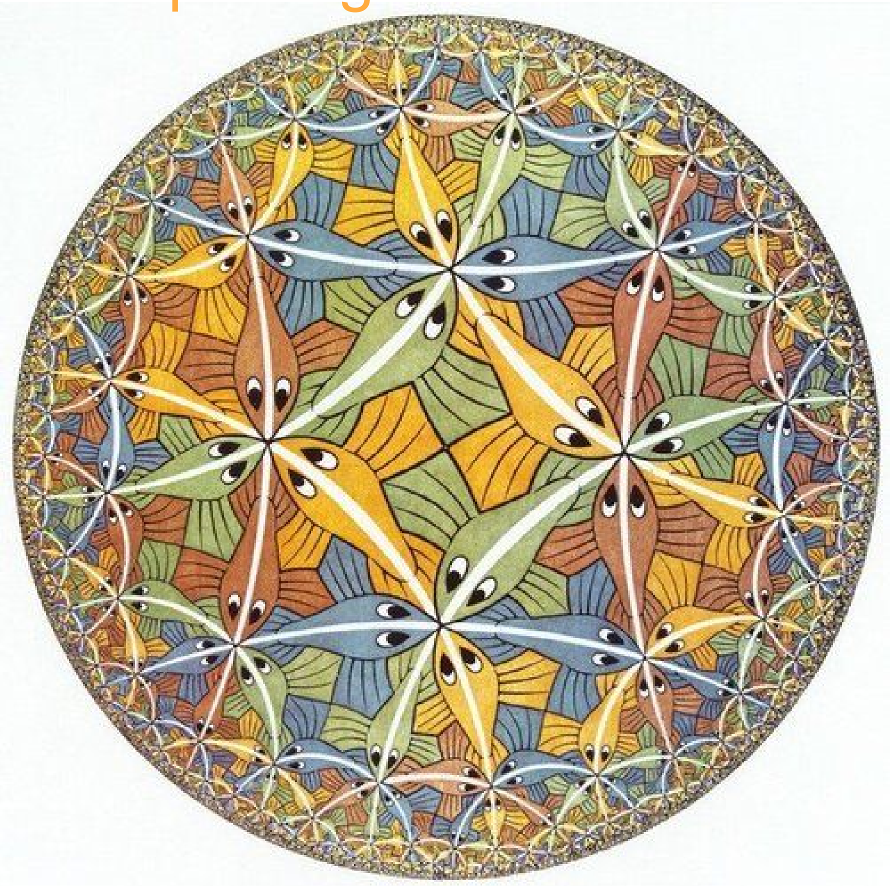
Insight

Hyperbolic space is suited to modelling **hierarchical** (tree-like) relationships

Power law distributions are often a good sign of this (eg: Zipf's law)

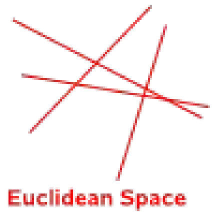
This is the "natural geometry" for language → more **efficient representations**

1. Basic Introduction to Hyperbolic Geometry
2. Mathematical Underpinnings
3. Results

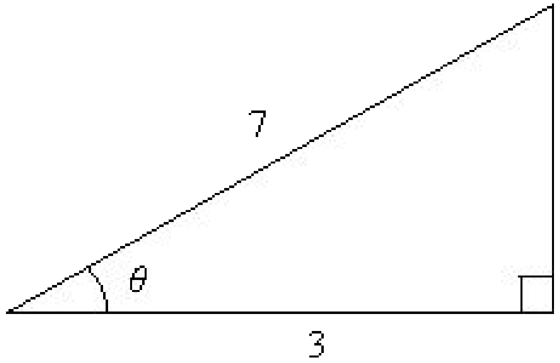


Distance between Two Points - Arccosine

Use inverse of cosine function to find angle (corresponds to distance)

