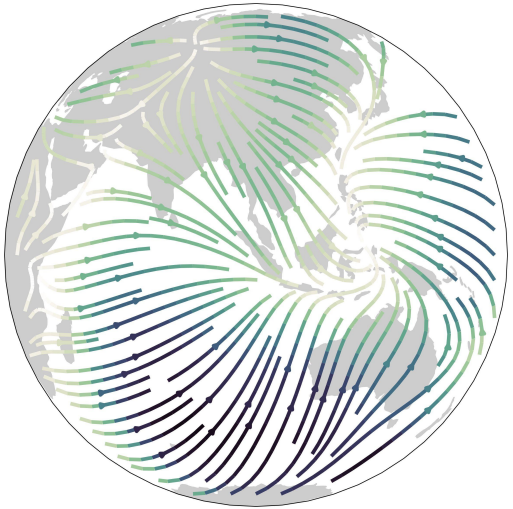
# Geometry and Deep generative modelling

*Geometry as an inductive bias and a manifold constraint*

*Emile Mathieu*

*Data-Centric Engineering Reading Group,*
*@ The Alan Turing Institute.*



UNIVERSITY OF
OXFORD

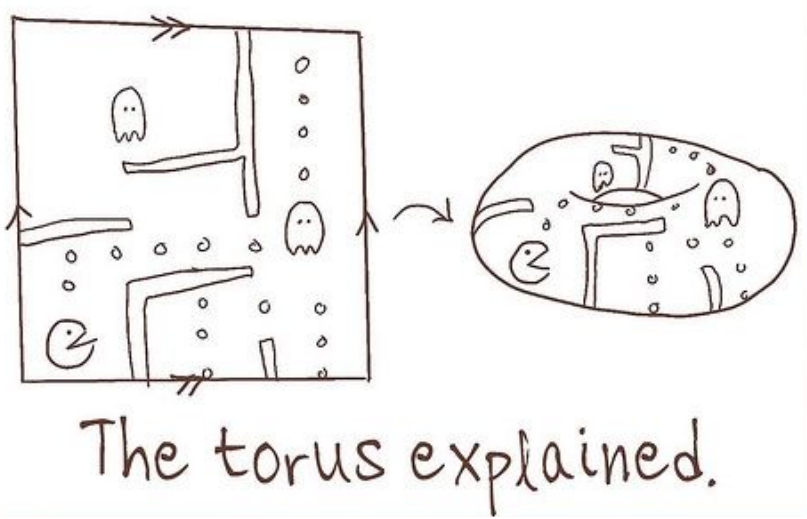
# Overview: Geometry and Deep generative models

- *Geometry as an inductive bias*
  - Variational Auto-Encoder for hierarchical data

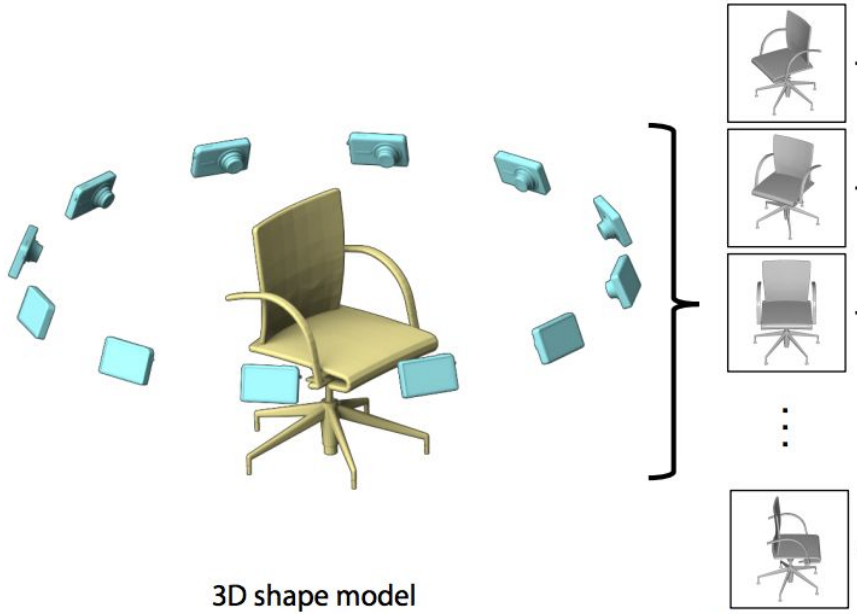
- *Geometry as a manifold constraint*
  - Normalizing Flow for manifold data

# Geometry as an inductive bias

# Data with loops

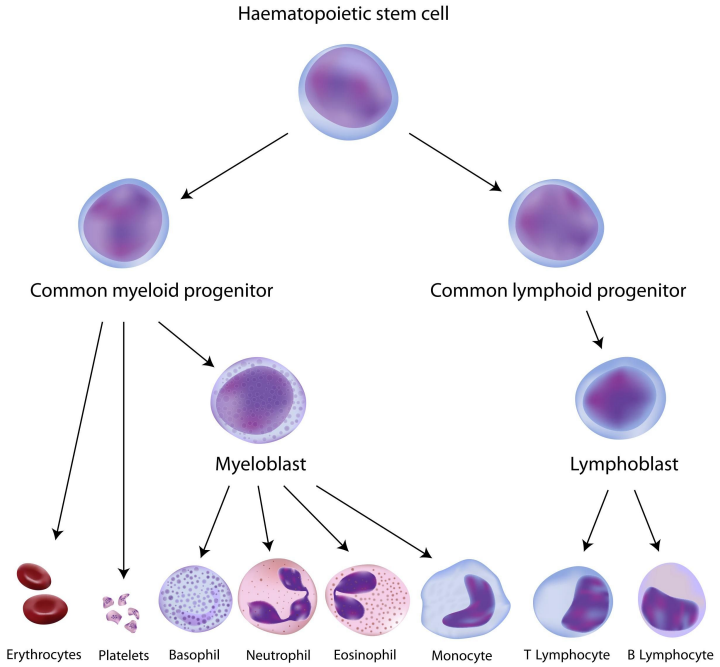



From Human Mathematics Blog, 2009

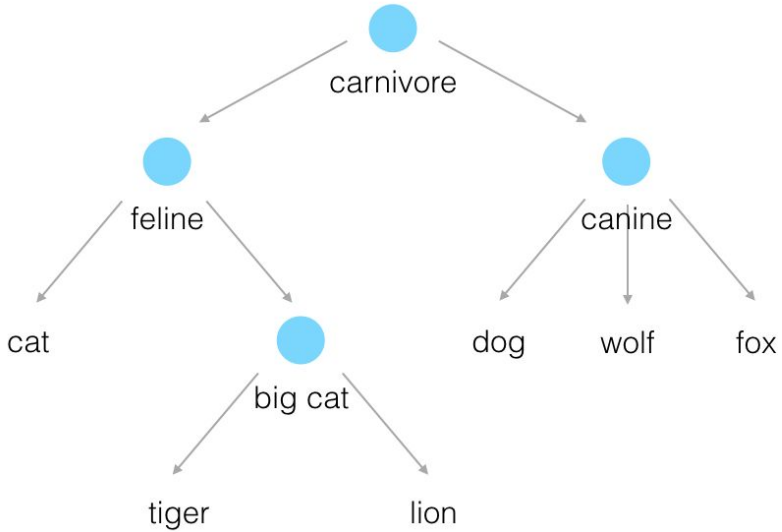



From Connelly Barnes, 3D Reconstruction and Understanding, 2017

# Data with hierarchy




From Tom Ulrich et al. 2014




Fragment of WordNet taxonomy graph

# Poincaré Variational Auto-Encoders

A generative model for data with underlying hierarchical structure

- Latent space endowed with hyperbolic geometry
- Derive necessary methods for two main Gaussian generalisations on hyperbolic space
- Decoder architecture that is geometry-aware

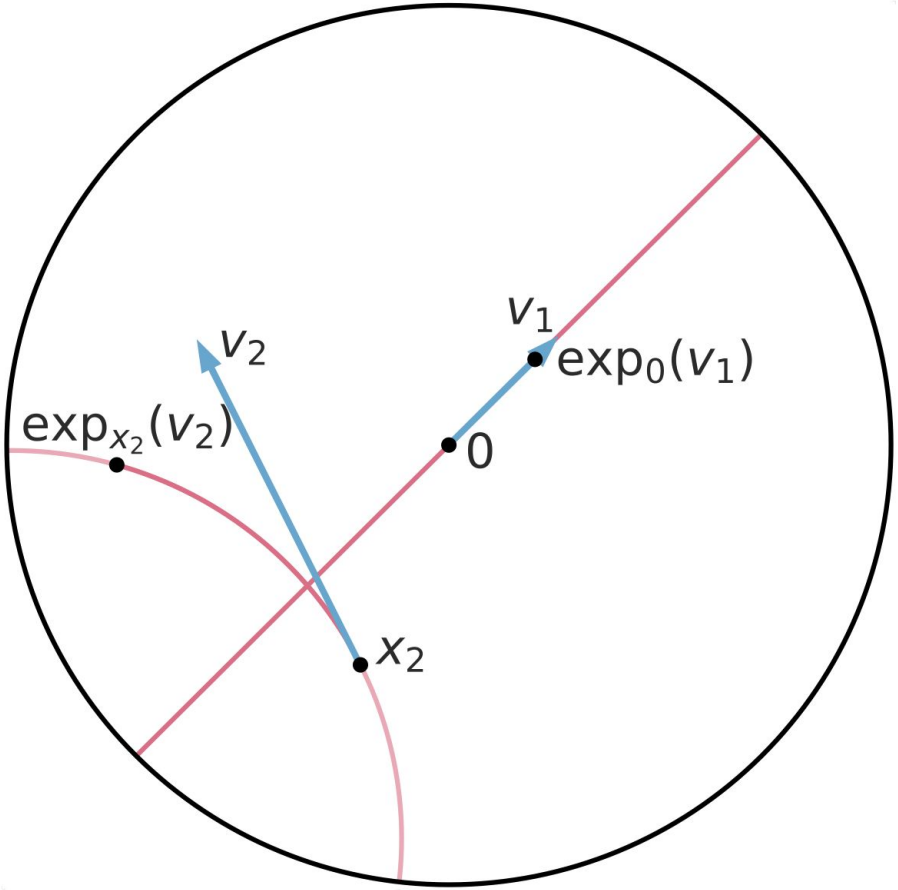
With C. Le Lan, C. Maddison, R. Tomioka, Y.W Teh

