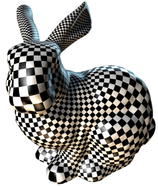
Manifold Fitting: An Invitation to Machine Learning
A Mathematician's view



Shing-Tung Yau
Yau Mathematics Science Center, Tsinghua University

Text-to-Video Generative Model: Sora

Figure: The Sora video generation system from OpenAI emerged unexpectedly, stunning the world.

Figure: Sora claims to be able to apply and understand the basic model of the world, generating content based on descriptions.

The Debate About Sora

Common Opinions

- AGI Arrival Advocates
  - Sora has understood the world model, with correct methodology;
  - Continuing to develop along this direction, general artificial intelligence is just around the corner;
- Brute Force Advocates
  - Sora still has flaws at present;
  - With increased computational power and data, intelligence will naturally emerge;
- Dead End Advocates
  - Simulating the world using pixel generation is both a waste of computational power and a dead end;
  - Text is discrete, symbols are finite, and relatively simple; it can be handled by generative AI;
  - Sensory inputs are continuous and complex, with much greater uncertainty and difficulty in prediction; generative AI is inadequate for handling them.

Our Opinion

- AGI Arrival Advocates
  - Sora approaches the intellectual level of preschool children, with the developed language and visual centers in the brain;
  - Still unable to approach the level of adult intelligence, abstract thinking (mathematical logic) has not yet been established.
- Brute Force Advocates
  - Combining large language models with large visual models has great potential;
  - Unable to model abstract thinking;
  - It is necessary to organically integrate connectionism with symbolicism;
- Dead End Advocates
  - Sora's direction is correct: combining large language models with large visual models
  - Further development: large language models + large visual models + large mathematical models (abstract thinking).

Sora as the World Simulator

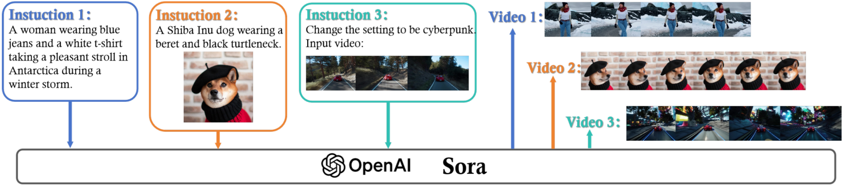


Figure: Sora generates videos based on text prompt.

Sora's physical paradox

Figure: The video generated by Sora depicts a person 'running in opposite directions'.

Sora's physical paradox

Figure: The video generated by Sora shows 'breaking through the glass with red wine'.

Sora's physical paradox

Figure: The video generated by Sora depicts 'clones of the puppy'.

Sora Framework

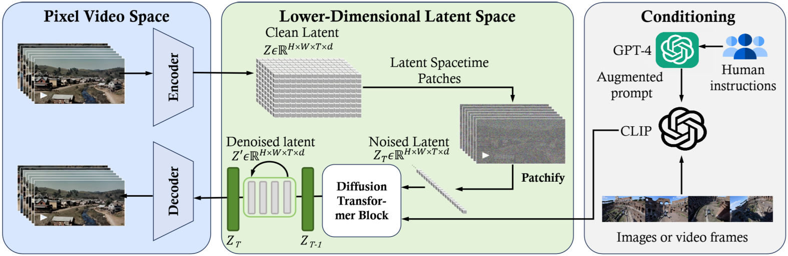
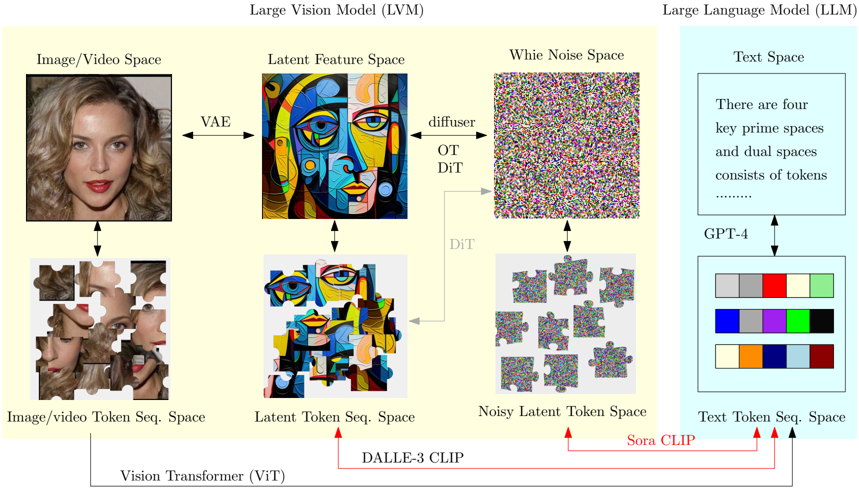