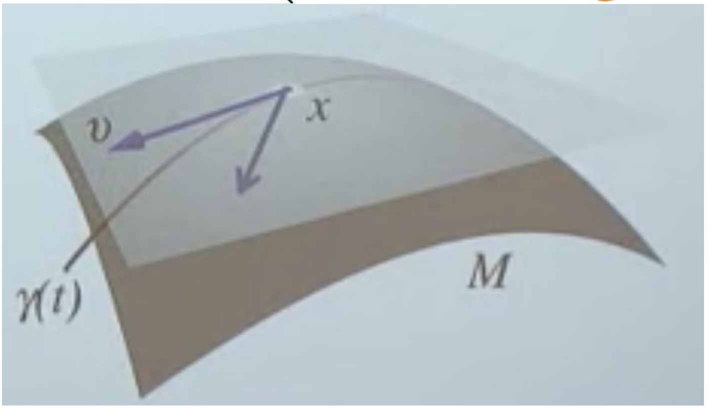


θ = arccos(cos(3/7))



γ(t) =

$d(\boldsymbol{u}, \boldsymbol{v}) = \operatorname{arcosh}\Big($ hmm ... $\Big)$



Distance to a Point - Metric Tensor

Metric tensor: How distance is defined in a space

Distance to a Point - Metric Tensor

Metric tensor: How distance is defined in a space

$g^E = I_d$ Euclidean metric tensor is identiy matrix of size d dimensions

Distance to a Point - Metric Tensor

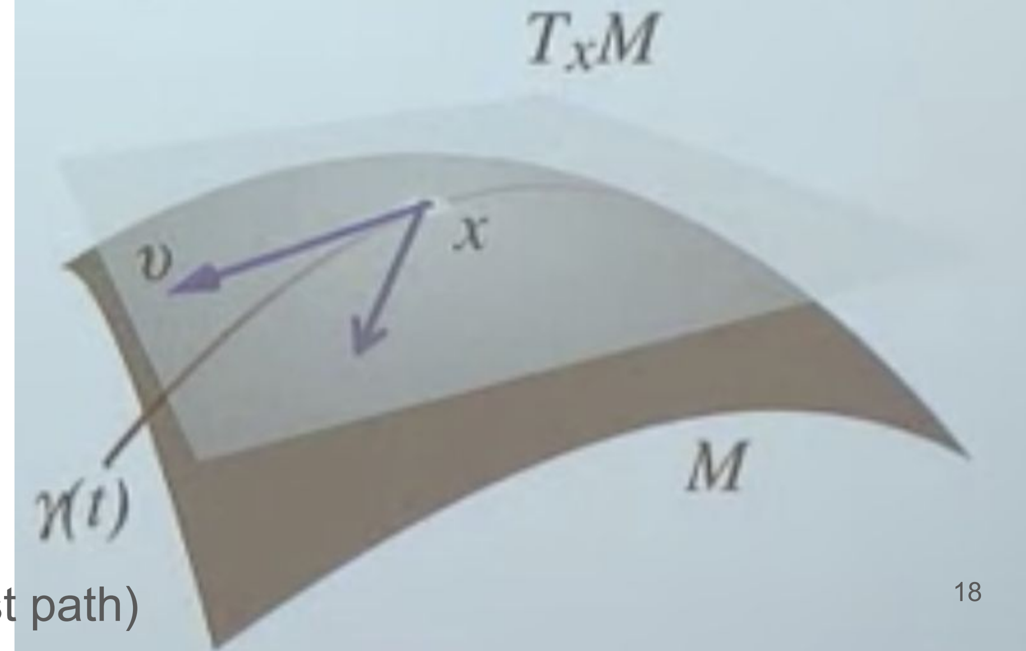
Metric tensor: How distance is defined in a space

Euclidean Space

$g^E = I_d$ Euclidean metric tensor is identiy matrix of size d dimensions

Riemannian Manifold:

Hyperbolic Space



M Hyperbolic manifold

T_xM Plane tangential to M (Euclidean)

$\gamma(t)$ Length of geodesic curve (shortest path)

