Data Science and Deep Learning (2024)

Lecture 2

Working with High-Dimensional Data

Stan Z. Li

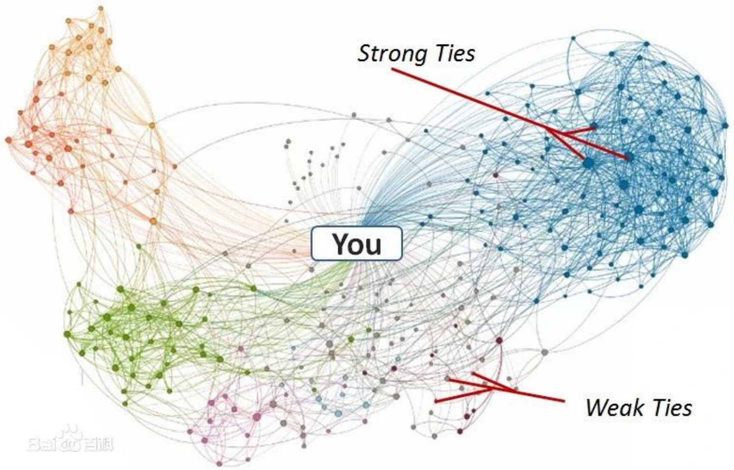
西湖大學
WESTLAKE UNIVERSITY

Outline

1. High-dimensional data
2. Lower-dimensional patterns/manifolds
3. Representational learning/dimension reduction
    - Linear projection
    - Nonlinear projection/neural networks transformation

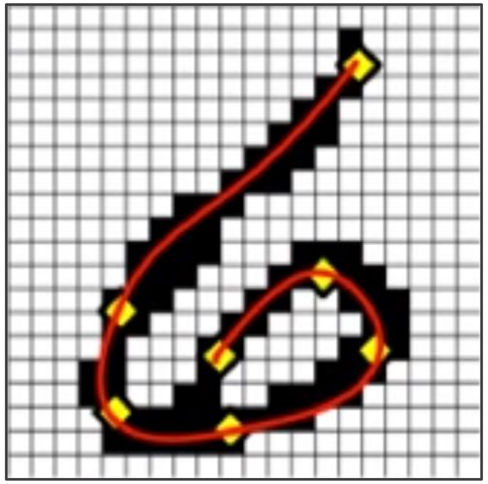
High-Dimensional Data

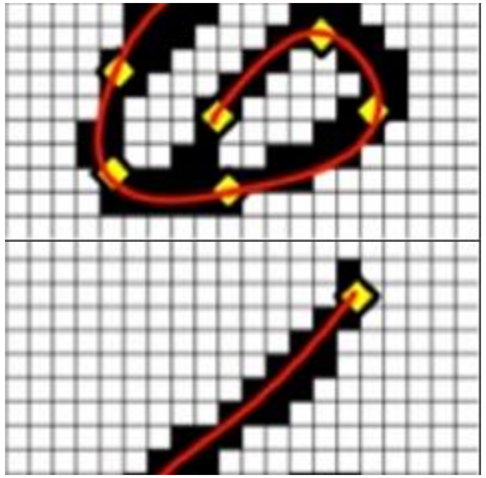
- Images, Videos, Text, Audio,
- Web pages, Social Networks
- Molecular Structures
- DNA Sequences
- Protein Sequence-Structures



Handwritten Digit images

- Image size 20x20 = 400
- Pixel values in {0,1}
- Image Space $\mathbb{S} = \{0,1\}^{400}$
- $\#\mathbb{S} = 2.58 \times 10^{120}$
- Only a tiny portion of $\mathbb{S}$ is of digits
- The digit pattern lives in a low dim subspace (manifold)





Face Image Data

- Image size 100x100 = 10^4 pixels
- RGB image size 3×10^4 pixels
- Dimensionality = 3×10^4
- Pixel values in {0,…,255}
- #Possibility = $256^{30,000} \cong$ infinity
- Only a tiny portion is of faces
- Face pattern lives in low dim subspace

