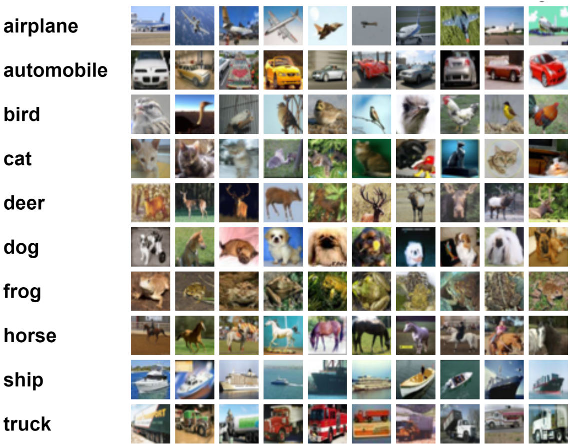
Introduction

Geometry of Data

August 22, 2023

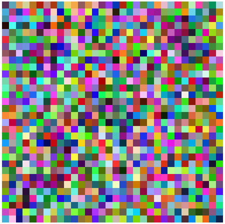
CIFAR-10



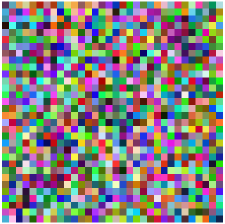
$32 \times 32 \times 3 = 3{,}072$ dimensions

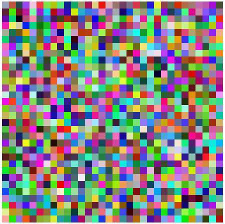
10 classes

Uniform Random Images

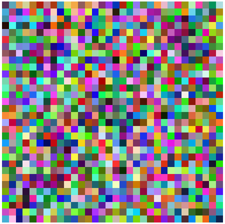


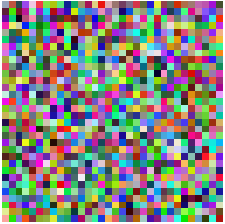
Uniform Random Images

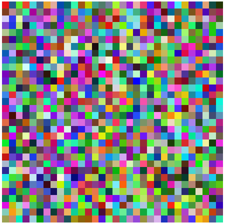




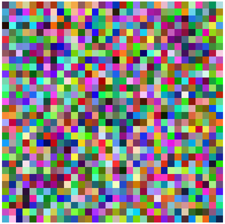
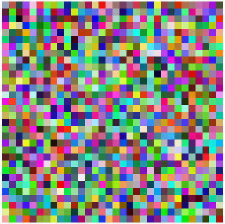
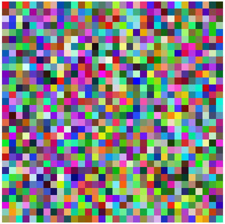
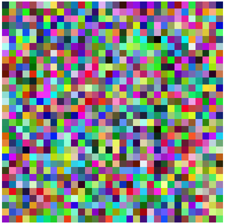
Uniform Random Images