Distance to a Point - Metric Tensor

Metric tensor: How distance is defined in a space

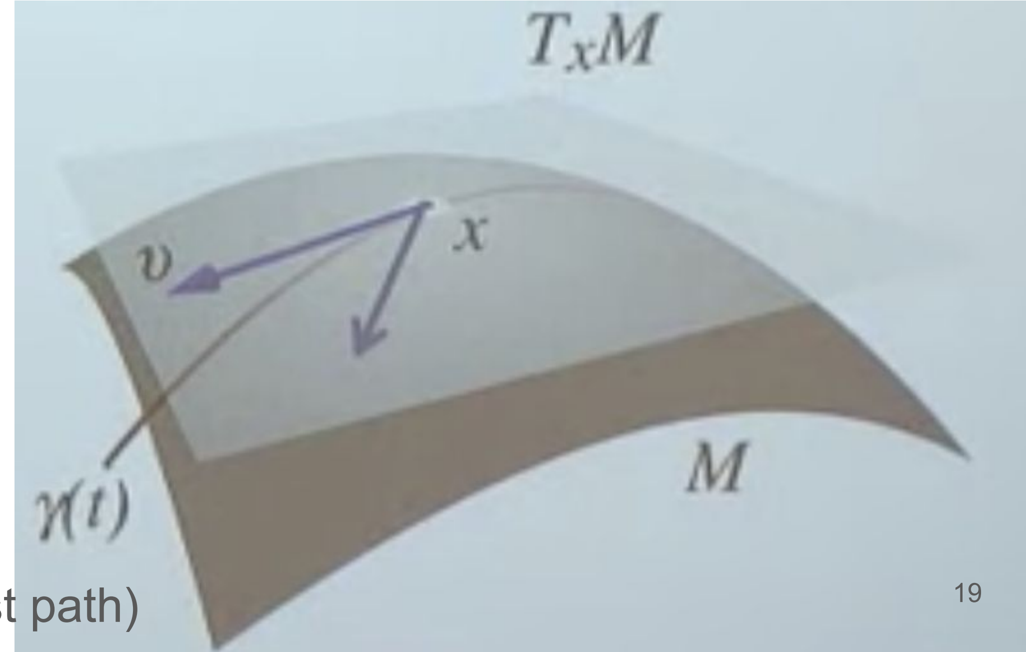
Euclidean Space

$g^E = I_d$ Euclidean metric tensor is identiy matrix of size d dimensions

Riemannian Manifold:

Hyperbolic Space

$$g_{\boldsymbol{x}} = \left(\frac{2}{1 - \|\boldsymbol{x}\|^2} \right)^2 g^E$$



M Hyperbolic manifold

T_xM Plane tangential to M (Euclidean)

$\boldsymbol{\gamma}(t)$ Length of geodesic curve (shortest path)

Poincaré Embedding Distance Function

$$d(\boldsymbol{u}, \boldsymbol{v}) = \operatorname{arcosh}\left(1 + 2\frac{\|\boldsymbol{u} - \boldsymbol{v}\|^2}{(1 - \|\boldsymbol{u}\|^2)(1 - \|\boldsymbol{v}\|^2)}\right)$$

Inverse of hyperbolic cosine
→ *distance*

Ball radius (they only use the Unit ball)

$$g_{\boldsymbol{x}} = \left(\frac{2}{1 - \|\boldsymbol{x}\|^2}\right)^2 g^E$$

Comes from Riemannian metric tensor

Distance changes smoothly
→ Differentiable

★ **Norm** (distance from origin) → **Hierarchy** of objects
Distance → **Similarity** of objects

Derivative of the Distance Funciton

$$\frac{\partial d(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \boldsymbol{\theta}} = \frac{4}{\beta\sqrt{\gamma^2 - 1}} \left(\frac{\|\boldsymbol{x}\|^2 - 2\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle + 1}{\alpha^2} \boldsymbol{\theta} - \frac{\boldsymbol{x}}{\alpha} \right)$$

$$\beta = 1 - \|\boldsymbol{x}\|^2$$

$$\alpha = 1 - \|\boldsymbol{\theta}\|^2$$

$$\gamma = 1 + \frac{2}{\alpha\beta} \|\boldsymbol{\theta} - \boldsymbol{x}\|^2$$

Loss Function

$$\mathcal{L}(\Theta) = \sum_{(u,v)\in\mathcal{D}} \log \frac{e^{-d(\boldsymbol{u},\boldsymbol{v})}}{\sum_{\boldsymbol{v}'\in\mathcal{N}(u)} e^{-d(\boldsymbol{u},\boldsymbol{v}')}}$$

Observed hypernymy pairs

Set of negative examples for *u*
(In training, 10 negative examples were sampled for each positive one)

Intuition: Unconnected nodes should not be close and connected nodes should not be distant

Higher Dimensions

Instead of Poincaré disc ($\mathcal{B}^2$), they use a d-dimensional Poincaré ball ($\mathcal{B}^d$)