100 pixels

100 pixels

Manifold Assumption

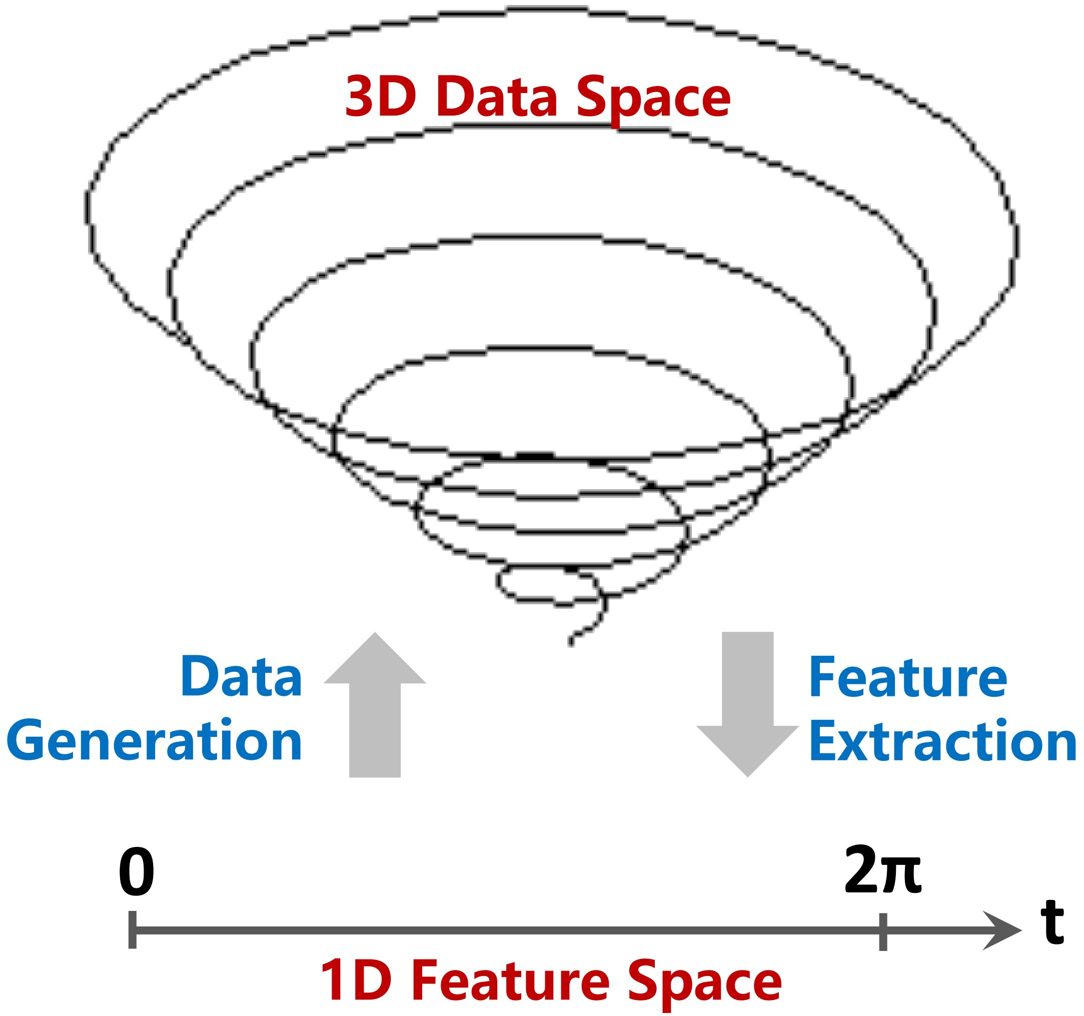
High-Dimensional Data: Images, Web pages, Gene sequences, ….

Dimension Reduction into Coordinate System of a Lower Dim

- **For representation learning （feature extraction)**
- **For data visualization – in 2D or 3D**

Manifold Assumption: an interesting pattern in high dimensional data resides on a low dimensional manifold

Manifold in Hi-D Data Space: 1D Curve in 3D Space



Conical Helix:

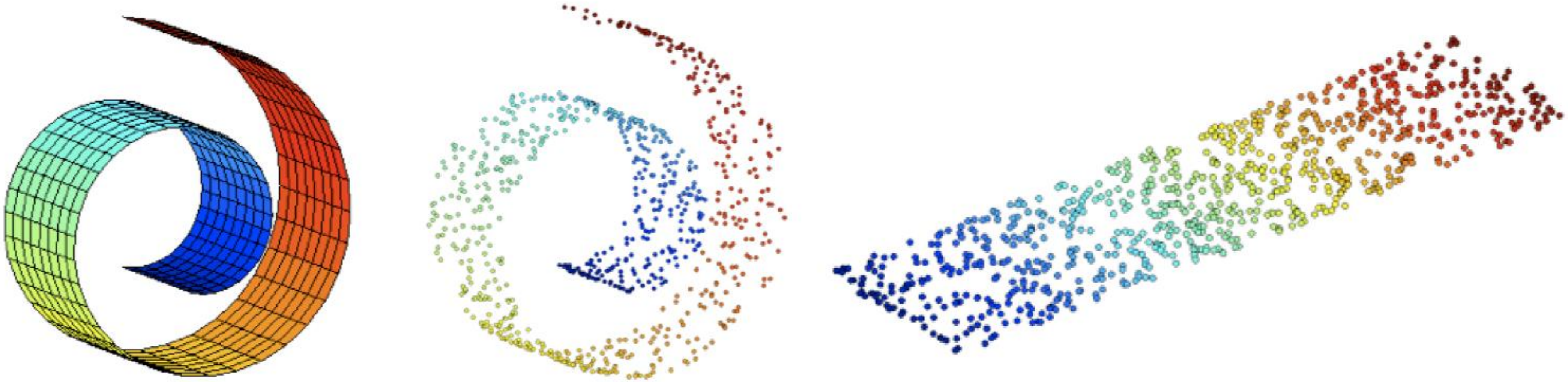
$x=t*\cos(6t)$, $y=t*\sin(6t)$, $z=t$

$0 \le t \le 2\pi$

1D line segment

Latent variable t

2D Manifold in 3D Space



Swiss Roll:

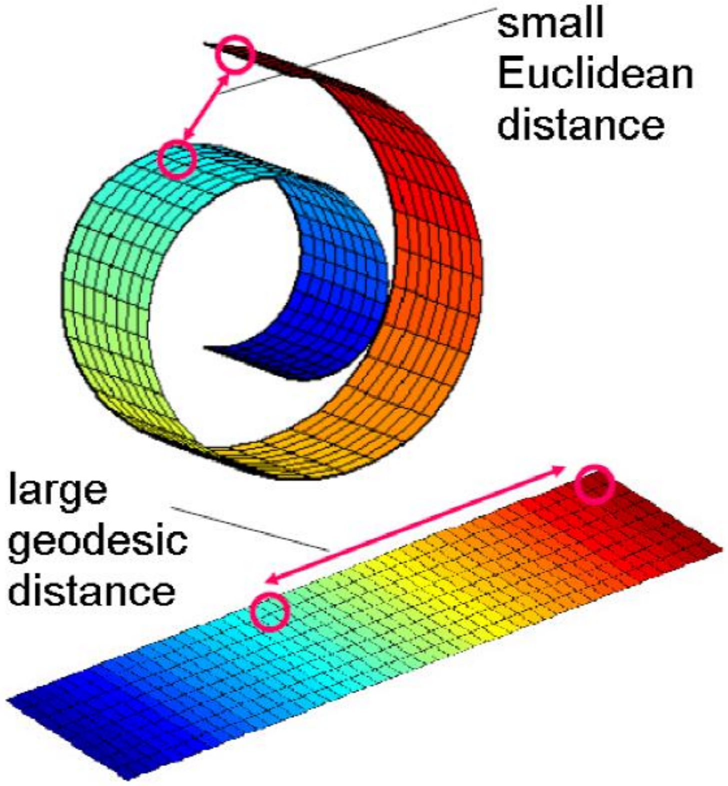
$x=\varphi\cos(\varphi)$, $y=\varphi\sin(\varphi)$, $z=\psi$