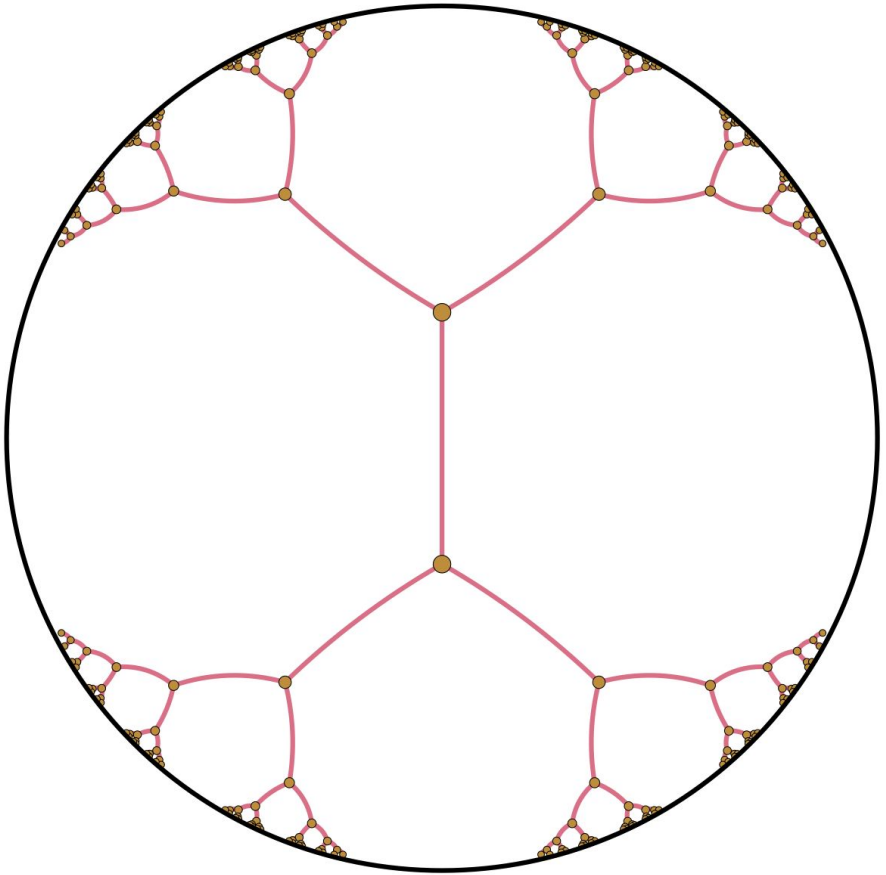
# Motivation

Theorem [Bourgain] A Euclidean space is unable to obtain low distortion for tree embeddings—even using an unbounded number of dimensions.

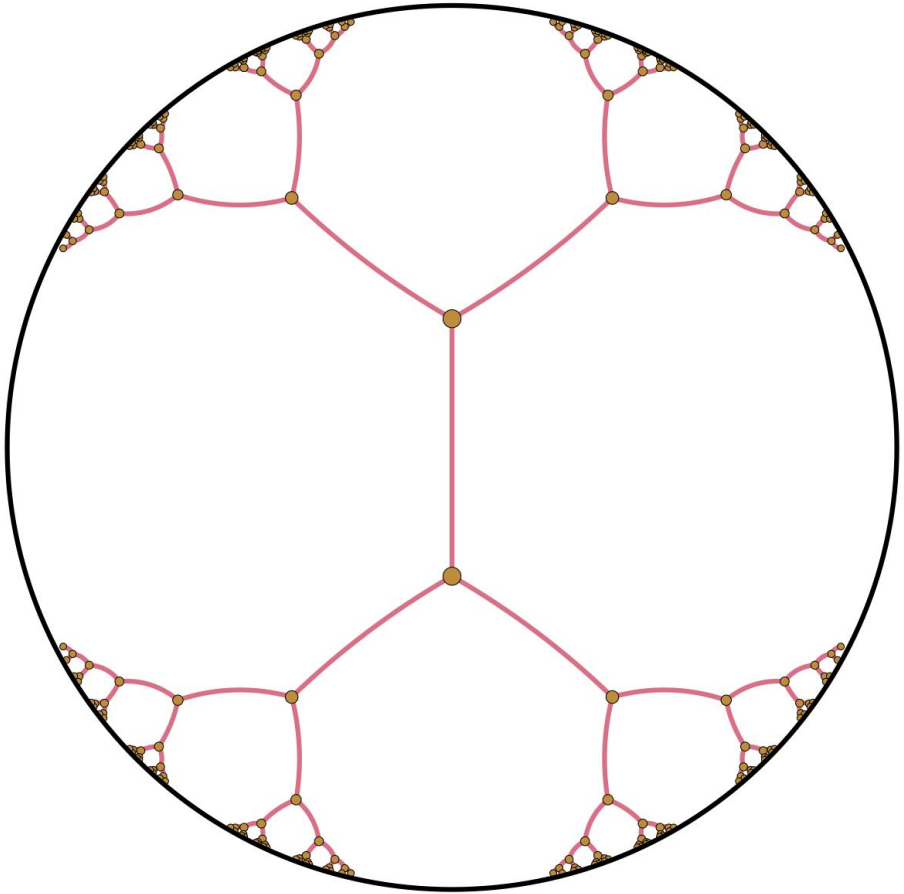
Theorem [R. Sarkar, 2011] Trees can be embedded with arbitrarily low distortion into the Poincaré disk.