- Can model multiple coexisting hierarchies
- Makes optimization easier

Computing embeddings:

$$\Theta' \leftarrow \arg\min_{\Theta} \mathcal{L}(\Theta) \qquad \text{s.t. } \forall\, \boldsymbol{\theta}_i \in \Theta : \|\boldsymbol{\theta}_i\| < 1$$

Set of all embeddings

An embedding on the Poincare ball

Loss function

Optimization

Poincare ball has a “Riemannian manifold structure”

- Smooth, therefore differentiable

Optimize with “stochastic Riemannian optimization”

- eg: RSGD or RSVRG
- Computational and memory complexity is linear in relation to the embedding dimension

Optimization

Updating Parameters with RSGD

$$\boldsymbol{\theta}_{t+1} = \mathfrak{R}_{\theta_t}\left(-\eta_t \nabla_R \mathcal{L}(\boldsymbol{\theta}_t)\right)$$

Retraction onto $\mathcal{B}$ at θ

$$\mathfrak{R}_\theta(\boldsymbol{v}) = \boldsymbol{\theta} + \boldsymbol{v}$$

Learning rate at time t

Riemannian gradient

In topology, a **retraction** means mapping onto a subspace while preserving position of all points

Angles in $\mathcal{B}^2$ correspond to angles in $\mathbb{R}^2$, so can rescale to ∇_E

Updating Parameters with RSGD (resulting equation)

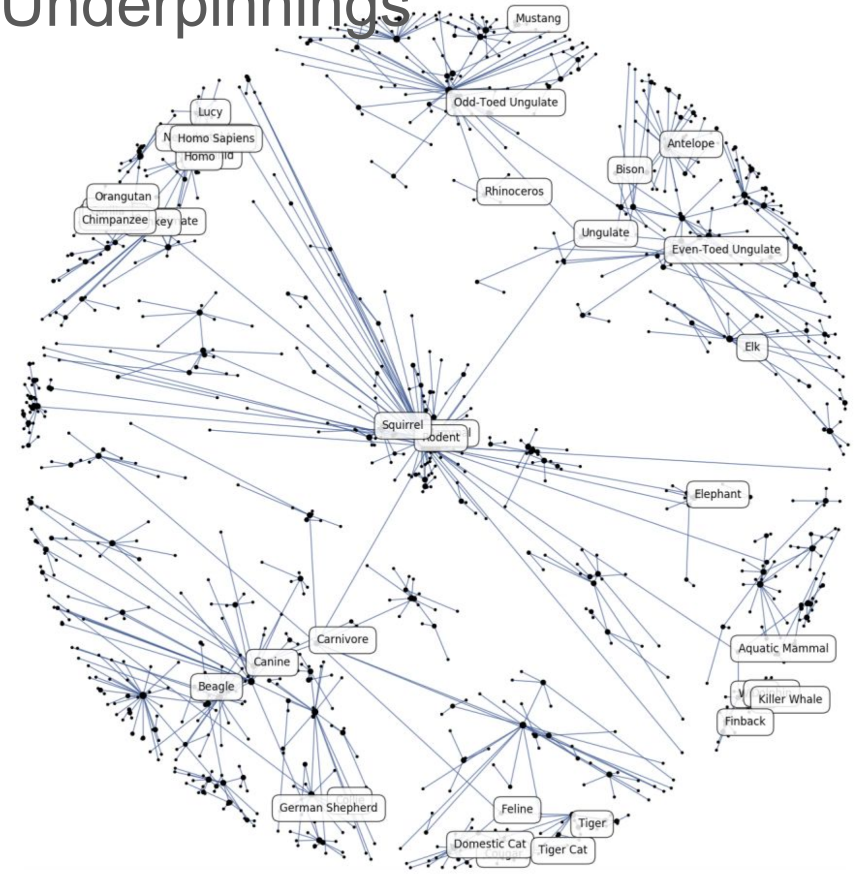
$$\boldsymbol{\theta}_{t+1} \leftarrow \operatorname{proj}\left(\boldsymbol{\theta}_t - \eta_t \frac{(1 - \|\boldsymbol{\theta}_t\|^2)^2}{4} \nabla_E\right)$$