Figure: The basic framework of Sora

Sora Framework

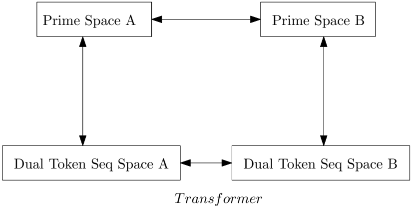


Sora Physical Paradox

- Sora combines with the large language model ChatGPT, greatly enhancing the system's performance
- Mistakes made by Sora when simulating the physical world:
  - Correlation vs Causality (joint distribution and mapping);
  - Local Rationality vs Global Absurdity (infinite correlation length);
  - Missing Critical States (regularity of transmission mapping, boundaries of data manifolds).

The methodology needs further development, incorporating abstract thinking and large mathematical models.

Framework



- Four primary spaces and four dual token sequence spaces;
- The classical mathematical description of transformations between primary spaces;
- The transformation between dual spaces is implemented by the transformers;
- Is the diagram commutative?

VAE, LDM, DDPM, CLIP, DiT, ViT, OT.

Why Exploring Manifolds?

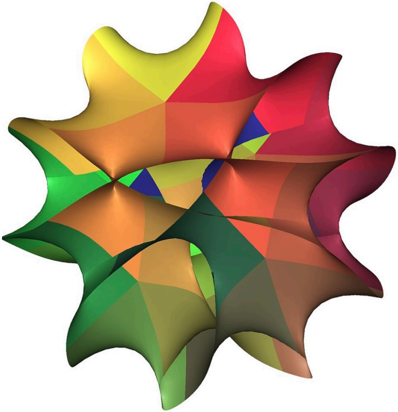


Figure: A 2D slice of a 6D Calabi–Yau quintic manifold

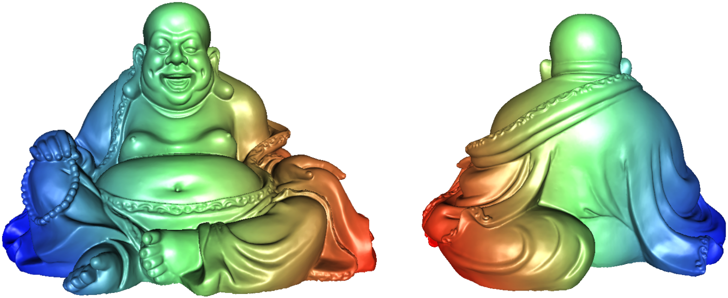
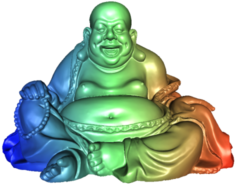
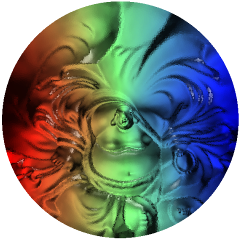


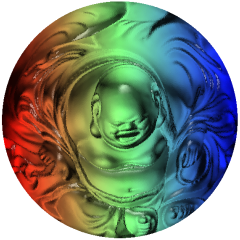
Figure: The happy Buddha surface is a two dimensional manifold embedding in the three dimensional Euclidean space.