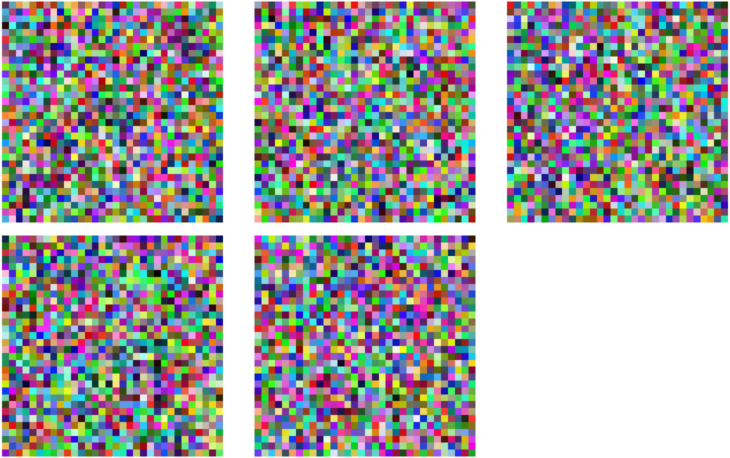




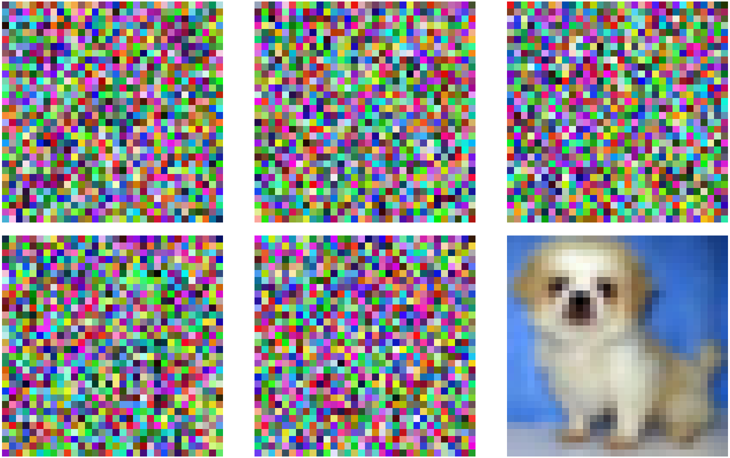
Uniform Random Images

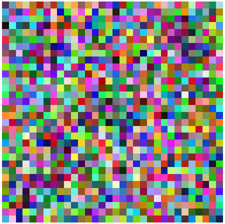
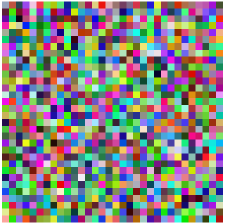
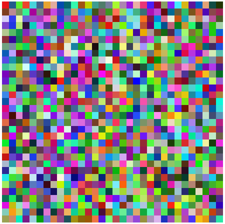
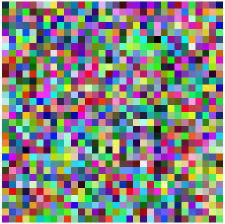
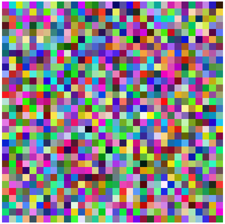
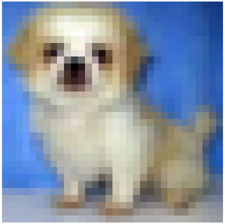
Uniform Random Images



Uniform Random Images



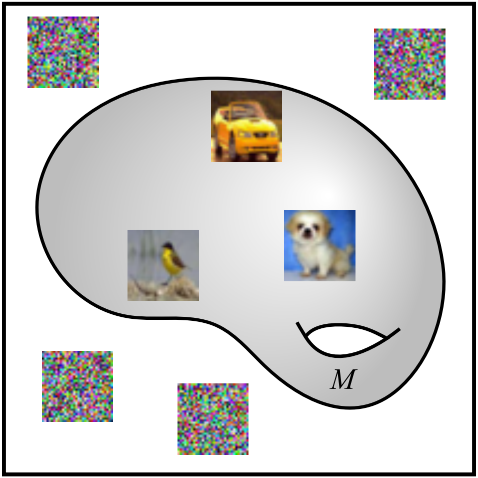
Uniform Random Images

just kidding!

Manifold Hypothesis

Real data lie near lower-dimensional manifolds



Manifold Learning

Manifold Learning

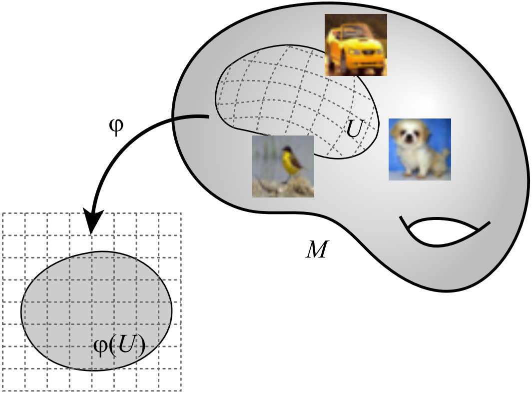
- Learn a model/representation for the data manifold

Manifold Learning

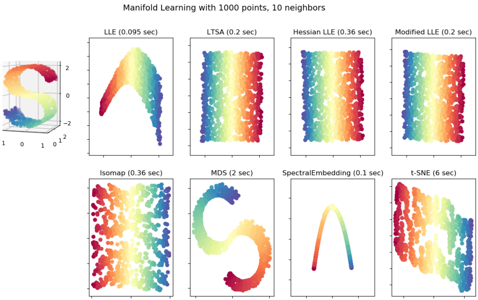
- Learn a model/representation for the data manifold
- Often involves finding a flat coordinate chart

Manifold Learning

- Learn a model/representation for the data manifold
- Often involves finding a flat coordinate chart

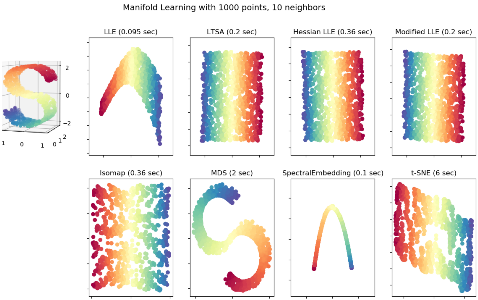


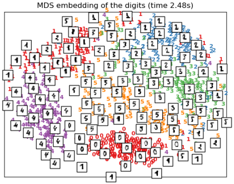
Manifold Learning



From scikit-learn.org

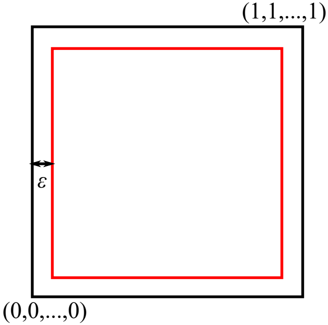
Manifold Learning





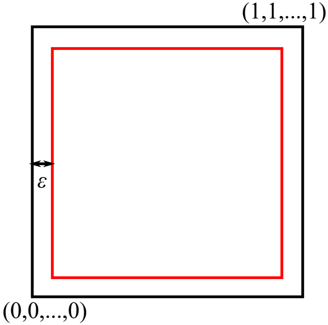
From scikit-learn.org

Volumes in High Dimensions



What is the volume of the unit d-cube shrunk by some small amount in each dimension?