To constrain results to the Poincaré ball

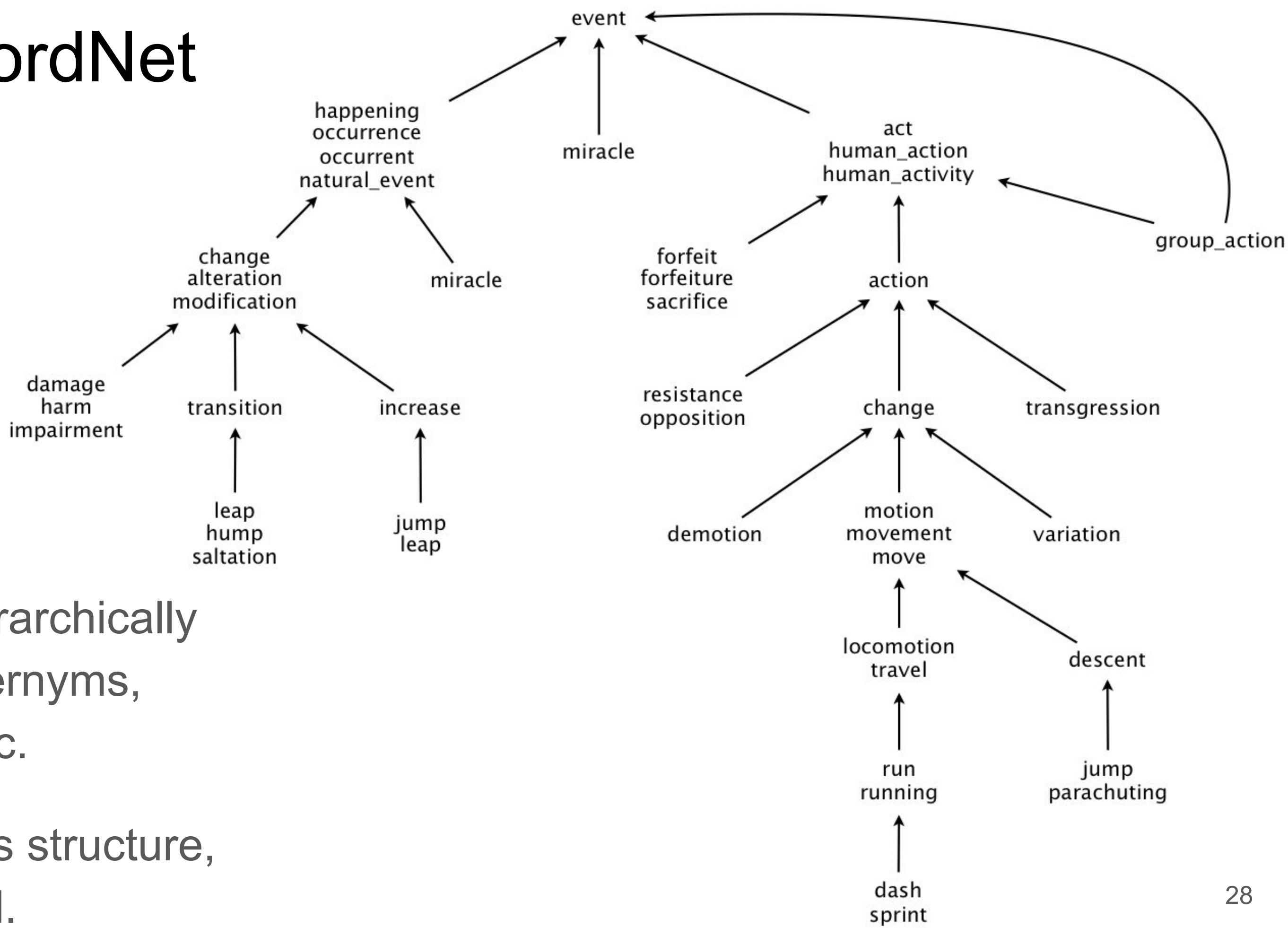
$$\operatorname{proj}(\boldsymbol{\theta}) = \begin{cases} \boldsymbol{\theta}/\|\boldsymbol{\theta}\| - \varepsilon & \text{if } \|\boldsymbol{\theta}\| \geq 1 \\ \boldsymbol{\theta} & \text{otherwise ,} \end{cases}$$

Euclidean gradient

1. Basic Introduction to Hyperbolic Geometry
2. Mathematical Underpinnings
3. Results



Example of WordNet



Lexical databse higherarchically organized into hypernyms, hyponyms, etc.