(a) manifold



(b) latent space

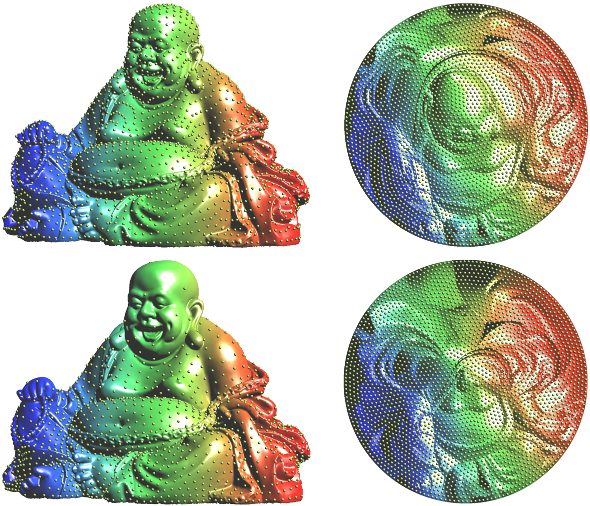


(c) another encoding

- *Encoding*: from the manifold to the latent space;
- *Decoding*: from the latent space to the manifold.

There are infinite many encoding maps.

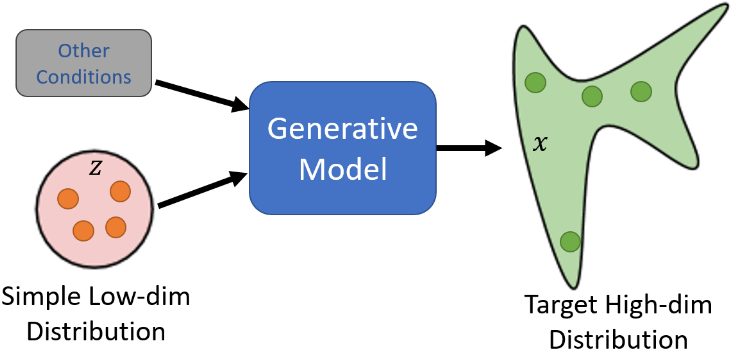
Distribution on Manifold



distributions on manifold latent distributions

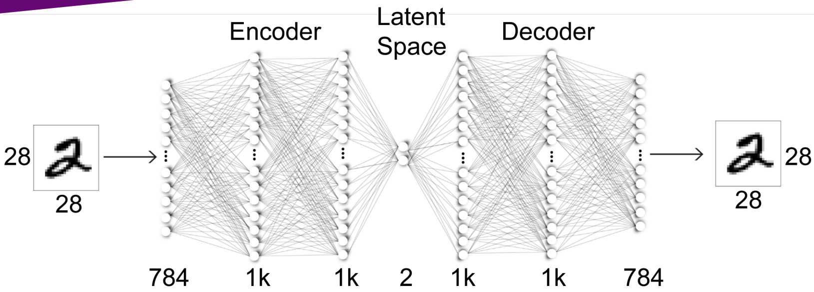
Distributions on data manifolds can be mapped to latent space distributions.

Generative Models



Generative models utilize specified conditions and draw samples from a simple distribution, such as white noise, as inputs. They then produce new samples that mimic a target distribution, including outputs like images, videos and more.