Volumes in High Dimensions

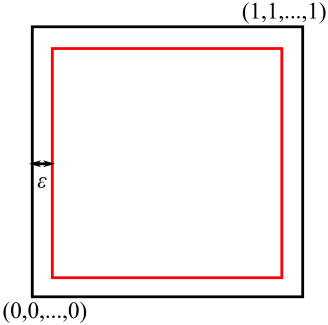


What is the volume of the unit d-cube shrunk by some small amount in each dimension?

$$V = (1 - 2\epsilon)^d$$

Approaches 0 as $d \to \infty$

Volumes in High Dimensions



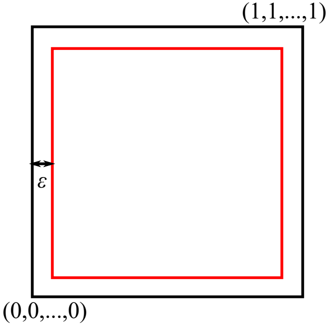
What is the volume of the unit d-cube shrunk by some small amount in each dimension?

$$V = (1 - 2\epsilon)^d$$

Approaches 0 as $d \to \infty$

Example: $256 \times 256 \times 3$ images, $\epsilon = \frac{1}{256}$

Volumes in High Dimensions



What is the volume of the unit d-cube shrunk by some small amount in each dimension?

$$V = (1 - 2\epsilon)^d$$

Approaches 0 as $d \to \infty$

Example: $256 \times 256 \times 3$ images, $\epsilon = \frac{1}{256}$

$$V \approx 2.0 \times 10^{-670}$$

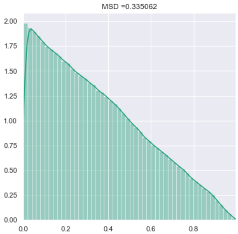
Distances in High Dimensions

Sample two points uniformly from the unit d-cube:
$X, Y \sim \mathrm{Unif}([0,1]^d)$

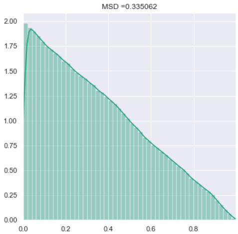
Distances in High Dimensions

Sample two points uniformly from the unit d-cube:
$X, Y \sim \mathrm{Unif}([0,1]^d)$

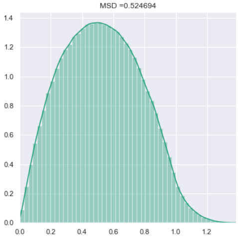
What is the distribution of the distance between them?
$D = \|X - Y\|$



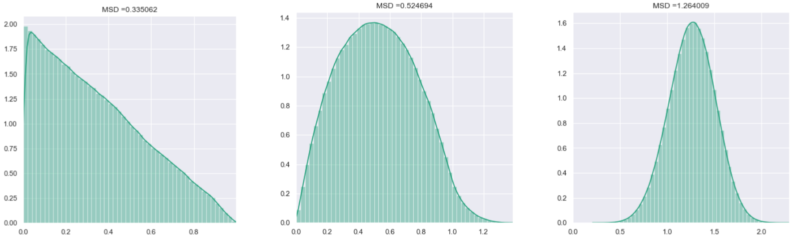
$$d = 1$$


MSD =0.335062
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8

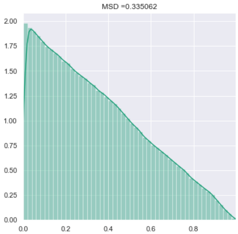
$d = 1$


MSD =0.524694
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2

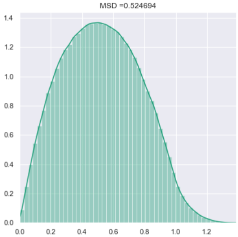
$d = 2$



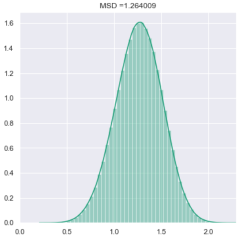
$d = 1$ $\qquad$ $d = 2$ $\qquad$ $d = 10$