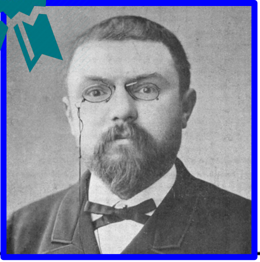
Their task is to infer this structure, unsupervised.

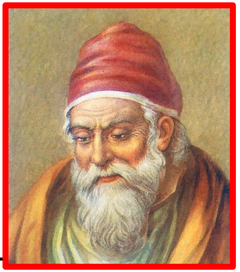
Table 1: Experimental results on the transitive closure of the WordNet noun hierarchy. Highlighted cells indicate the best Euclidean embeddings as well as the Poincaré embeddings which acheive equal or better results. Bold numbers indicate absolute best results.

			Dimensionality					
			5	10	20	50	100	200
WordNet Reconstruction	**Euclidean**	Rank	3542.3	2286.9	1685.9	1281.7	1187.3	1157.3
		MAP	0.024	0.059	0.087	0.140	0.162	0.168
	Translational	Rank	205.9	179.4	95.3	92.8	92.7	91.0
		MAP	0.517	0.503	0.563	0.566	0.562	0.565
	Poincaré	Rank	4.9	4.02	3.84	3.98	3.9	**3.83**
		MAP	0.823	0.851	0.855	0.86	0.857	**0.87**
WordNet Link Pred.	**Euclidean**	Rank	3311.1	2199.5	952.3	351.4	190.7	81.5
		MAP	0.024	0.059	0.176	0.286	0.428	0.490
	Translational	Rank	65.7	56.6	52.1	47.2	43.2	40.4
		MAP	0.545	0.554	0.554	0.56	0.562	0.559
	Poincaré	Rank	5.7	**4.3**	4.9	4.6	4.6	4.6
		MAP	0.825	0.852	0.861	**0.863**	0.856	0.855

MAP: Mean average precision

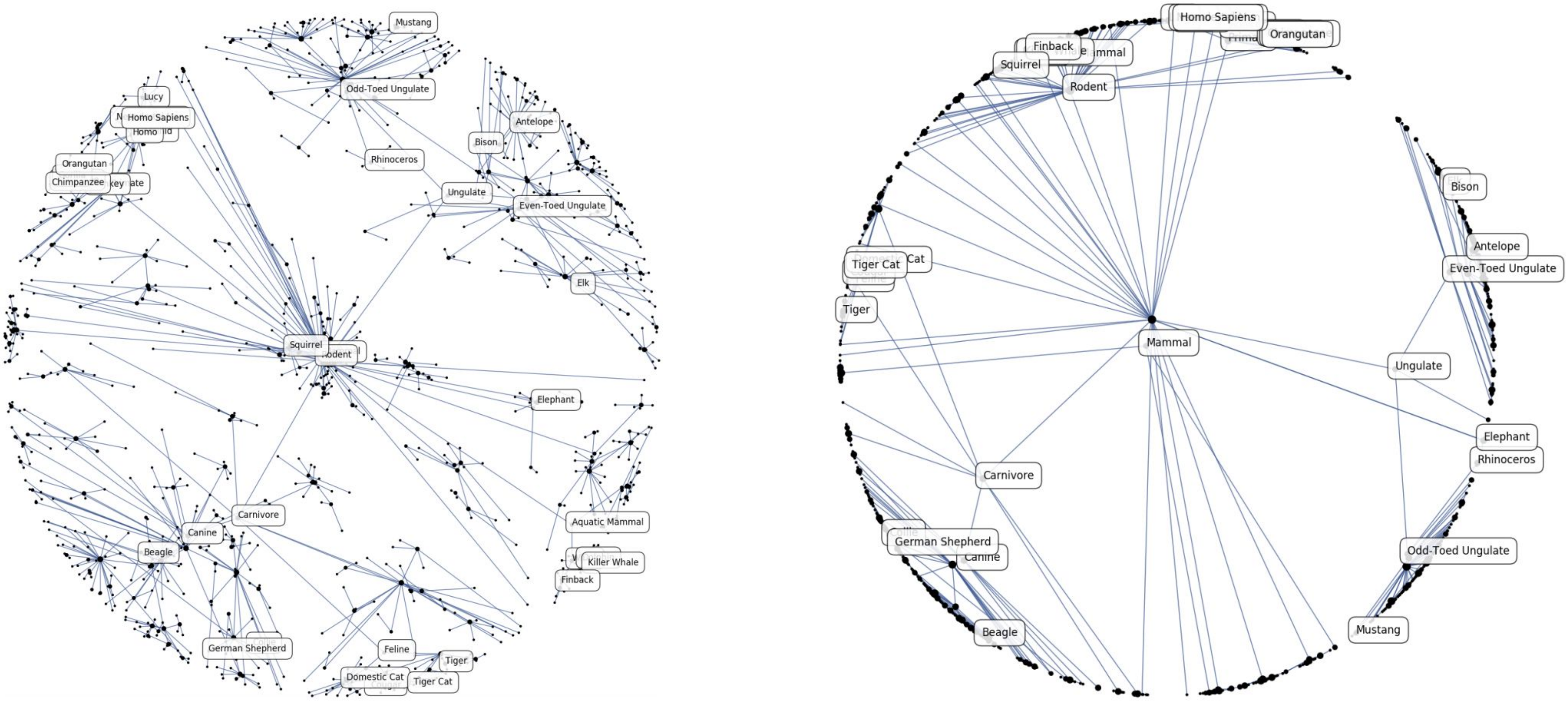
Even just
5 dimensions Poincaré
outperforms
200 dimensions Euclidean