# Poincaré Ball: A model of Hyperbolic geometry



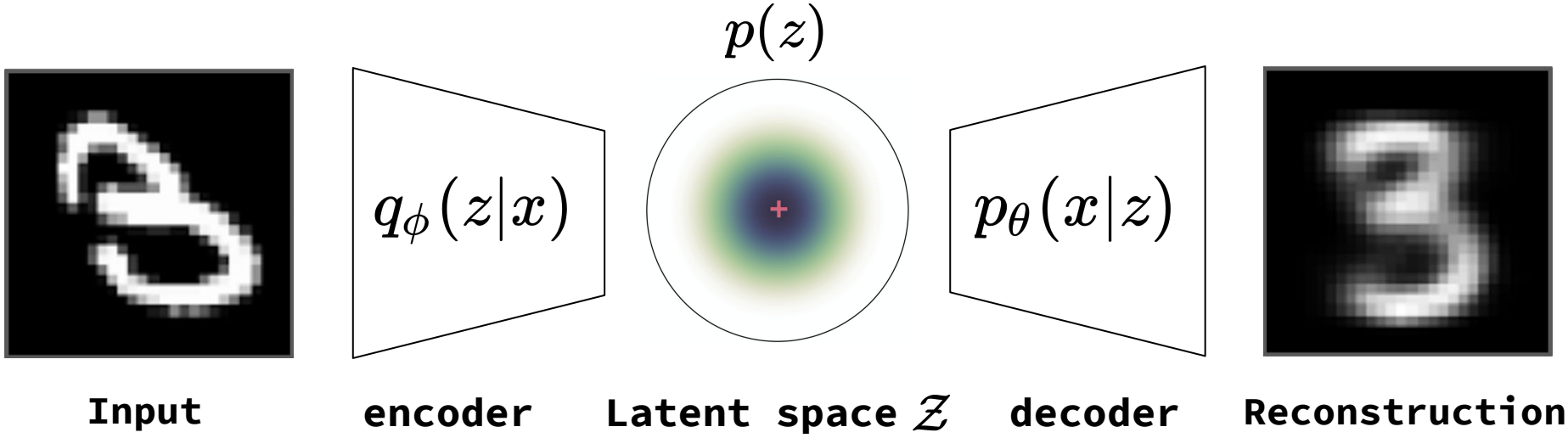



# Poincaré Ball: A model of Hyperbolic geometry

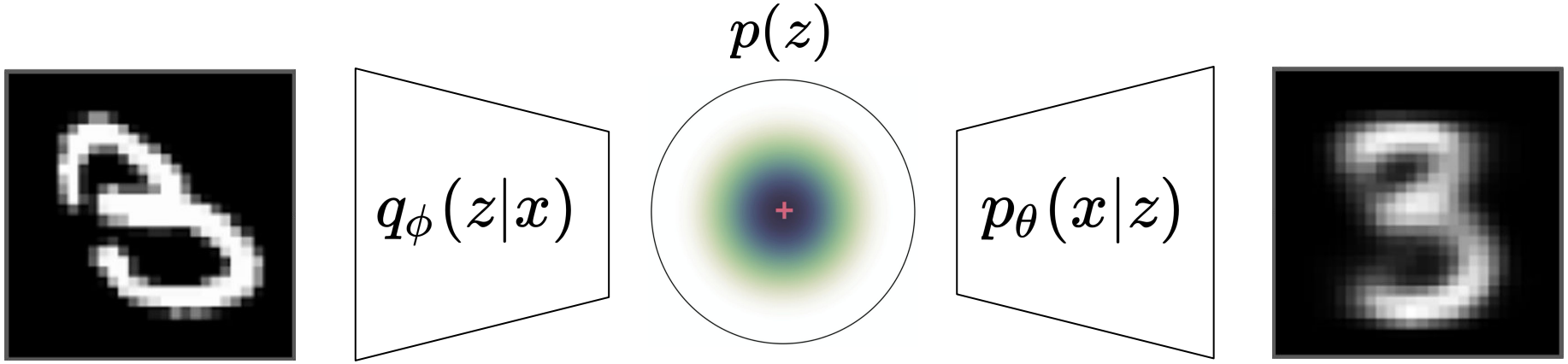


# Model

# Poincaré Variational Auto-Encoder

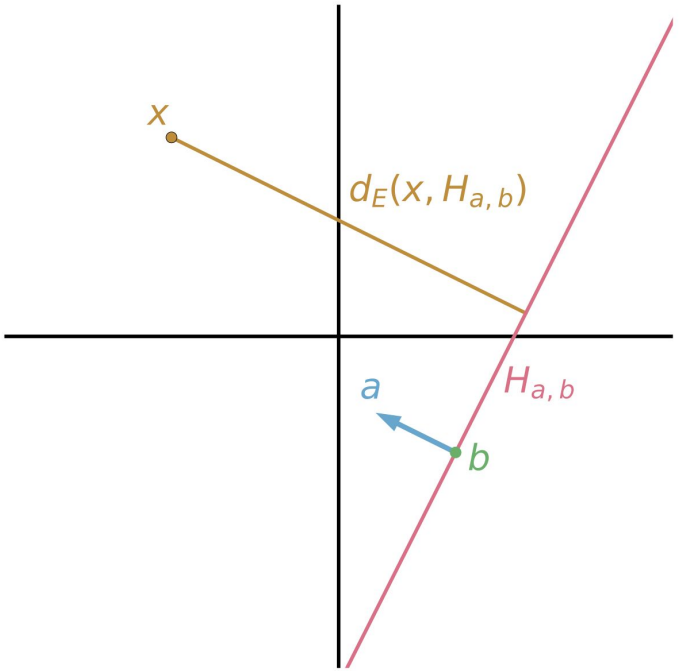


# Poincaré Variational Auto-Encoder



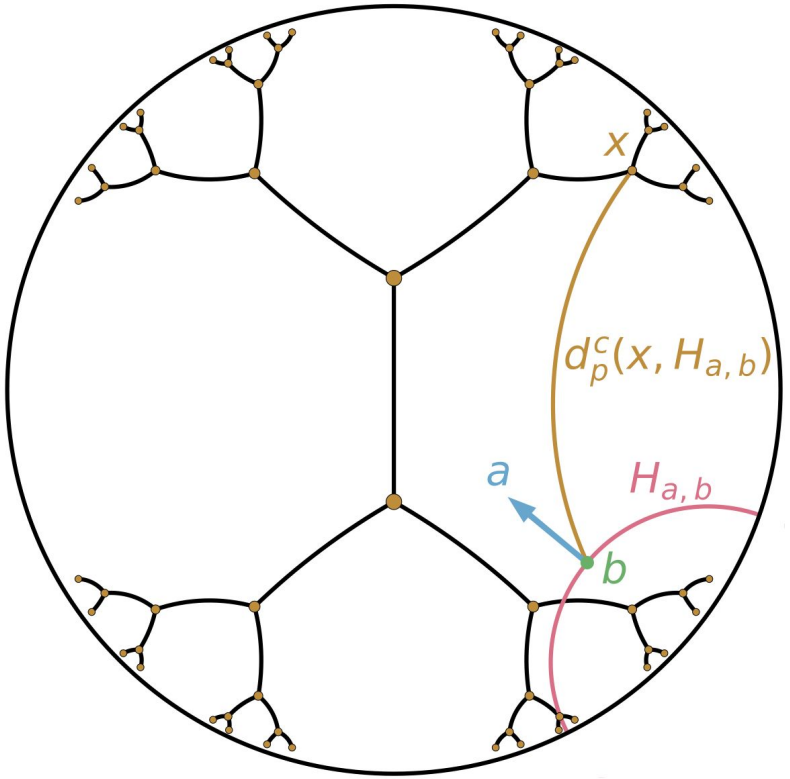

$$\log p(\boldsymbol{x}) \geq \int_{\mathcal{M}} \ln \left( \frac{p_\theta(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \right) q_\phi(\boldsymbol{z}|\boldsymbol{x}) \, d\mathcal{M}(\boldsymbol{z})$$

# **Decoder**: From latents to observations



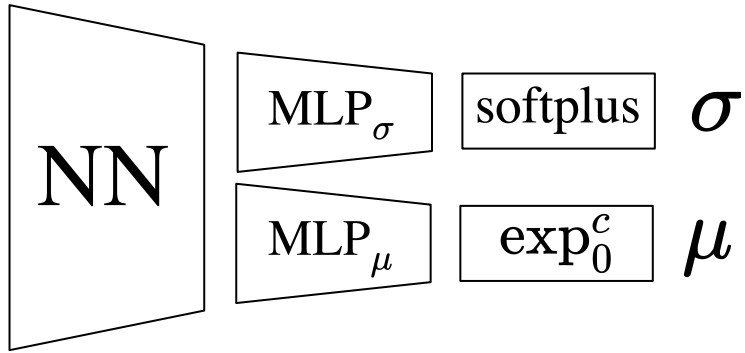

$$f_{\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{x}) = \langle \boldsymbol{a}, \boldsymbol{x} - \boldsymbol{b} \rangle$$

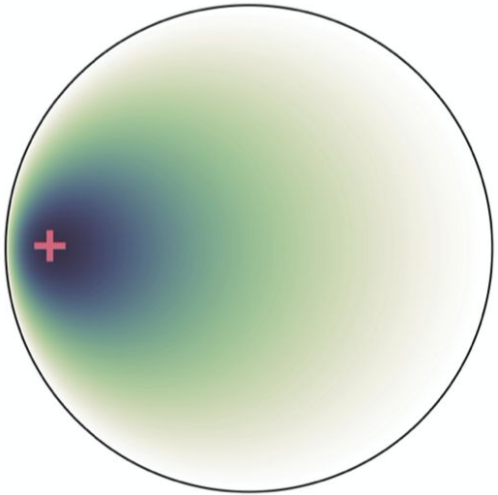
$$f_{\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{x}) = \operatorname{sign}(\langle \boldsymbol{a}, \boldsymbol{x} - \boldsymbol{b} \rangle) \, \|\boldsymbol{a}\| \, d_E(\boldsymbol{x}, H^c_{\boldsymbol{a},\boldsymbol{b}})$$




$$f^c_{\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{x}) = \operatorname{sign}(\langle \boldsymbol{a}, \log^c_{\boldsymbol{x}}(\boldsymbol{b}) \rangle_{\boldsymbol{b}}) \, \|\boldsymbol{a}\|_{\boldsymbol{b}} \, d^c_p(\boldsymbol{x}, H^c_{\boldsymbol{a},\boldsymbol{b}})$$

# Encoder: From observations to latents



$$\exp_0^c : \mathbb{R}^d \to \mathbb{B}_c^d$$

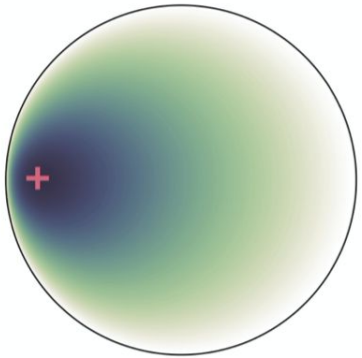
$$u \mapsto \tanh\left(\frac{\sqrt{c}}{2}\|u\|_2\right)\frac{u}{\sqrt{c}|u\|_2}$$

$$q_\phi(z|x) = q(z;\mu,\sigma)$$

# Generalizing the Normal distribution

## Riemannian Normal

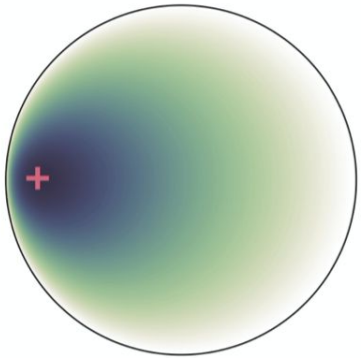
$$\frac{d\nu^{\mathrm{R}}(\boldsymbol{x}|\boldsymbol{\mu},\sigma^2)}{d\mathcal{M}(\boldsymbol{x})} \propto \exp\left(-\frac{d_p^c(\boldsymbol{\mu},\boldsymbol{x})^2}{2\sigma^2}\right)$$



# Generalizing the Normal distribution

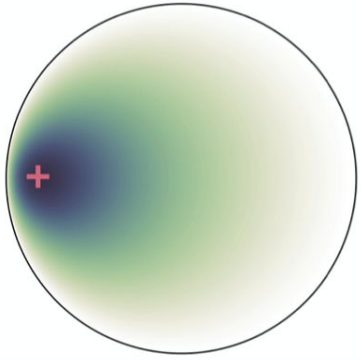
## Riemannian Normal

$$\frac{d\nu^{\mathrm{R}}(\boldsymbol{x}|\boldsymbol{\mu},\sigma^2)}{d\mathcal{M}(\boldsymbol{x})} \propto \exp\left(-\frac{d_p^c(\boldsymbol{\mu},\boldsymbol{x})^2}{2\sigma^2}\right)$$

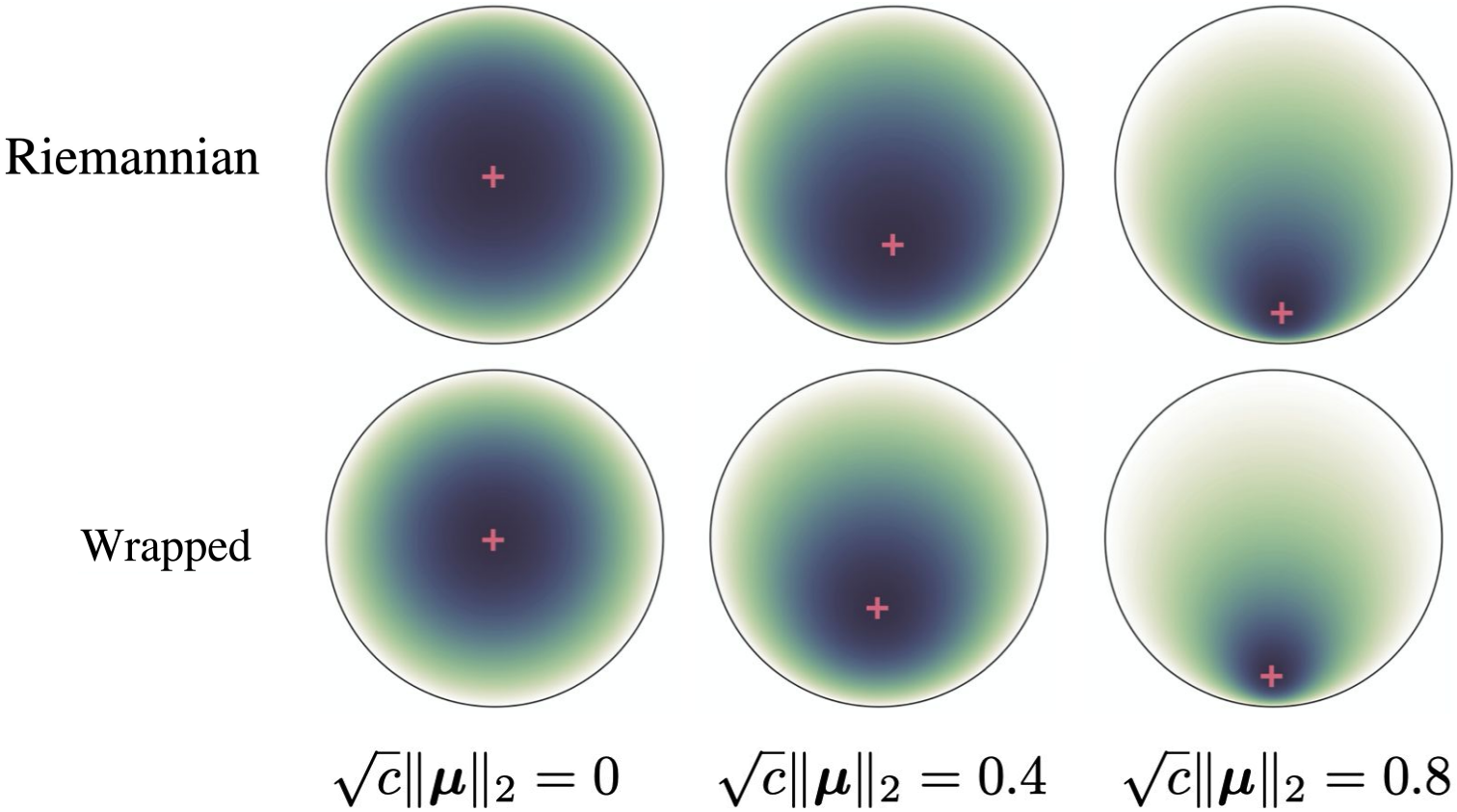


## Wrapped Normal

$$\boldsymbol{x} = \exp_{\boldsymbol{\mu}}^c\left(\frac{\boldsymbol{v}}{\lambda_{\boldsymbol{\mu}}^c}\right) \quad \boldsymbol{v} \sim \mathcal{N}(\cdot|\mathbf{0},\sigma^2)$$