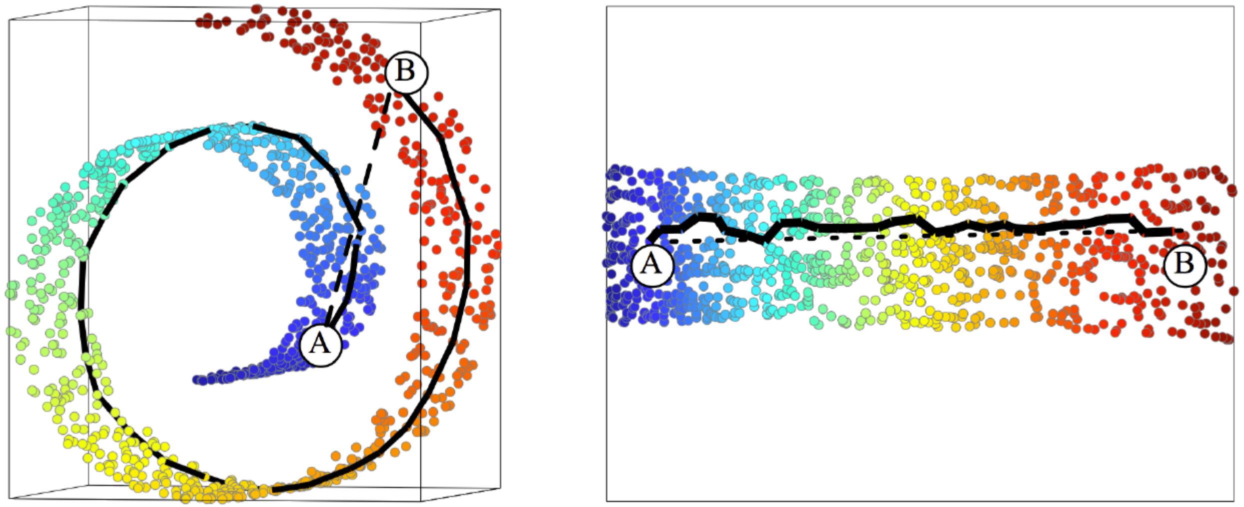
$1.5\pi \le \varphi \le 4.5\pi$, $0 \le \psi \le 10$