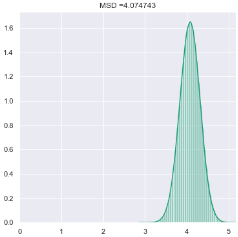
$d = 1$



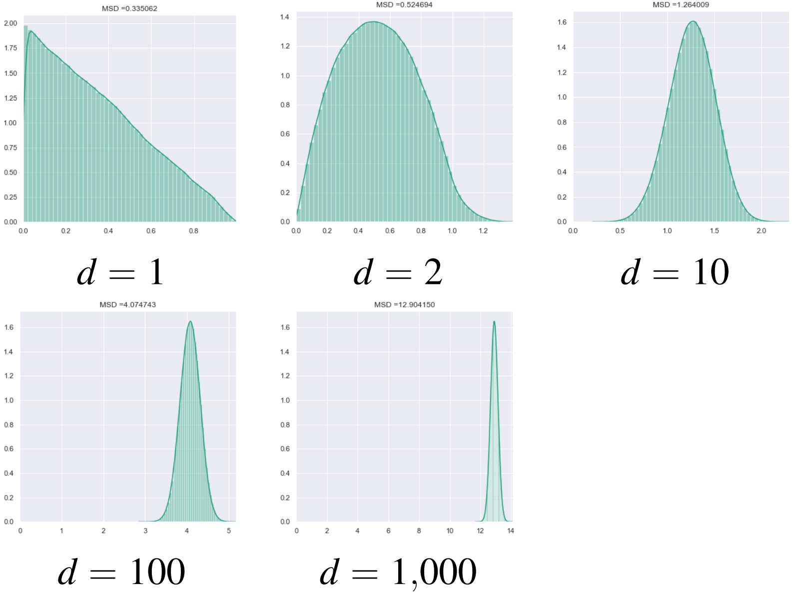
$d = 2$

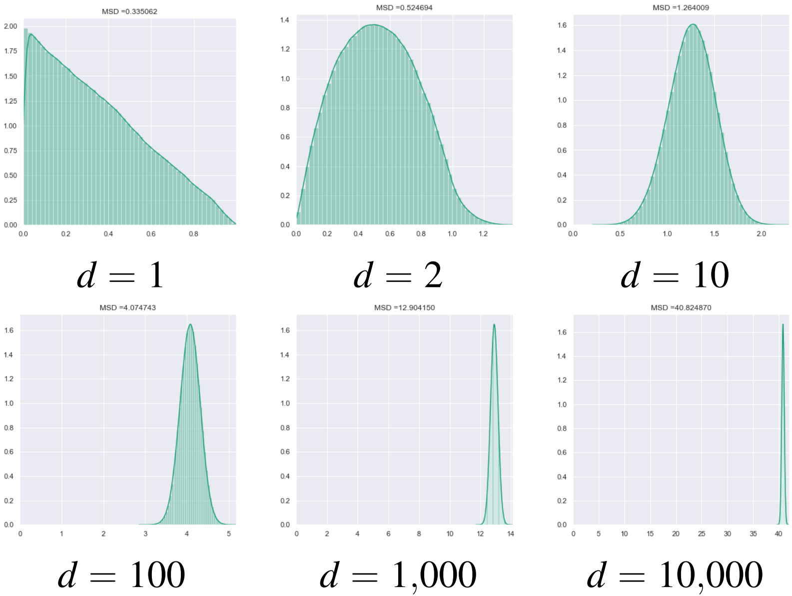


$d = 10$

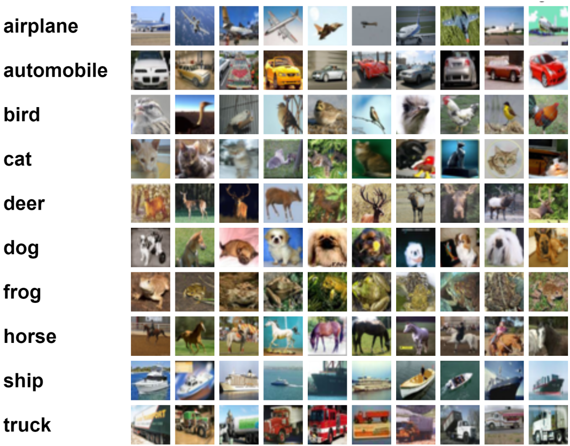


$d = 100$


MSD =0.335062
$d = 1$
MSD =0.524694
$d = 2$
MSD =1.264009
$d = 10$
MSD =4.074743
$d = 100$
MSD =12.904150
$d = 1{,}000$


MSD =0.335062
d = 1
MSD =0.524694
d = 2
MSD =1.264009
d = 10
MSD =4.074743
d = 100
MSD =12.904150
d = 1,000
MSD =40.824870
d = 10,000

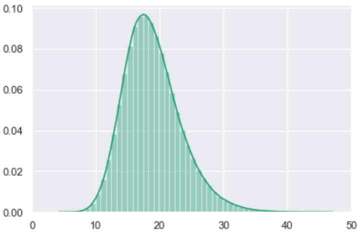
CIFAR-10



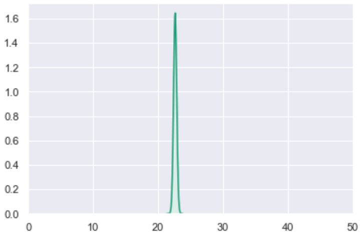
$32 \times 32 \times 3 = 3{,}072$ dimensions

10 classes

Distances in Real Data



CIFAR-10



$\mathrm{Unif}([0, 1]^{3072})$

Manifold-valued Data

- Manifold already known, not learned

Manifold-valued Data

- Manifold already known, not learned
- Manifold arises from natural non-linear constraints on data

Manifold-valued Data

- Manifold already known, not learned
- Manifold arises from natural non-linear constraints on data
- Linear data analyses (in fact, vector space operations) violate these constraints

Directional Data

Data living on a circle (S^1) or sphere (S^2), etc.