MAP: Mean average precision

			Dimensionality					
			5	10	20	50	100	200
WordNet Reconstruction	**Euclidean**	Rank	3542.3	2286.9	1685.9	1281.7	1187.3	1157.3
		MAP	0.024	0.059	0.087	0.140	0.162	0.168
	Translational	Rank	205.9	179.4	95.3	92.8	92.7	91.0
		MAP	0.517	0.503	0.563	0.566	0.562	0.565
	Poincaré	Rank	4.9	4.02	3.84	3.98	3.9	**3.83**
		MAP	0.823	0.851	0.855	0.86	0.857	**0.87**
WordNet Link Pred.	**Euclidean**	Rank	3311.1	2199.5	952.3	351.4	190.7	81.5
		MAP	0.024	0.059	0.176	0.286	0.428	0.490
	Translational	Rank	65.7	56.6	52.1	47.2	43.2	40.4
		MAP	0.545	0.554	0.554	0.56	0.562	0.559
	Poincaré	Rank	5.7	**4.3**	4.9	4.6	4.6	4.6
		MAP	0.825	0.852	0.861	**0.863**	0.856	0.855



(a) Intermediate embedding after 20 epochs

(b) Embedding after convergence

Trained only on the mammals subtree of **WordNet**

Blue lines represent ground truth from Wordnet

Additional Results

Also evaluated on:

Link prediction in social networks (not NLP-related)

and **lexical entailment** (using HyperLex)

Quantifying what *degree* X is a type of Y, via ratings on a scale of [0, 10]

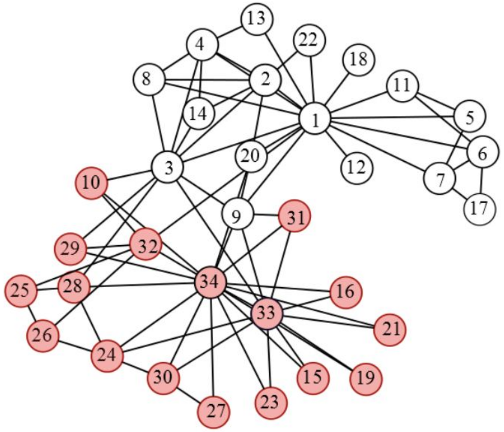
Table 3: Spearman's ρ for Lexical Entailment on HYPERLEX.

	FR	SLQS-Sim	WN-Basic	WN-WuP	WN-LCh	Vis-ID	Euclidean	Poincaré
ρ	0.283	0.229	0.240	0.214	0.214	0.253	0.389	0.512

From *Neural Embeddings of Graphs in Hyperbolic Space*, Chamberlain et. al.