Autoencoder



Autoencoder consists of two symmetric deep neural networks. The widths of the input and output layers are equal to the dimension of the ambient space $\mathbb{R}^n$, and the width of the bottleneck layer is equal to the dimension of the latent space $\mathbb{R}^d$. The first half of the network computes an encoding mapping φ_θ, and the second half of the network computes the decoding mapping ψ_η. Densely sample on the data manifold $\Sigma \subset \mathbb{R}^n$ to obtain $\{x_1, x_2, \dots, x_N\}$, optimize the loss function,

$$\min_{\theta,\eta} \mathcal{L}(\theta, \eta) = \min_{\theta,\eta} \sum_{i=1}^{N} \|x_i - \psi_\eta \circ \varphi_\theta(x_i)\|^2.$$

if the loss is near 0, then the restriction of $\psi_\eta \circ \varphi_\theta$ on Σ is identity, φ_θ and ψ_η are homeomorphisms.

Generative Adversarial Networks (GANs)

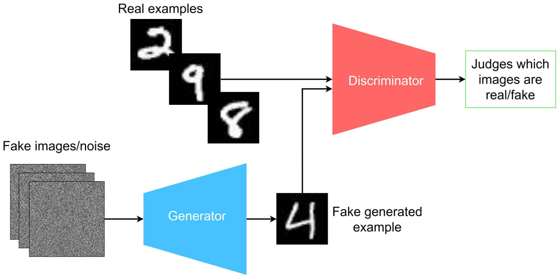
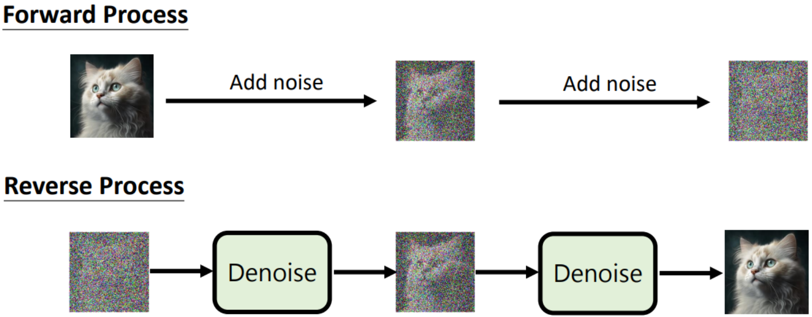


Figure: GANs: more creative than Autoencoders.

GANs have two neural networks contesting with each other in a zero-sum game framework. The generator tries to produce data resembling a certain training dataset, and the discriminator tries to distinguish between genuine data fake data produced by the generator. Through this competitive process, GANs learn to generate new data similar to the training set.

Image source: `https://developer.ibm.com/articles/generative-adversarial-networks-explained/`

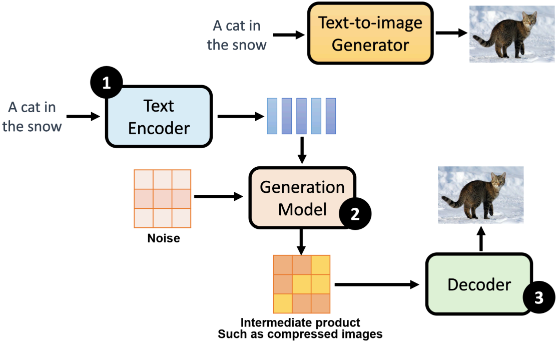
Diffusion Model



Diffusion models add Gaussian noise to an input image over T steps (forward process), distinct from neural network operations. They're trained to reverse this noise, enabling new data generation through the reverse diffusion process.

Image source: `https://speech.ee.ntu.edu.tw/~hylee/ml/ml2023-course-data/DDPM%20(v7).pdf`

Modern ‘Large’ Image Generators



Train three sub-modules (separately) on billions of text-image pairs:

1. Text embedding: LLMs (ChatGPT), BERT...
2. Generation in latent spaces: Diffusion Models, Variational Autoencoders...
3. Decode the intermediate: Autoencoder, Super-Resolution...

High Degree of Freedom – Is the ‘manifold’ still in charge?