Directional Data

Data living on a circle (S^1) or sphere (S^2), etc.

- Orientation of molecules in protein structure

Directional Data

Data living on a circle (S^1) or sphere (S^2), etc.

- Orientation of molecules in protein structure
- Direction of robot or autonomous vehicle

Directional Data

Data living on a circle (S^1) or sphere (S^2), etc.

- Orientation of molecules in protein structure
- Direction of robot or autonomous vehicle
- Position on the earth

Directional Data

Data living on a circle (S^1) or sphere (S^2), etc.

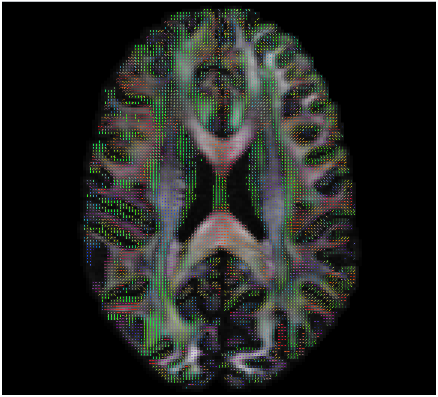
- Orientation of molecules in protein structure
- Direction of robot or autonomous vehicle
- Position on the earth
- Motion capture: orientation of joints

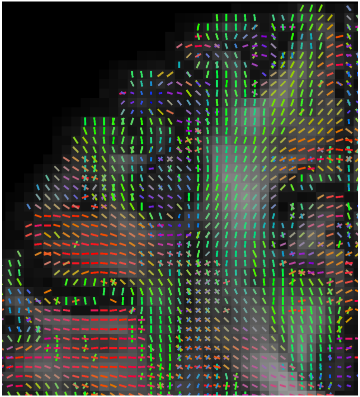
Directional Data

Data living on a circle (S^1) or sphere (S^2), etc.

- Orientation of molecules in protein structure
- Direction of robot or autonomous vehicle
- Position on the earth
- Motion capture: orientation of joints
- Time (time of day, day of the year, etc.)

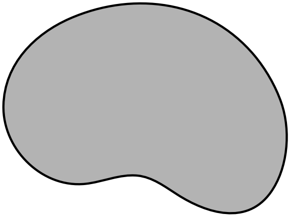
Directional Data: Diffusion MRI





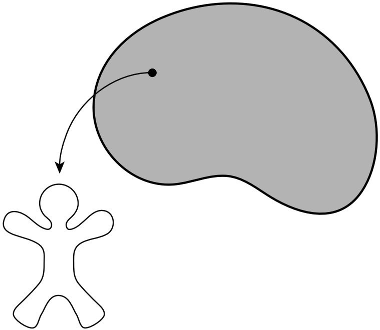
Voxel features are directions of axons in brain

Shape Manifolds



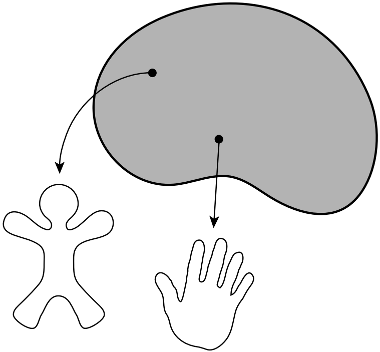
A shape is a point in a high-dimensional, nonlinear manifold, called a **shape space**.

Shape Manifolds



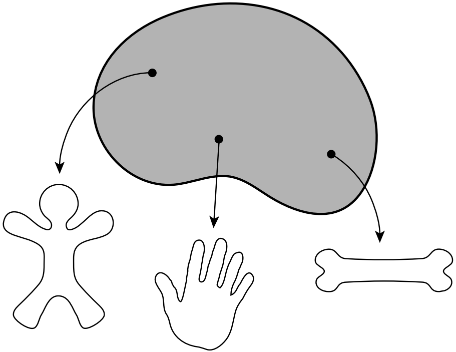
A shape is a point in a high-dimensional, nonlinear manifold, called a **shape space**.

Shape Manifolds



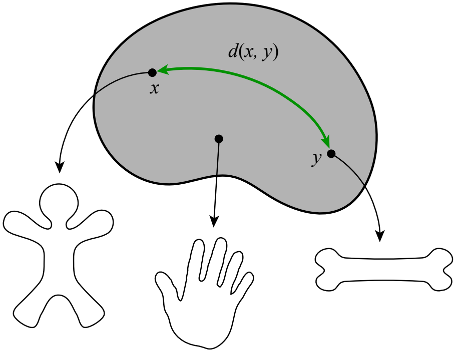
A shape is a point in a high-dimensional, nonlinear manifold, called a **shape space**.