Manifold: 2D rectangle

generated by two latent variables φ, ψ

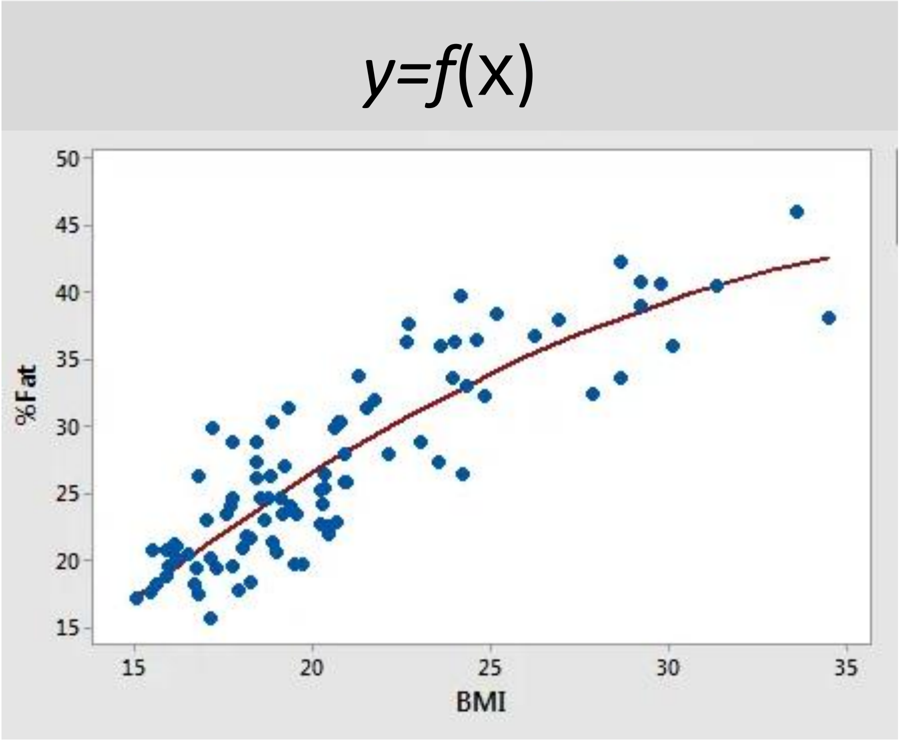
Geodesic Distance on Manifolds



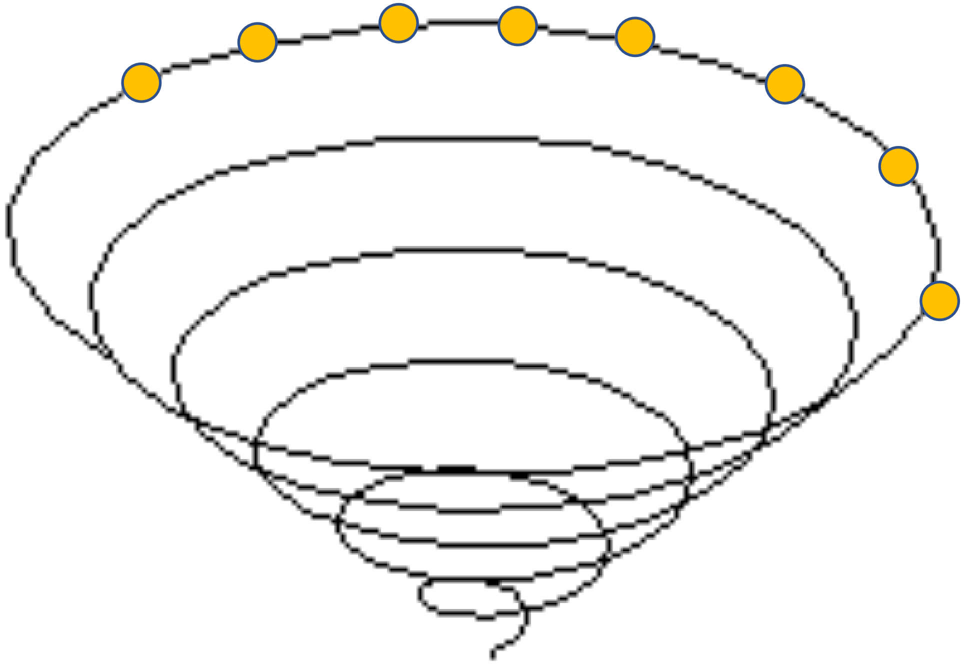


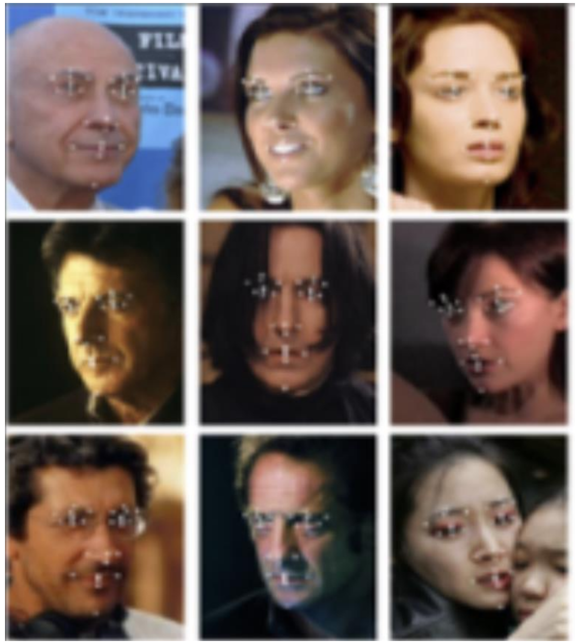
Data Samples on Manifold