Image source: `speech.ee.ntu.edu.tw/~hylee/ml/ml2023-course-data/StableDiffusion%20(v2).pdf`

Part I: Manifold Fitting

The fundamental principle of data science:

Each natural concept corresponds to a *dataset*, where each *sample* is a point in the dataset. The dataset is distributed near a low-dimensional manifold, which is called the *data manifold* Σ. The data manifold Σ is embedded in a high-dimensional *ambient space* $\mathbb{R}^n$. The dataset can be abstracted as a probability distribution μ on the data manifold Σ.

Namely, sample can be modeled* as

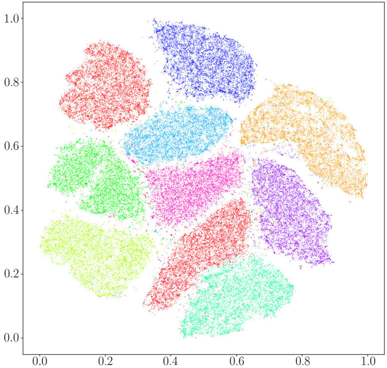
$$y_i = x_i + \xi_i \quad \text{for} \quad i = 1, 2, ..., N$$

- $x_i \in \Sigma \subset \mathbb{R}^n$: unobserved sample from $\mu(\Sigma)$;
- $\xi_i \in \mathbb{R}^n$, $\xi_i \sim \phi_\sigma^{(n)}$: ambient space noise;
- $y_i \in \mathbb{R}^n$, $y_i \sim \mu \star \phi_\sigma^{(n)}$: observation.

---

*Yao Z., Su J., Li B., and Yau S.T., *Manifold Fitting*, arXiv:2304.07680, 2023.

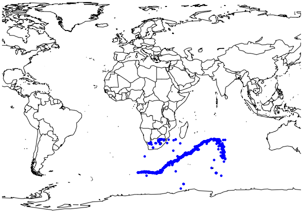
LeCun's MNIST dataset

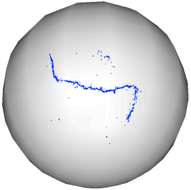


Hinton's t-SNE embedding

The hand written digits image set can be explained well as from a distribution on a 2D surface embedded in the space of gray images.

Sample lies on known manifolds:





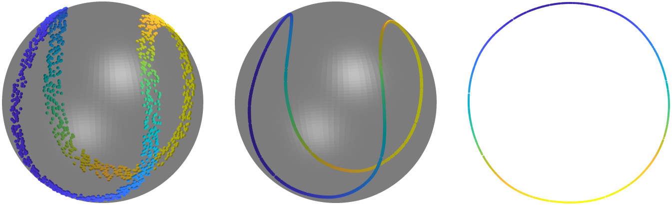
- Asymptotics: Central limit theorem on manifolds...
- Regression: Regression on Lie group, Fréchet regression...
- Classification: Principal boundaries*, classify CY manifolds*...
- Extensions of PCA:
  - Nonlinear generalization: principal flows*, principal submanifolds*...
  - PCA on shape spaces: principal nested spheres, torus PCA...

*Recent work by Yao, Z., Yau, S.T. and other collaborators

Subspace of Maximal Variation

Find a reasonably smooth d-dim subspace Γ of Σ such that

- for any point $x \in \Sigma$, the tangent space $T_x\Gamma \subset T_x\Sigma$;
- $T_x\Gamma$ is 'approximately' the span space of the leading d eigenvectors of the local covariance matrix at x;
- Γ roughly passing the 'center' of the observations.



For a stripe dataset on a sphere, we can find a curve that captures its most of the variation and, if necessary, find a proper encoding for it.

Connection to SYZ conjecture

- The concept of finding a sub-manifold of data lying on manifolds is naturally rooted in a seemingly unrelated conjecture, namely, the SYZ conjecture[†]. Without diving into too many mathematical statements, the conjecture offers a geometrical way of breaking a complicated space (manifold) into its constituent parts.