# Generalizing the Normal distribution



# Reparametrization “trick”

## Riemannian Normal

$$\frac{d\nu^{\mathrm{R}}(\boldsymbol{x}|\boldsymbol{\mu},\sigma^2)}{d\mathcal{M}(\boldsymbol{x})} \propto \exp\left(-\frac{d_p^c(\boldsymbol{\mu},\boldsymbol{x})^2}{2\sigma^2}\right)$$

## Wrapped Normal

$$\boldsymbol{x} = \exp_{\boldsymbol{\mu}}^c\left(\frac{\boldsymbol{v}}{\lambda_{\boldsymbol{\mu}}^c}\right) \quad \boldsymbol{v} \sim \mathcal{N}(\cdot|\mathbf{0},\sigma^2)$$

Reparametrization

$$\boldsymbol{x} = \exp_{\boldsymbol{\mu}}^c\left(\frac{r}{\lambda_{\boldsymbol{\mu}}^c}\boldsymbol{\alpha}\right) \quad \texttt{with} \quad r = d_p^c(\boldsymbol{\mu},\boldsymbol{x})$$

# Reparametrization “trick”

## Riemannian Normal

$$\frac{d\nu^{\mathrm{R}}(\boldsymbol{x}|\boldsymbol{\mu},\sigma^2)}{d\mathcal{M}(\boldsymbol{x})} \propto \exp\left(-\frac{d_p^c(\boldsymbol{\mu},\boldsymbol{x})^2}{2\sigma^2}\right)$$

Reparametrization

$$\boldsymbol{x} = \exp_{\boldsymbol{\mu}}^c\left(\frac{r}{\lambda_{\boldsymbol{\mu}}^c}\boldsymbol{\alpha}\right) \quad \texttt{with} \quad r = d_p^c(\boldsymbol{\mu},\boldsymbol{x})$$

$$\rho^{\mathrm{R}}(r) \propto \mathbb{1}_{\mathbb{R}_+}(r) e^{-\frac{r^2}{2\sigma^2}} \left(\frac{\sinh(\sqrt{c}r)}{\sqrt{c}}\right)^{d-1}$$

## Wrapped Normal

$$\boldsymbol{x} = \exp_{\boldsymbol{\mu}}^c\left(\frac{\boldsymbol{v}}{\lambda_{\boldsymbol{\mu}}^c}\right) \quad \boldsymbol{v} \sim \mathcal{N}(\cdot|\boldsymbol{0},\sigma^2)$$

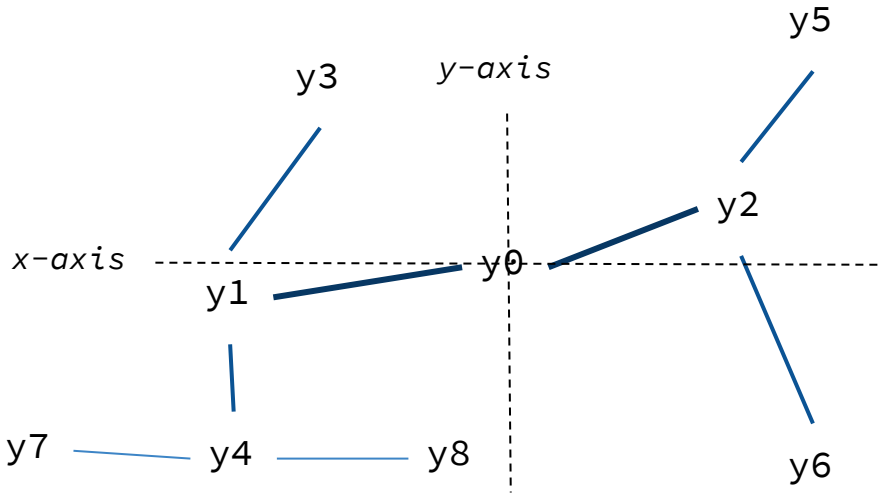
$$\rho^{\mathrm{W}}(r) \propto \mathbb{1}_{\mathbb{R}_+}(r)\, e^{-\frac{r^2}{2\sigma^2}} r^{d-1}$$

# Experiments

# Synthetic dataset: Branching diffusion process

Nodes $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) \in \mathbb{R}^n$

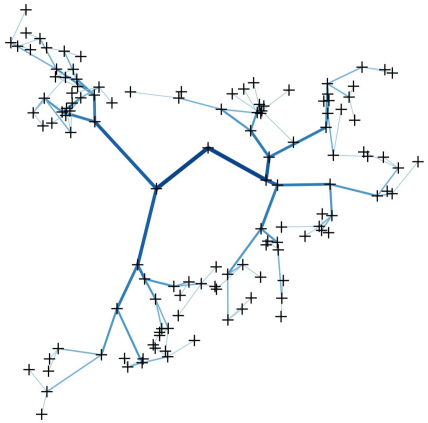
$\boldsymbol{y}_i \sim \mathcal{N}\left(\cdot \,|\boldsymbol{y}_{\pi(i)}, \sigma_0^2\right) \quad \forall i \in 1, \ldots, N$



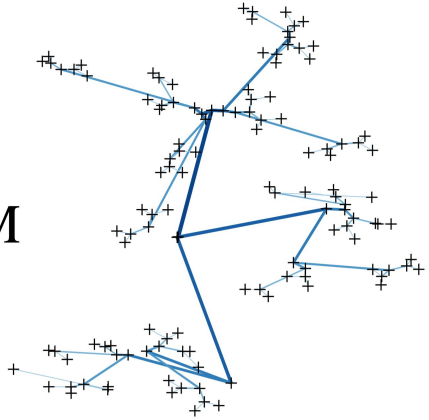

| | | Models | | | | | |
|---|---|---|---|---|---|---|---|
| | $\sigma_0$ | $\mathcal{N}$-VAE | $\mathcal{P}^{0.1}$-VAE | $\mathcal{P}^{0.3}$-VAE | $\mathcal{P}^{0.8}$-VAE | $\mathcal{P}^{1.0}$-VAE | $\mathcal{P}^{1.2}$-VAE |
| $\mathcal{L}_{\text{IWAE}}$ | 1 | $57.1_{\pm 0.2}$ | $57.1_{\pm 0.2}$ | $57.2_{\pm 0.2}$ | $56.9_{\pm 0.2}$ | $56.7_{\pm 0.2}$ | $56.6_{\pm 0.2}$ |
| $\mathcal{L}_{\text{IWAE}}$ | 1.7 | $57.0_{\pm 0.2}$ | $56.8_{\pm 0.2}$ | $56.6_{\pm 0.2}$ | $55.9_{\pm 0.2}$ | $55.7_{\pm 0.2}$ | $\mathbf{55.6}_{\pm 0.2}$ |

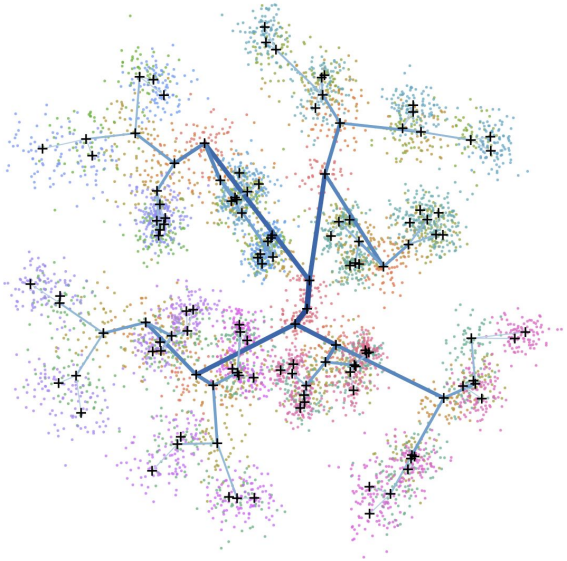
# Synthetic dataset: Branching diffusion process