***y*=*f*(x) sampled to $\{(x_i, y_i) \mid i = 1,\ldots,n\}$**

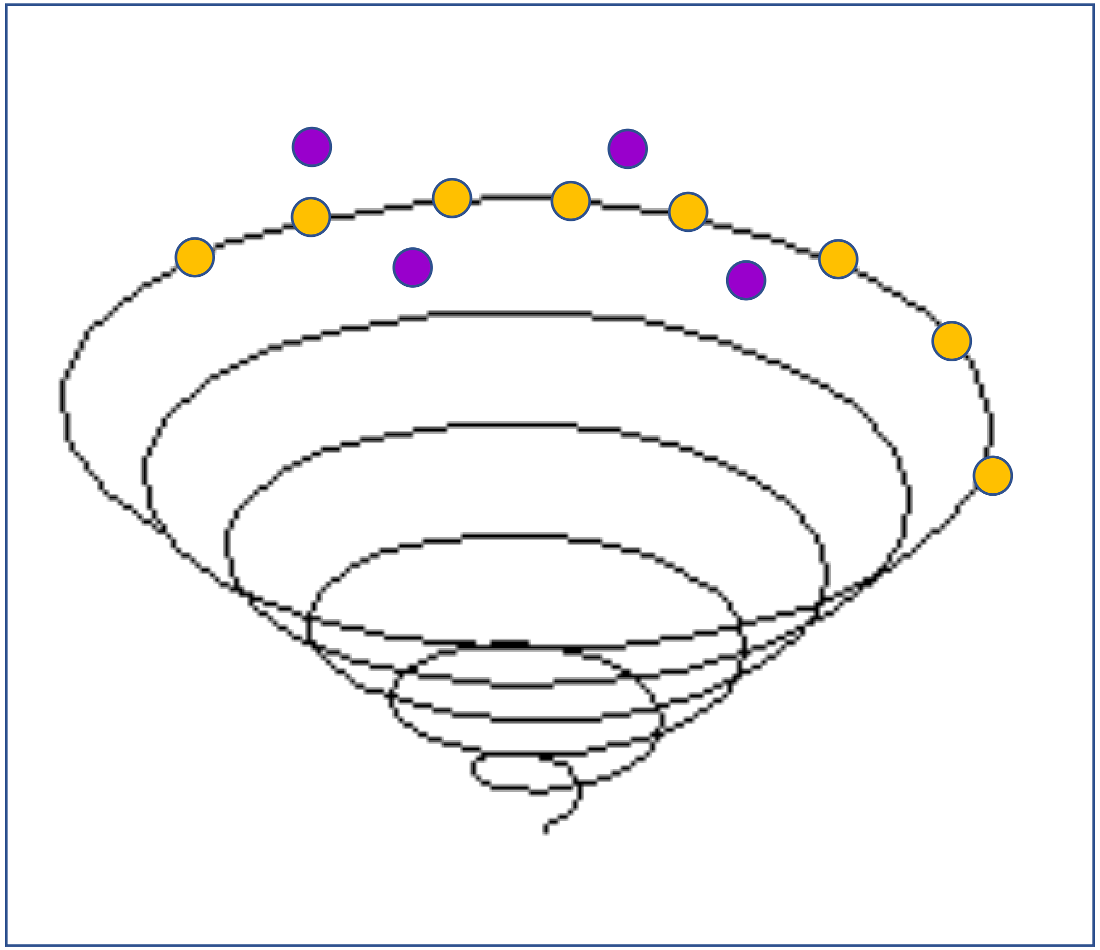


Samples on Face Manifold in Data Space

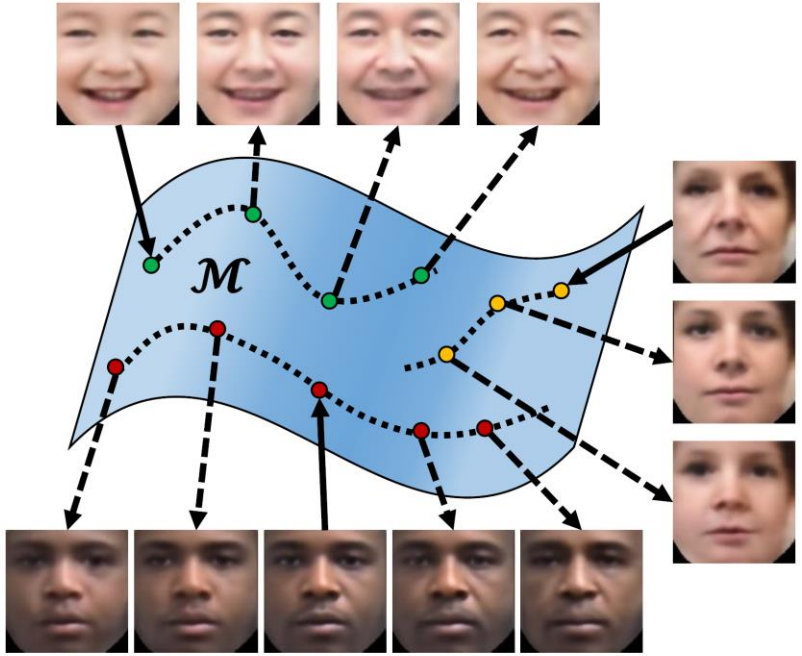




Samples Close to the Face Manifold



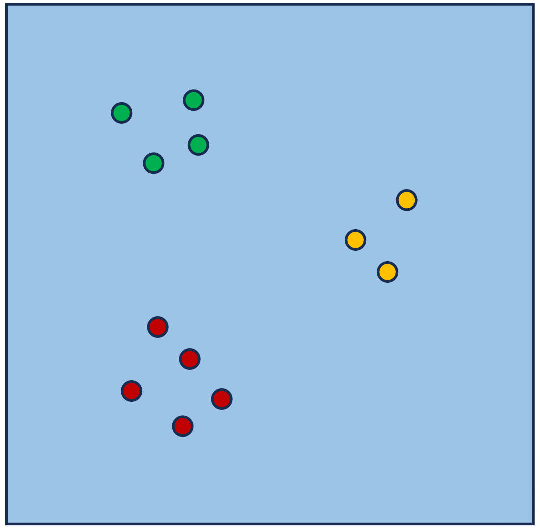
Low-Dim Manifold/Surface in High-Dim Space