Shape Manifolds



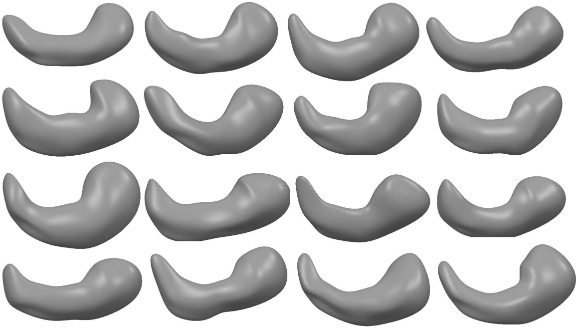
A shape is a point in a high-dimensional, nonlinear manifold, called a **shape space**.

Shape Manifolds

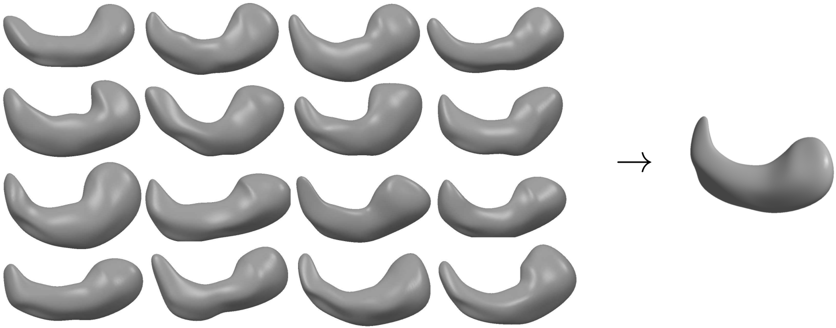


A metric space structure provides a comparison between two shapes.

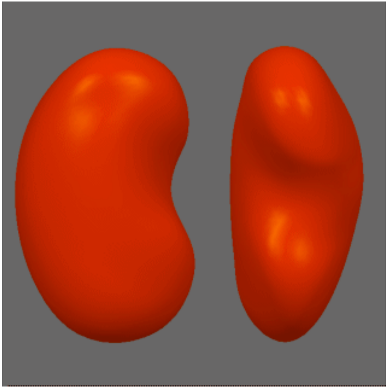
Shape Statistics: Averages

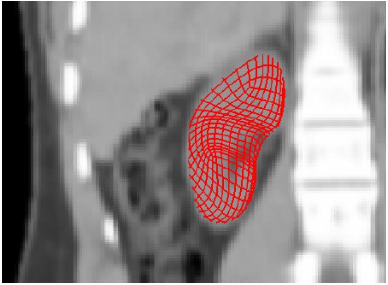


Shape Statistics: Averages



Shape Statistics: Variability





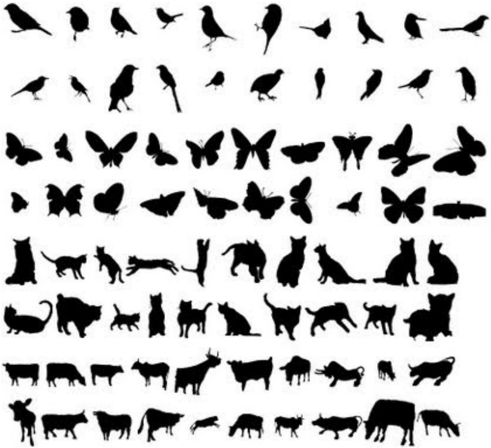
Shape priors in segmentation

Shape Application: Bird Identification

American Crow

Common Raven

Shape Statistics: Classification



http://sites.google.com/site/xiangbai/animaldataset

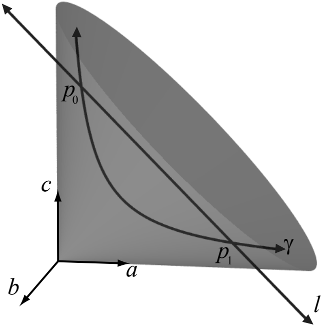
Information Geometry

Parameters of a probability model live on manifolds

Information Geometry

Parameters of a probability model live on manifolds

Example: covariance matrix of a 2D Gaussian distribution:



$\Sigma \in \mathrm{PD}(2)$ is of the form

$$\Sigma = \begin{pmatrix} a & b \\ b & c \end{pmatrix},$$

$$ac - b^2 > 0, \quad a > 0.$$

(positive-definite constraint)

Applications in AI

Latest trends in Artificial Intelligence from a Manifold lens:

- Unsupervised Learning
  - Automatic discovery of intrinsic structure of data, i.e. manifold

Applications in AI

Latest trends in Artificial Intelligence from a Manifold lens:

- Unsupervised Learning
  - Automatic discovery of intrinsic structure of data, i.e. manifold
- Self-supervised Learning
  - Embeds data on a manifold with a known metric

Applications in AI

Latest trends in Artificial Intelligence from a Manifold lens:

- Unsupervised Learning
  - Automatic discovery of intrinsic structure of data, i.e. manifold
- Self-supervised Learning
  - Embeds data on a manifold with a known metric
- Graph Neural Networks (GNN)
  - Graphs are discrete representations of underlying manifold