PCA

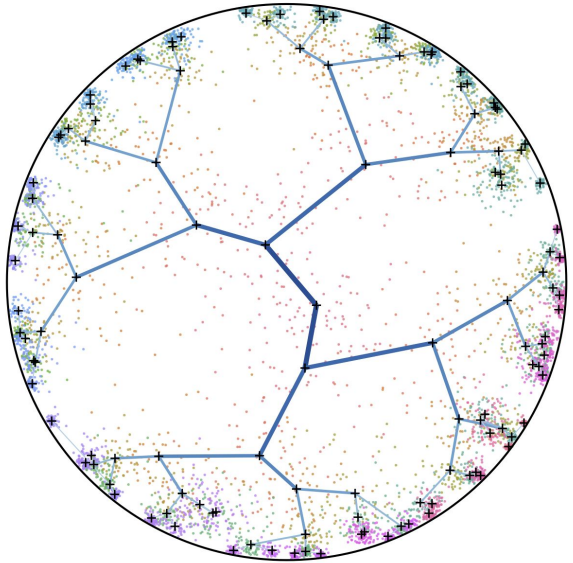


GPLVM



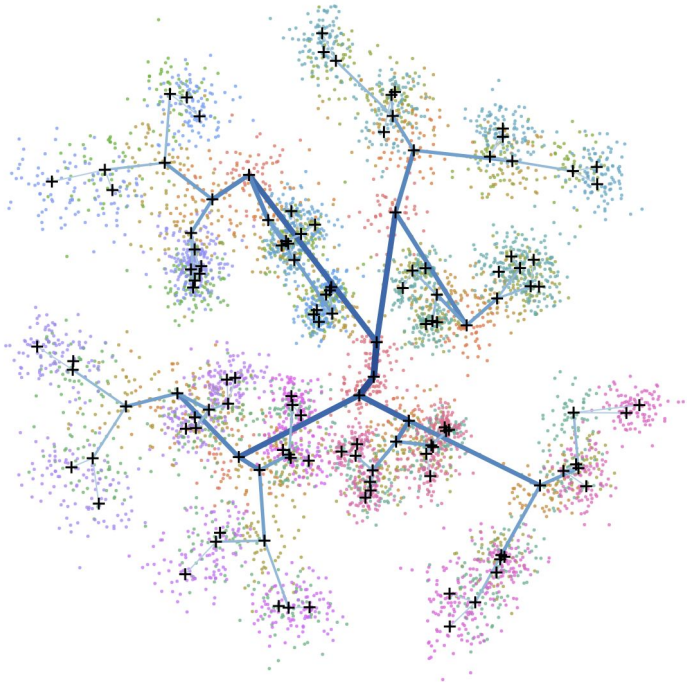


$\mathcal{N}$-VAE

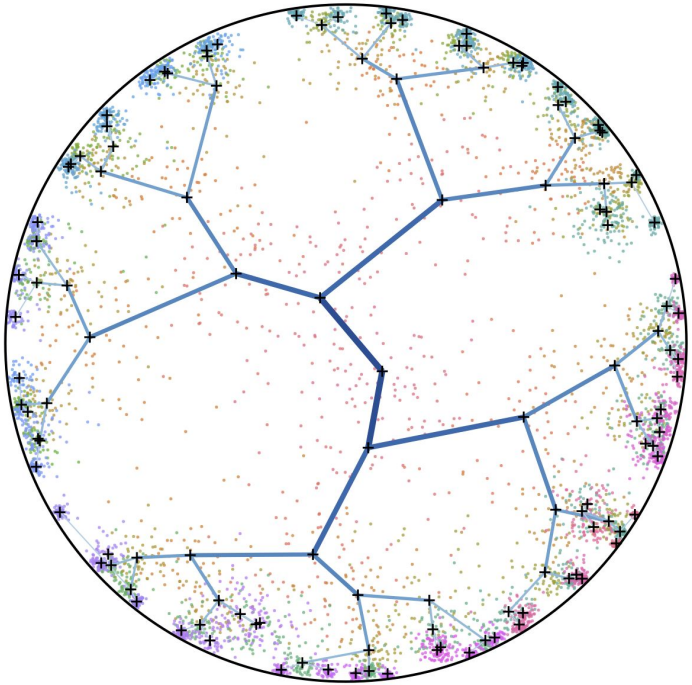


$\mathcal{P}^1$-VAE

# Synthetic dataset: Branching diffusion process

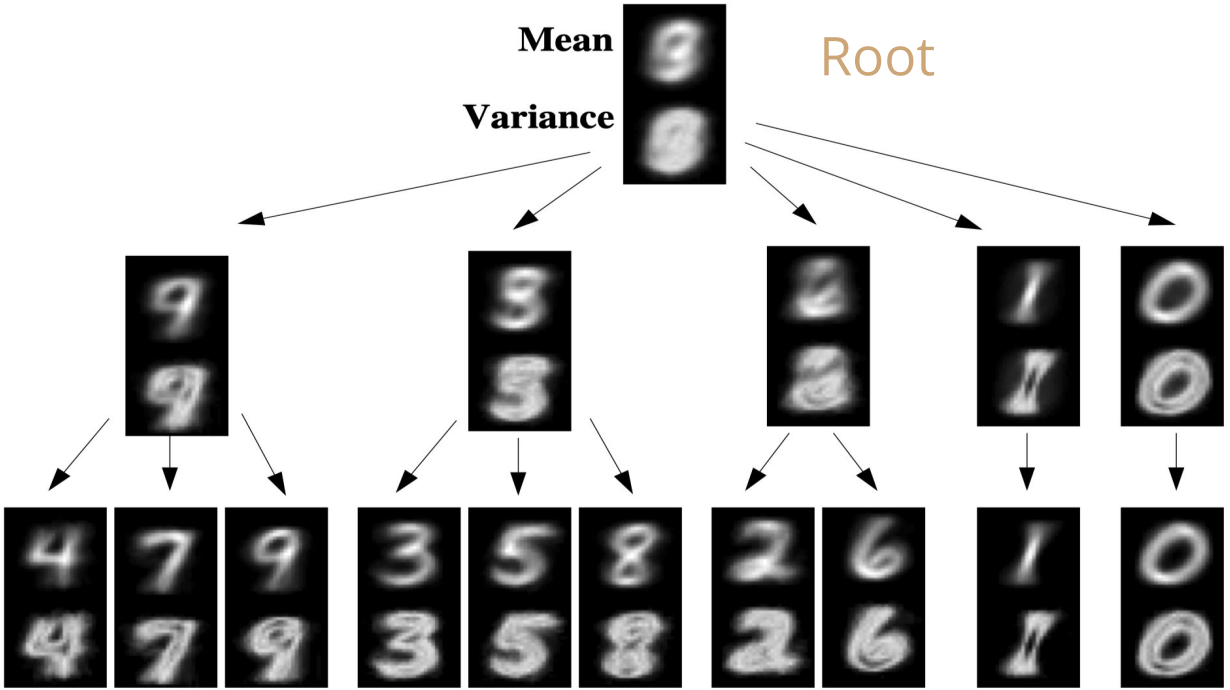


$\mathcal{N}$-VAE

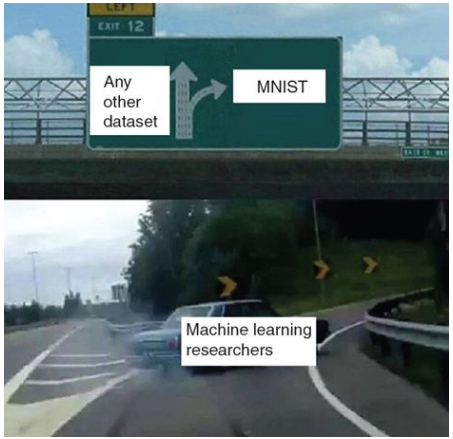


$\mathcal{P}^1$-VAE

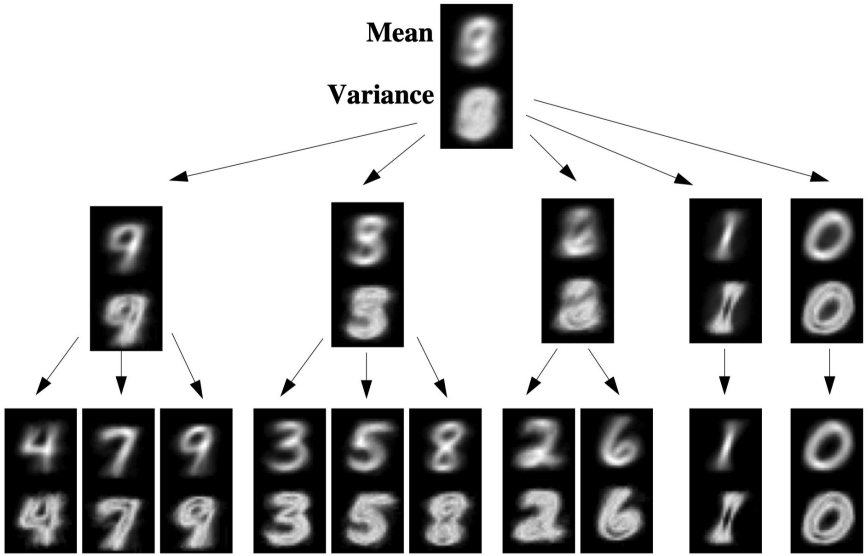
# MNIST: Handwritten digits

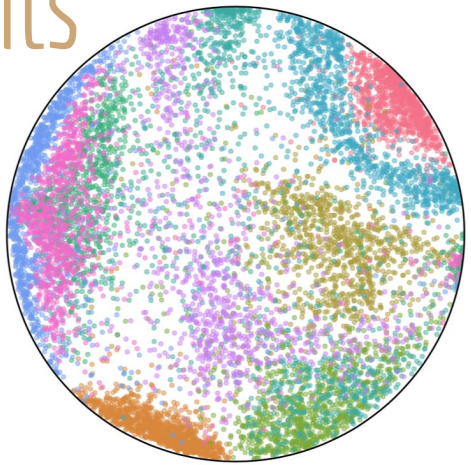


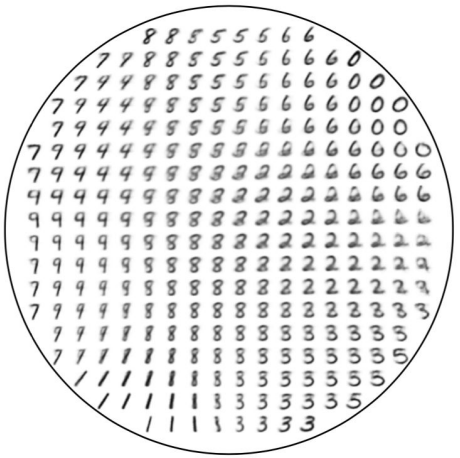

From Salakhutdinov et al., One-Shot Learning with a Hierarchical Nonparametric Bayesian Model