Samples on low-dim but complex manifold in high-dim data space

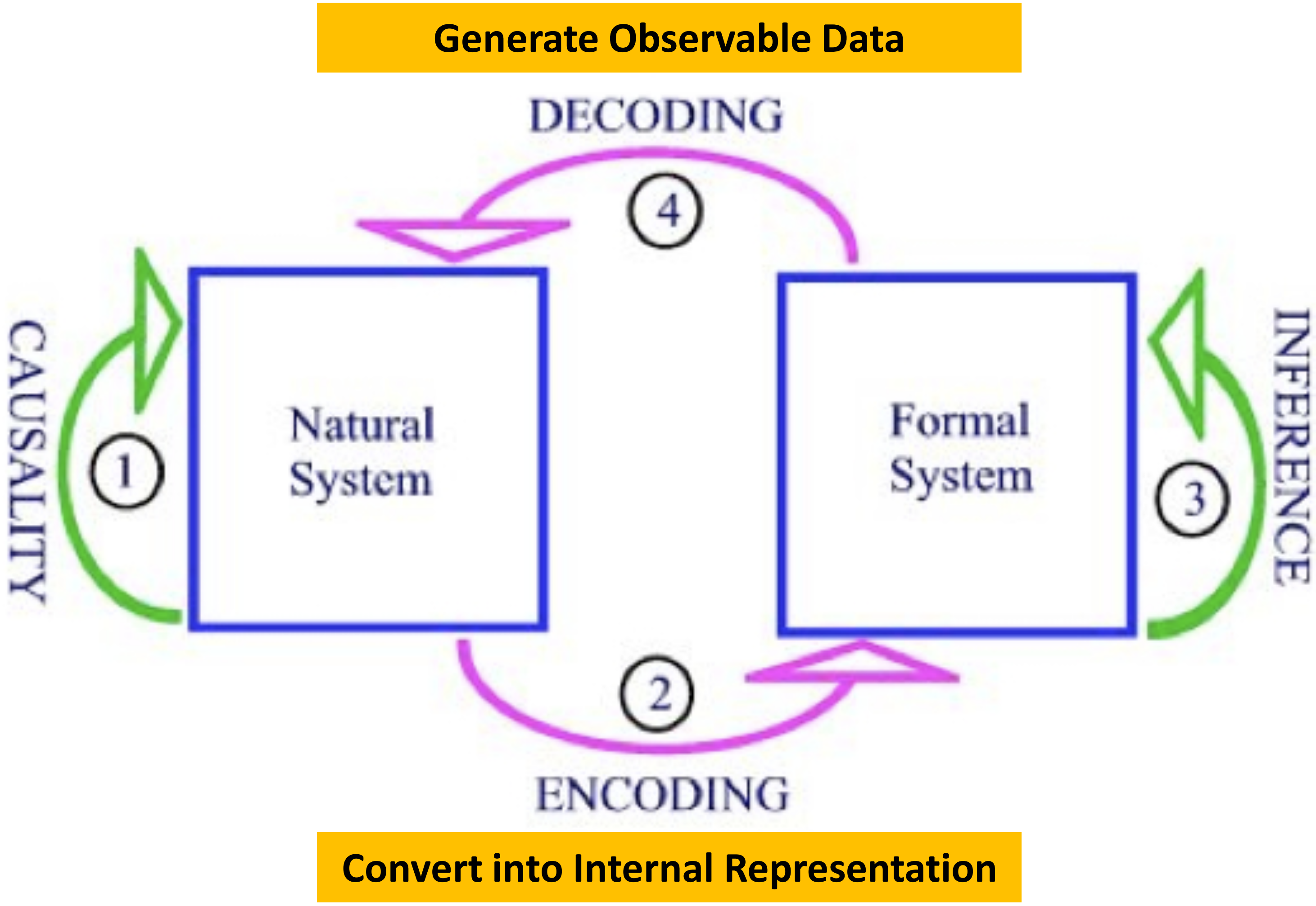
Feature Extraction By Encoder Neural Network

Data Generation by Decoder Neural Network



Features in lower-dim Euclidean embedding space

Scientific Modeling



Thanks