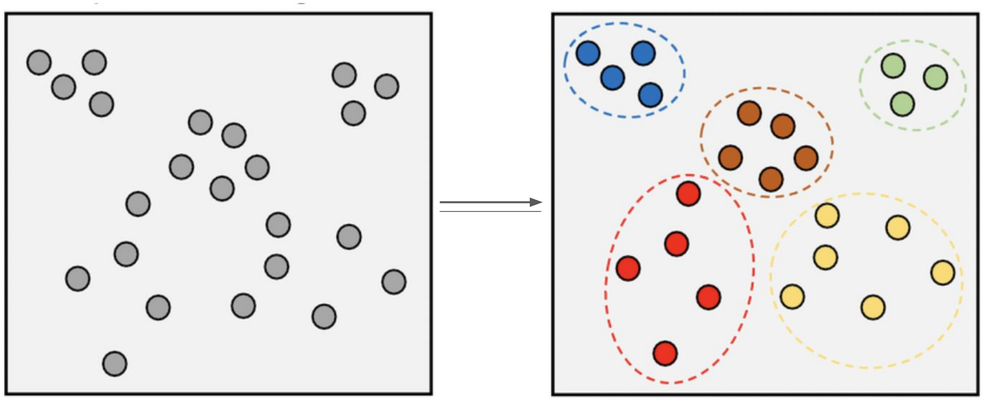
Applications in AI

Latest trends in Artificial Intelligence from a Manifold lens:

- Unsupervised Learning
  - Automatic discovery of intrinsic structure of data, i.e. manifold
- Self-supervised Learning
  - Embeds data on a manifold with a known metric
- Graph Neural Networks (GNN)
  - Graphs are discrete representations of underlying manifold
- Generative Modeling
  - VAEs learn the manifold as their latent representation
  - Diffusion models simulate a noising process through manifolds

Unsupervised Learning

Learns the intrinsic structure by leveraging patterns present in the data <u>without explicit labels</u>.



Clustering

These clusters correspond to modes on the underlying manifold

Self-supervised Learning

Contrastive (Self-)supervised methods project the data to a known manifold to minimize the distance between <u>positive samples</u>

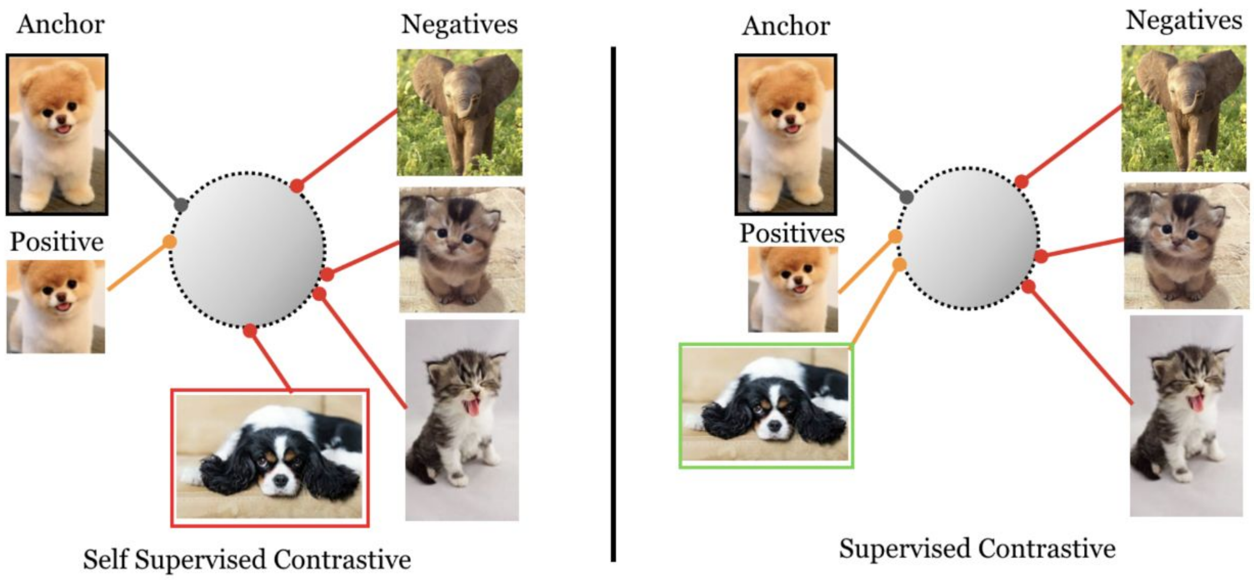
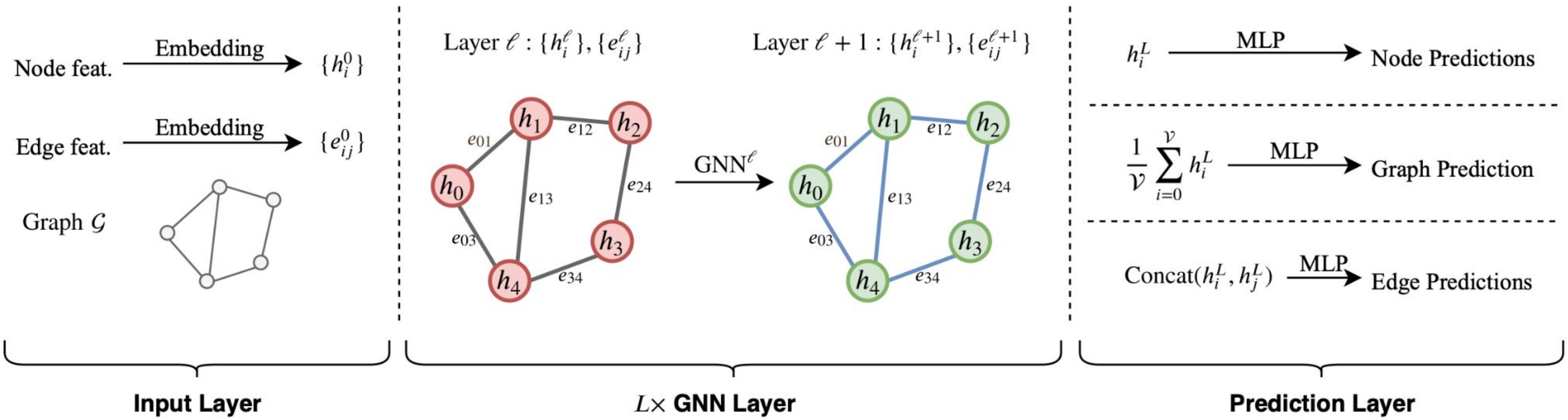