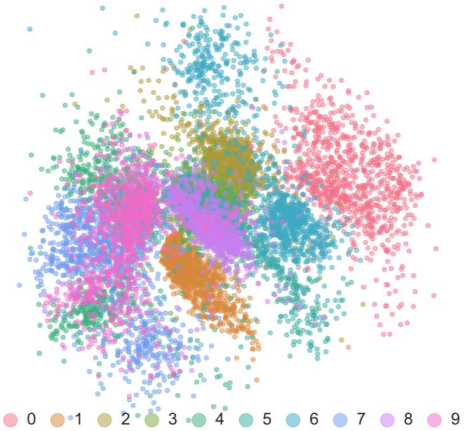


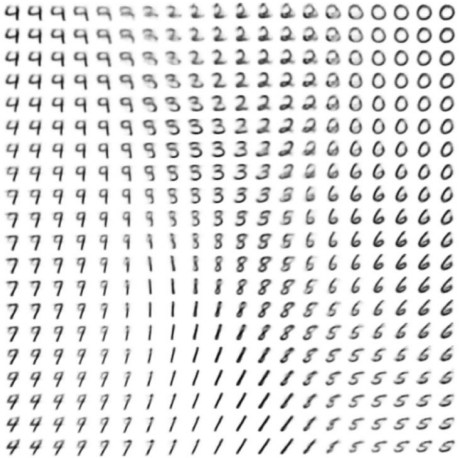
# MNIST: Handwritten digits







# MNIST: Handwritten digits

| | | Dimensionality | | | |
|---|---|---|---|---|---|
| | c | 2 | 5 | 10 | 20 |
| $\mathcal{N}$-VAE | (0) | $144.5_{\pm 0.4}$ | $114.7_{\pm 0.1}$ | $100.2_{\pm 0.1}$ | $97.6_{\pm 0.1}$ |
| $\mathcal{P}$-VAE (Wrapped) | 0.1 | $143.9_{\pm 0.5}$ | $115.5_{\pm 0.3}$ | $100.2_{\pm 0.1}$ | $97.2_{\pm 0.1}$ |
| | 0.2 | $144.2_{\pm 0.5}$ | $115.3_{\pm 0.3}$ | $100.0_{\pm 0.1}$ | $97.1_{\pm 0.1}$ |
| | 0.7 | $143.8_{\pm 0.6}$ | $115.1_{\pm 0.3}$ | $100.2_{\pm 0.1}$ | $97.5_{\pm 0.1}$ |
| | 1.4 | $144.0_{\pm 0.6}$ | $114.7_{\pm 0.1}$ | $100.7_{\pm 0.1}$ | $98.0_{\pm 0.1}$ |
| $\mathcal{P}$-VAE (Riemannian) | 0.1 | $143.7_{\pm 0.6}$ | $115.2_{\pm 0.2}$ | $99.9_{\pm 0.1}$ | $\mathbf{97.0}_{\pm 0.1}$ |
| | 0.2 | $143.8_{\pm 0.4}$ | $114.7_{\pm 0.3}$ | $\mathbf{99.7}_{\pm 0.1}$ | $97.4_{\pm 0.1}$ |
| | 0.7 | $143.1_{\pm 0.4}$ | $\mathbf{114.1}_{\pm 0.2}$ | $101.2_{\pm 0.2}$ | * |
| | 1.4 | $\mathbf{142.5}_{\pm 0.4}$ | $115.5_{\pm 0.3}$ | * | * |

| Digits | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{N}$-VAE | 89 | 97 | 81 | 75 | 59 | 43 | 89 | **78** | 68 | **57** | 73.6 |
| $\mathcal{P}^{1.4}$-VAE | **94** | 97 | **82** | **79** | **69** | **47** | **90** | 77 | 68 | 53 | **75.6** |

# MNIST: Handwritten digits

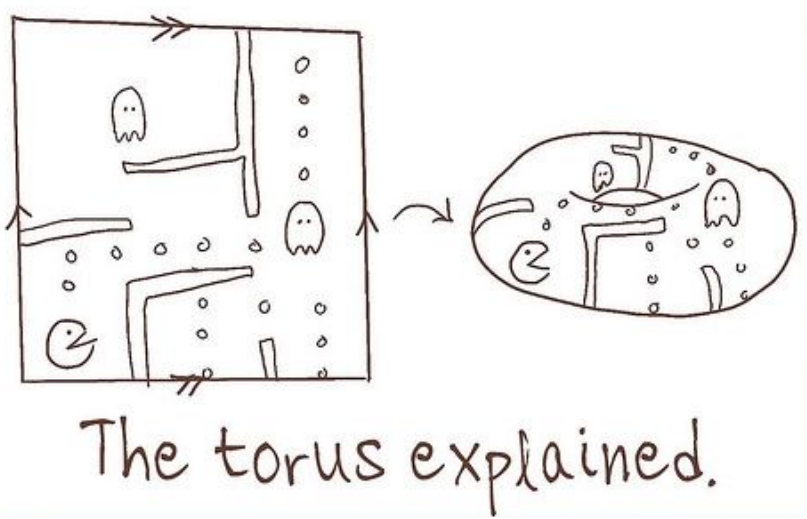
| | | Dimensionality | | | |
|---|---|---|---|---|---|
| | c | 2 | 5 | 10 | 20 |
| $\mathcal{N}$-**VAE** | (0) | $144.5_{\pm 0.4}$ | $114.7_{\pm 0.1}$ | $100.2_{\pm 0.1}$ | $97.6_{\pm 0.1}$ |
| $\mathcal{P}$-**VAE (Wrapped)** | 0.1 | $143.9_{\pm 0.5}$ | $115.5_{\pm 0.3}$ | $100.2_{\pm 0.1}$ | $97.2_{\pm 0.1}$ |
| | 0.2 | $144.2_{\pm 0.5}$ | $115.3_{\pm 0.3}$ | $100.0_{\pm 0.1}$ | $97.1_{\pm 0.1}$ |
| | 0.7 | $143.8_{\pm 0.6}$ | $115.1_{\pm 0.3}$ | $100.2_{\pm 0.1}$ | $97.5_{\pm 0.1}$ |
| | 1.4 | $144.0_{\pm 0.6}$ | $114.7_{\pm 0.1}$ | $100.7_{\pm 0.1}$ | $98.0_{\pm 0.1}$ |
| $\mathcal{P}$-**VAE (Riemannian)** | 0.1 | $143.7_{\pm 0.6}$ | $115.2_{\pm 0.2}$ | $99.9_{\pm 0.1}$ | $\mathbf{97.0}_{\pm 0.1}$ |
| | 0.2 | $143.8_{\pm 0.4}$ | $114.7_{\pm 0.3}$ | $\mathbf{99.7}_{\pm 0.1}$ | $97.4_{\pm 0.1}$ |
| | 0.7 | $143.1_{\pm 0.4}$ | $\mathbf{114.1}_{\pm 0.2}$ | $101.2_{\pm 0.2}$ | * |
| | 1.4 | $\mathbf{142.5}_{\pm 0.4}$ | $115.5_{\pm 0.3}$ | * | * |

# Recap

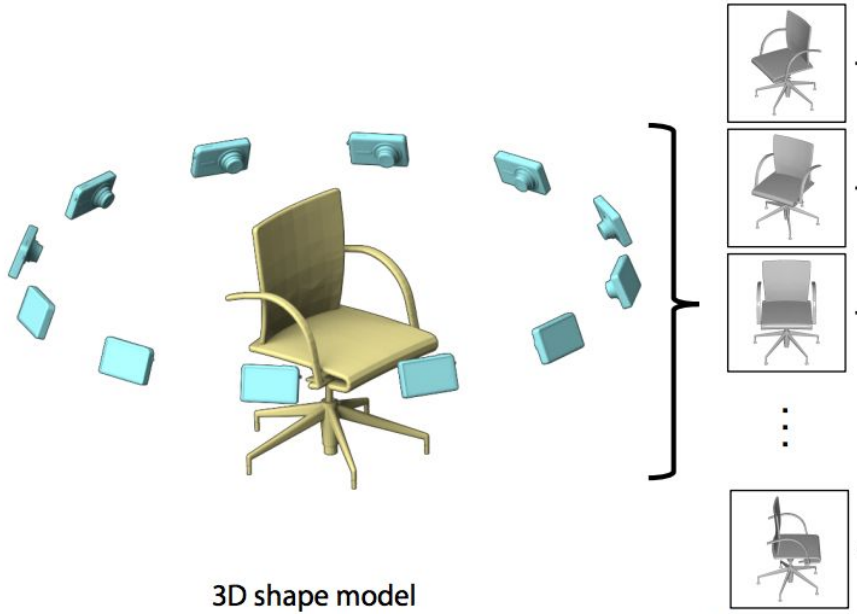
- Deep generative model for hierarchical data
- Leverage a hyperbolic latent space to induce hierarchical representations
- Mostly useful for low latent dimensionality

# Geometry as a manifold constraint

# Data with loops



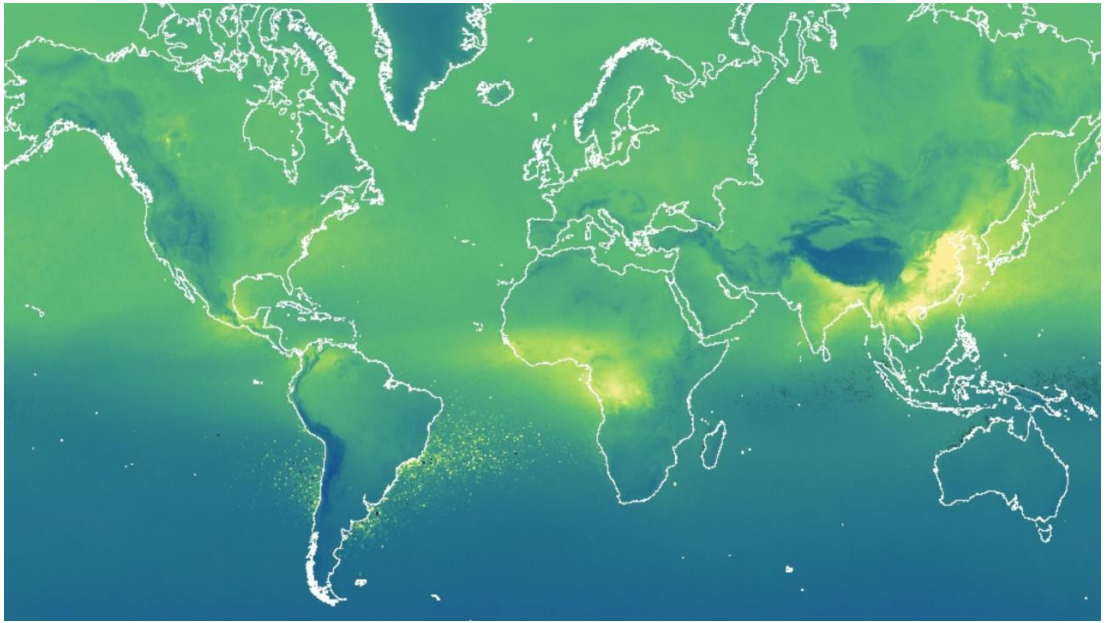

From Human Mathematics Blog, 2009




From Connelly Barnes, 3D Reconstruction and Understanding, 2017

# Earth science

$$\mathcal{M} = \mathbb{S}^2$$



*Concentrations of Carbon monoxide (CO) and water vapor.*
*From The Sentinel-5 Precursor, European Space Agency.*

# Riemannian Continuous Normalizing Flows

Learning expressive probability distributions on manifolds

- Flows on manifolds as solution of ODEs
- Continuous change of variables for manifold-valued variables
- Vector field parametrization