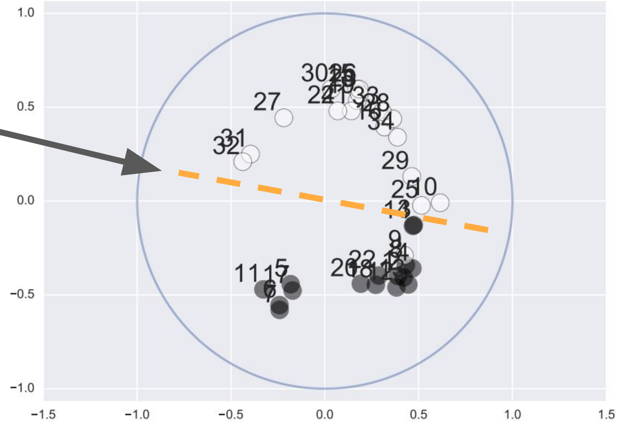
https://arxiv.org/pdf/1705.10359.pdf

May 2017

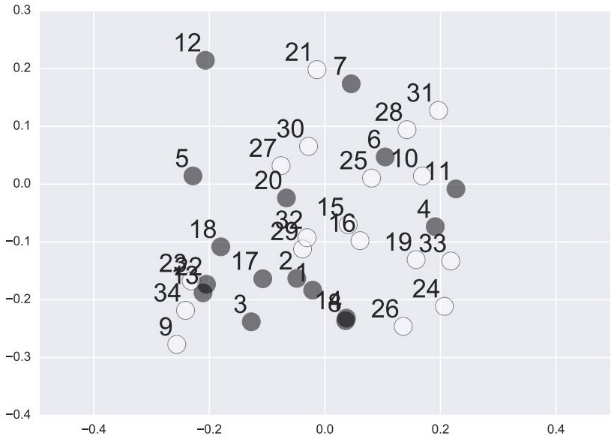


(a) Zachary's karate network. The network is split into two factions.

The two factions are easily seperable in this representation



(b) Two-dimensional hyperbolic embedding of the karate network in the Poincaré disk.



(c) Two dimensional Deepwalk embedding of the karate network.

Future Direction

“Expand the applications of Poincaré embeddings”

eg: To multi-relational data

“Derive models that are tailored to specific applications”

→ Applications to less hierarchical datasets?

Summary

The point of embeddings is to **organize words according to their semantic/functional similarity**

Summary

The point of embeddings is to **organize words according to their semantic/functional similarity**

And language has **natural hierarchies** (as do taxonomies, knowledge graphs, etc.)

Summary

The point of embeddings is to **organize words according to their semantic/functional similarity**

And language has **natural hierarchies** (as do taxonomies, knowledge graphs, etc.)

Euclidean space is not suited to representing hierarchies (*“doing things the hard way”*)

Summary

The point of embeddings is to **organize words according to their semantic/functional similarity**

And language has **natural hierarchies** (as do taxonomies, knowledge graphs, etc.)

Euclidean space is not suited to representing hierarchies (*“doing things the hard way”*)

But **Hyperbolic space is suited** to this (*“the natural geometry for language”*)

Summary

The point of embeddings is to **organize words according to their semantic/functional similarity**

And language has **natural hierarchies** (as do taxonomies, knowledge graphs, etc.)

Euclidean space is not suited to representing hierarchies (*“doing things the hard way”*)

But **Hyperbolic space is suited** to this (*“the natural geometry for language”*)

Poincaré disk: Distances grow to infinity as you approach the edge, differentiable & optimizable

Summary

The point of embeddings is to **organize words according to their semantic/functional similarity**

And language has **natural hierarchies** (as do taxonomies, knowledge graphs, etc.)

Euclidean space is not suited to representing hierarchies (*"doing things the hard way"*)

But **Hyperbolic space is suited** to this (*"the natural geometry for language"*)

Poincaré disk: Distances grow to infinity as you approach the edge, differentiable & optimizable

Embeddings on **Poincaré ball perform better**, even with an order of magnitude **lower dimensionality**

Useful References

- *Poincaré Embeddings for Learning Hierarchical Representations*, Nickel et. al.

May 2017, https://arxiv.org/pdf/1705.08039.pdf

- From *Neural Embeddings of Graphs in Hyperbolic Space*, Chamberlain et. al.

May 2017, https://arxiv.org/pdf/1705.10359.pdf

- *Implementing Poincaré Embeddings*, Jain (gensim)

Dec. 2017, https://rare-technologies.com/implementing-poincare-embeddings/

https://radimrehurek.com/gensim/models/poincare.html