Image src: Khosla, Prannay, et al. "Supervised contrastive learning." Advances in neural information processing systems 33 (2020): 18661-18673.

Graph Neural Networks

Graphs are discrete approximations of continuous manifolds.

Where nodes are data points and edges are relationships



Essentially, GNNs help characterize the manifold discreetly by learning an embedded representation of the graphical data

Image src: https://graphdeeplearning.github.io/post/benchmarking-gnns/

Generative Modeling | VAE

Autoencoders learn a lower-dimensional latent space that helps navigate the high-dimensional manifold of real data

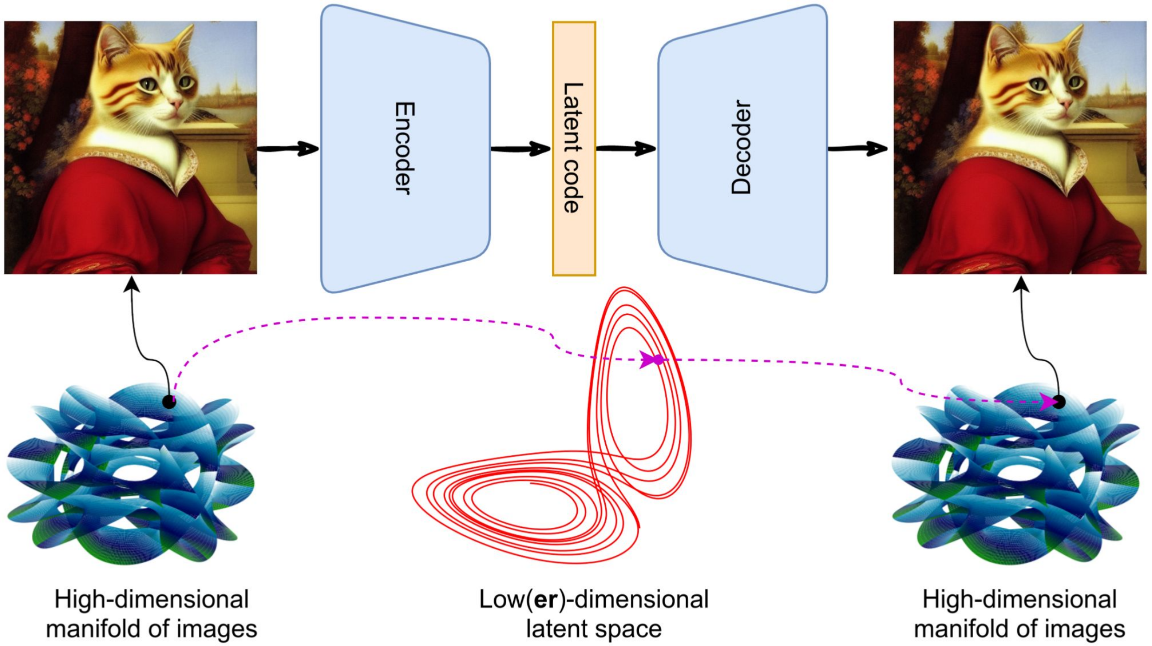


Image src: https://synthesis.ai/2023/02/07/generative-ai-i-variational-autoencoders/

Generative Modeling | Diffusion Models

Diffusion models are just nested VAEs & use geometry of underlying manifolds to simulate the process of spreading noise through them

- **Forward / noising process**
  - Sample data $p(\mathbf{x}_0)$ → turn to noise

- **Reverse / denoising process**
  - Sample noise $p_T(\mathbf{x}_T)$ → turn into data

These models can be conditioned on text i.e. can generate images given their descriptions e.g. OpenAI's Dall-E2, Stable Diffusion etc.

Image src: https://scholar.harvard.edu/binxuw/classes/machine-learning-scratch