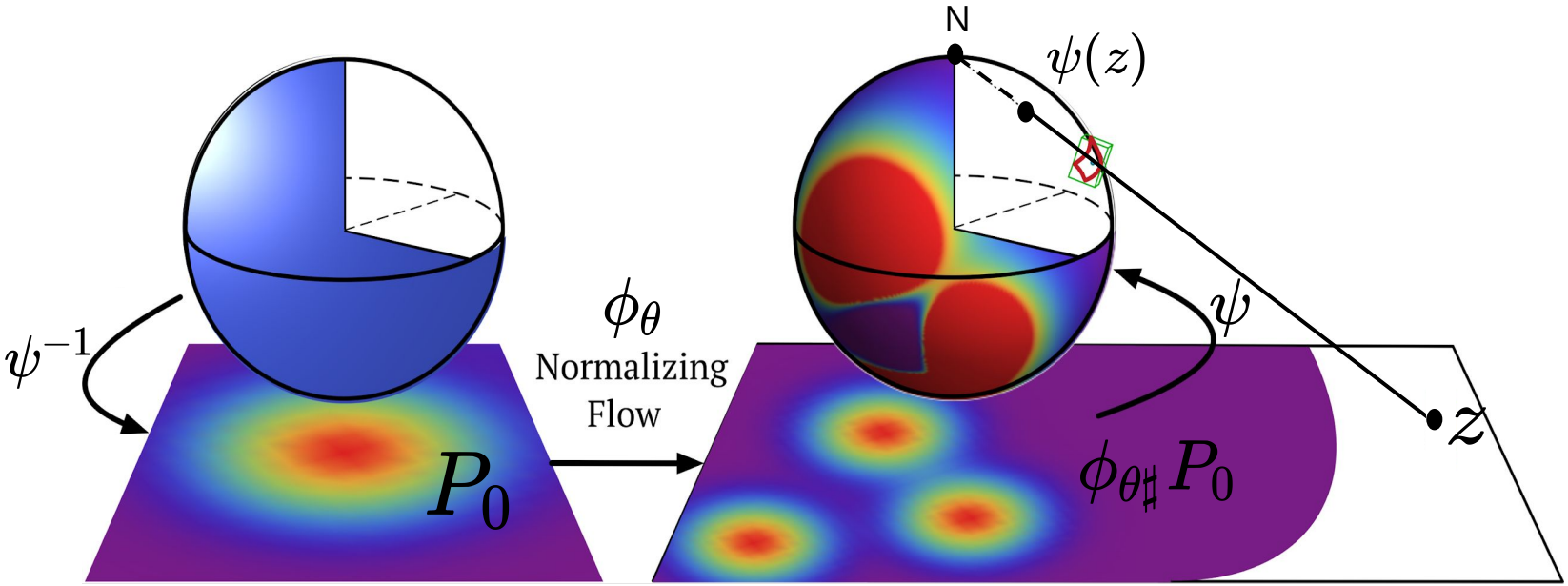
With Max Nickel

# Motivation

# Stereographic projection: An injection

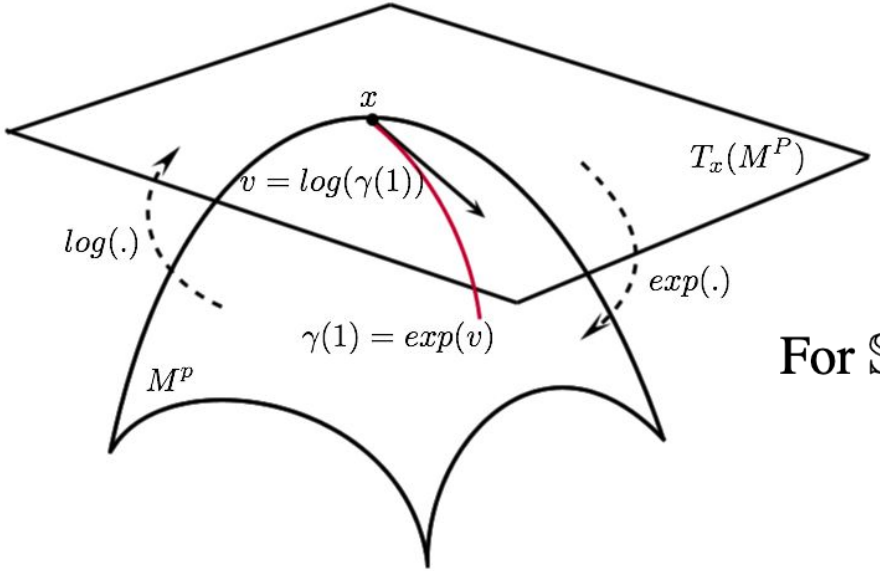
$$\psi : \mathbb{R}^d \to \mathcal{M}$$

$$P_\theta = (\psi \circ \phi_\theta)_\sharp P_0$$




*From Gemici et al, 2016*

# Exponential map: A surjection



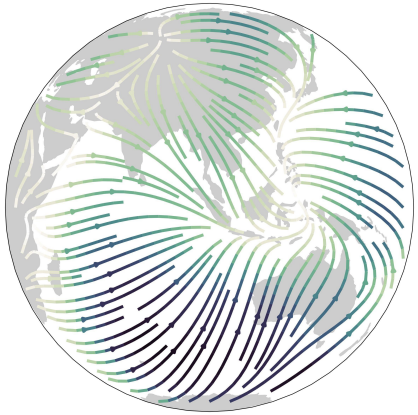

*From Matthieu Simeoni, 2013*

$$\exp_x : \mathcal{T}_x\mathcal{M} \cong \mathbb{R}^d \to \mathcal{M}$$

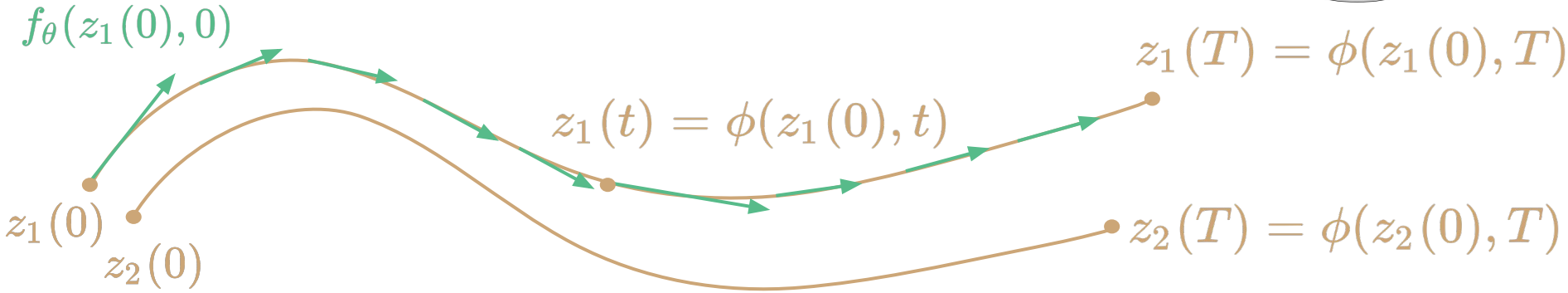
$$\text{For } \mathbb{S}^d, \exp_x(v + 2k\pi) = \exp_x(v)\ \forall k \in \mathbb{Z}$$

# Model

# Ordinary Differential Equation & Flow



**ODE** $\frac{dz(t)}{dt} = f_\theta(z(t), t)$ vector field



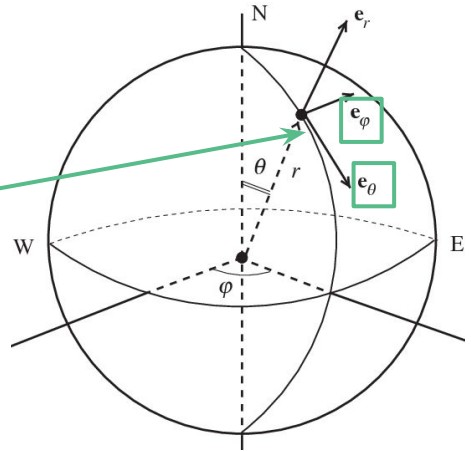

**Flow** $\phi(z) := \phi(z, T)$ $\phi : \mathcal{M} \to \mathcal{M}$ $\mathcal{C}^1$ - diffeomorphism if $f_\theta$ is $\mathcal{C}^1$ and bounded

**Model** $P_\theta = \phi_\sharp P_0$

# Vector field parametrization: local coordinates

**Basis**

$$f_\theta(z,t) = \sum_i^d f_\theta^i(z,t)\frac{\partial}{\partial z^i}(z)$$




**Divergence**

$$div(f_\theta(z,t)) = \frac{1}{\sqrt{|G(z)|}}\sum_i^d \frac{\partial}{\partial z^i}\left(\sqrt{|G(z)|}f_\theta^i(z,t)\right)$$

**Example**

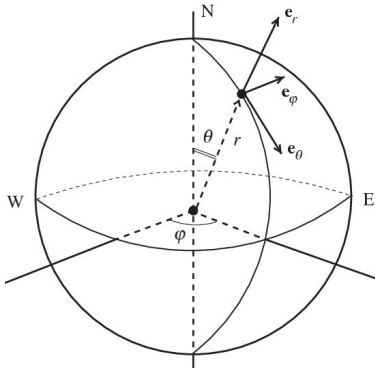
e.g. $z = (\theta, \phi)$ polar coordinates for $\mathbb{S}^2$

$$\sqrt{|G(\theta,\varphi)|} = \sin(\theta)$$

# Vector field parametrization: local coordinates

**Basis**

$$f_\theta(z,t) = \sum_i^d f_\theta^i(z,t)\frac{\partial}{\partial z^i}(z)$$




**Divergence**

$$div(f_\theta(z,t)) = \frac{1}{\sqrt{|G(z)|}}\sum_i^d \frac{\partial}{\partial z^i}\left(\sqrt{|G(z)|}f_\theta^i(z,t)\right)$$

**Example**

$$z(\theta,\varphi) = (\sin(\theta)\cos(\varphi), \sin(\theta)\sin(\varphi), \cos(\theta))$$

$$\sqrt{|G(\theta,\varphi)|} = \sin(\theta)$$

$$\mathrm{div}(f(\theta,\varphi)) = \frac{1}{\sin(\theta)}\frac{\partial}{\partial\theta}\left(\sin(\theta)f^\theta(\theta,\varphi)\right) + \frac{1}{\sin(\theta)}\frac{\partial}{\partial\varphi}\left(f^\varphi(\theta,\varphi)\right)$$

# Likelihood

## Continuous change of variable

$$\frac{\partial \log p_\theta(\boldsymbol{z}(t))}{\partial t} = -\operatorname{div}(\boldsymbol{f}_\theta(\boldsymbol{z}(t), t)) = -|G(\boldsymbol{z}(t))|^{-\frac{1}{2}} \operatorname{tr}\left(\frac{\partial \sqrt{|G(\boldsymbol{z}(t))|}\boldsymbol{f}_\theta(\boldsymbol{z}(t), t)}{\partial \boldsymbol{z}}\right) \quad (1)$$

## Stochastic estimator

$$\operatorname{div}(\boldsymbol{f}_\theta(\boldsymbol{z}(t), t)) = |G(\boldsymbol{z}(t))|^{-\frac{1}{2}} \mathbb{E}_{p(\boldsymbol{\epsilon})}\left[\boldsymbol{\epsilon}^\intercal \frac{\partial \sqrt{|G(\boldsymbol{z}(t))|}\boldsymbol{f}_\theta(\boldsymbol{z}(t), t)}{\partial \boldsymbol{z}} \boldsymbol{\epsilon}\right]$$

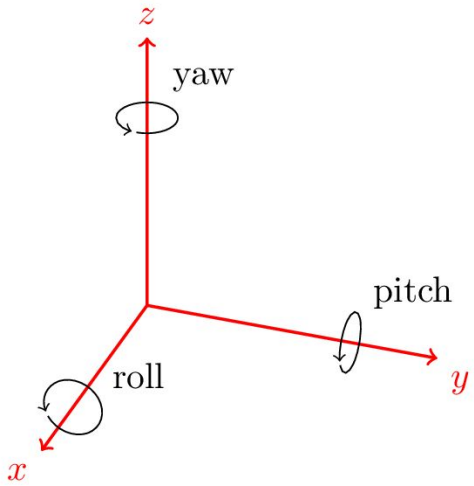
# Vector field parametrization: divergence-free basis

**Basis** $f_i(z) = \frac{d}{dt}\Big|_{t=0} exp(t\,\xi_i) \cdot z$

→ $div(f_i(z)) = 0$

→ $(f_i(z)\,; i = 1, \ldots, n)$ span the tangent space $T_z M$

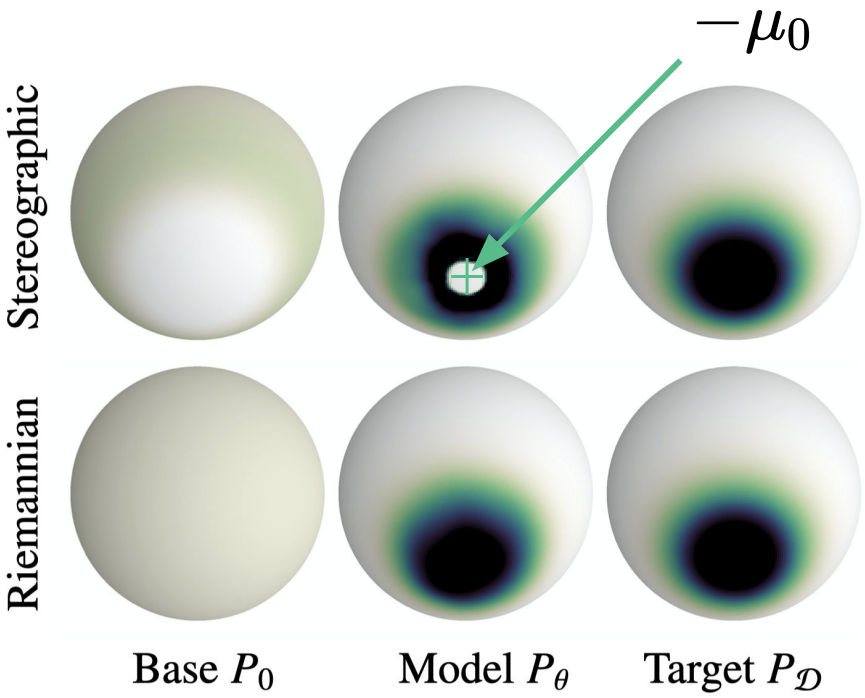
→ $f_\theta(z,t) = \sum_i^n f_\theta^i(z,t) f_i(z)$




**Divergence** $div(f_\theta(z,t)) = \sum_i^n \langle \frac{\partial}{\partial z} f^i(z,t), f_i(z) \rangle$

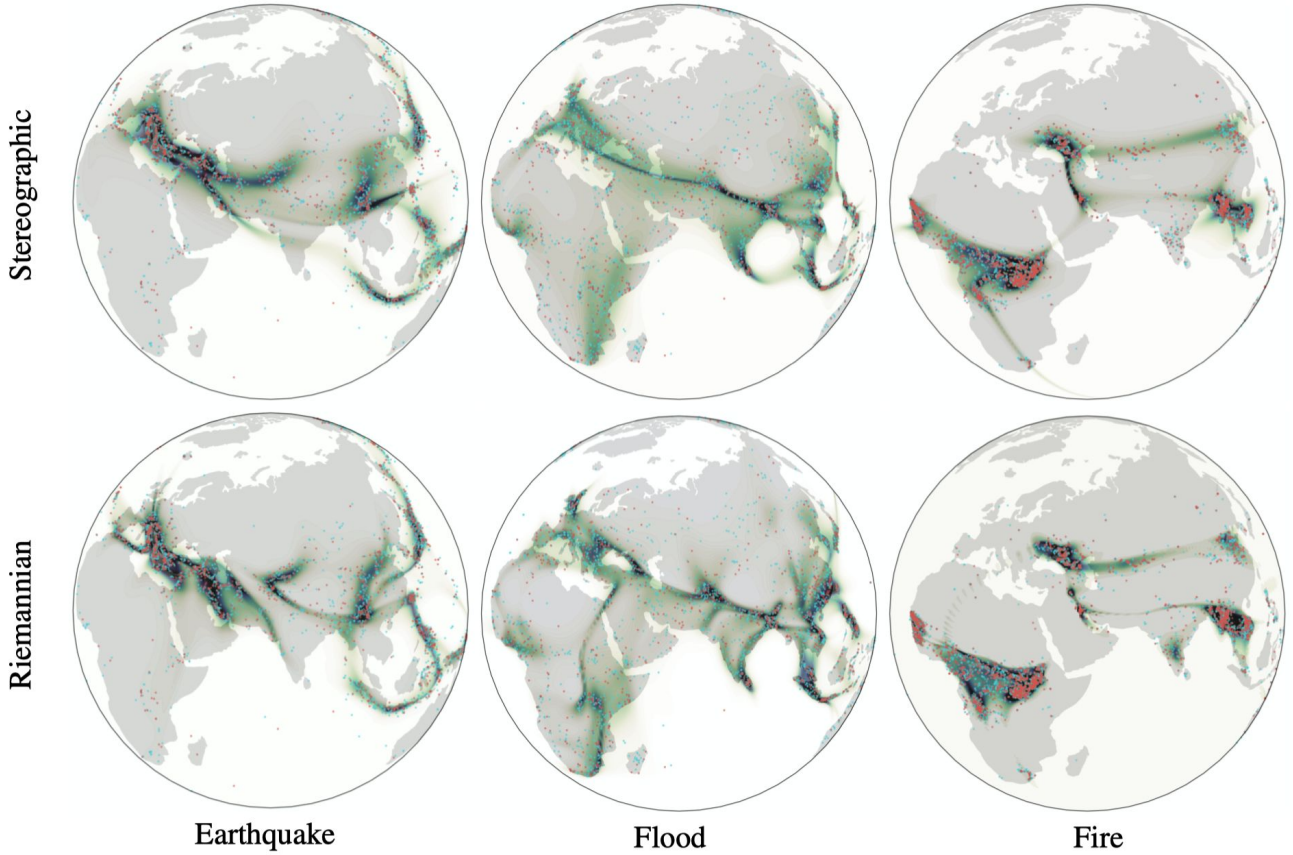
# Experiments

# Experiments: Von Mises-Fisher

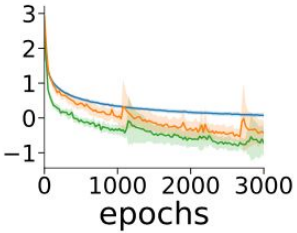
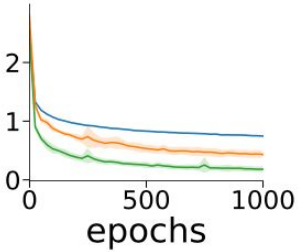
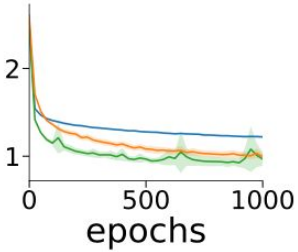
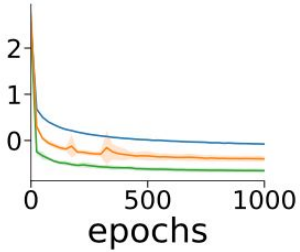



| Loss | model $\kappa$ | Stereographic | Riemannian |
|---|---|---|---|
| $\mathcal{L}^{\text{Like}}$ | 100 | $63.60_{\pm 3.56}$ | $\mathbf{-1.78}_{\pm 0.01}$ |
| | 50 | $32.68_{\pm 3.15}$ | $\mathbf{-1.09}_{\pm 0.01}$ |
| | 10 | $6.45_{\pm 2.42}$ | $\mathbf{0.52}_{\pm 0.01}$ |
| $\mathcal{L}^{\text{KL}}$ | 100 | $1.56_{\pm 0.34}$ | $\mathbf{0.04}_{\pm 0.02}$ |
| | 50 | $0.68_{\pm 0.16}$ | $\mathbf{0.03}_{\pm 0.02}$ |
| | 10 | $0.12_{\pm 0.01}$ | $\mathbf{0.01}_{\pm 0.00}$ |

# Experimental Results: Earth Science



# Experimental Results: Earth Science

| | Volcano | Earthquake | Flood | Fire |
|---|---|---|---|---|
| **Mixture vMF** ■ | $-0.31_{\pm 0.07}$ | $0.59_{\pm 0.01}$ | $1.09_{\pm 0.01}$ | $-0.23_{\pm 0.02}$ |
| **Stereographic** ■ | $-0.64_{\pm 0.20}$ | $0.43_{\pm 0.06}$ | $0.99_{\pm 0.04}$ | $-0.40_{\pm 0.06}$ |
| **Riemannian** ■ | $-0.97_{\pm 0.15}$ | $\mathbf{0.18}_{\pm 0.05}$ | $\mathbf{0.90}_{\pm 0.03}$ | $\mathbf{-0.66}_{\pm 0.05}$ |
| Learning curves |  |  |  |  |
| Data size | 829 | 6124 | 4877 | 12810 |

# Recap

- Extended neural ODEs to smooth manifolds
- The generated flow is a diffeomorphism
- Generalize better than the non-surjective stereographic approach

# Recap: Geometry and Deep generative models

- *Geometry as an inductive bias*
  - Hyperbolic geometry for hierarchical data

- *Geometry as a manifold constraint*
  - Normalizing Flows for manifold-valued data

# Thank you for your attention!

Questions?