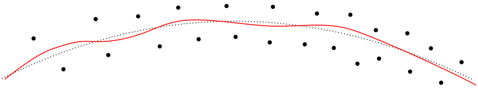
- The problem is related to principal sub-manifold, which is an empirical calculation of such decomposition under some scenarios from the noisy data[‡].

[†] Strominger, A., Yau, S.T., & Zaslow, E. *Mirror symmetry is T-duality.* Nuclear Physics B, 1996.

[‡] Yao, Z., Li, B., Tran, V.D. & Zhang, Z. *Principal Sub-manifolds: New Theory and Methods*. Manuscript, 2023.

Data manifold is unknown:

- Manifold Learning (Dimension Reduction):
  Find a low dimensional linear representation.
- Manifold Fitting: Fit manifolds in the same space.



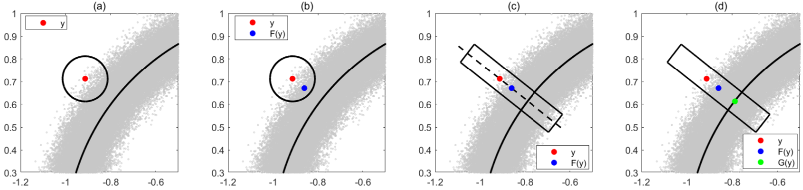
Genovese et al (2012 a,b), *Fefferman et al (2016)*, *Mohammed/Narayanan (2017)*, *Yao/Xia (2019)*, *Fefferman et al (2021)*, *Yao/Su/Li/Yau (2023)*[§], *Yao/Su/Yau (2024)*[¶], *Yao/Li/Lu/Yau (2024)*[‖].

§ *Manifold fitting*, arXiv:2304.07680.

¶ *Manifold fitting with CycleGAN*, PNAS, 2024.

‖ *Single-Cell Analysis via Manifold Fitting: A New Framework for RNA Clustering and Beyond*, revise at PNAS.

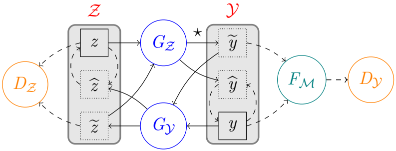
Overall Idea of Manifold Fitting



- Find the dominant direction;
- 'Push' noisy points along the direction;
- Obtain a manifold estimator $\widehat{\Sigma}$ that
  - is a smooth manifold;
  - closely represents Σ;
  - remains in the same ambient space.

CycleGAN/Manifold Fitting Framework†

GANs can be modified to fit data manifolds:



- $\mathcal{Z} \subset \mathbb{R}^d$: latent space
- $\mathcal{Y} \subset \mathbb{R}^n$: ambient space
- $G_\mathcal{Z}$, $G_\mathcal{Y}$: generators
- $D_\mathcal{Z}$, $D_\mathcal{Y}$: discriminators
- $F_\mathcal{M}$: manifold fitting sub-module

Main objective*: Let $Z \sim \mathrm{Unif}(0,1)^d$,

$$\min_{G_\mathcal{Z} \in \mathcal{C}(G_\mathcal{Z})} \mathrm{Div}(G_\mathcal{Z}(Z) \star \phi_\sigma, \nu),$$

and the data manifold can be estimated with the generators.

---

*ν: the distribution of y_i; ϕ_σ: the noise distribution; $\star$: convolution; Div: certain divergence.

†Yao, Z., Su, J., and Yau, S.T., *Manifold fitting with CycleGAN*, PNAS, Jan, 2024.

Role of Generators

Solve (non-sample version):

$$G^*_{\mathcal{Z}}, G^*_{\mathcal{Y}} = \arg\max_{G_{\mathcal{Z}}, G_{\mathcal{Y}}} \min_{F_{\mathcal{M}}, D_{\mathcal{X}}, D_{\mathcal{Y}}} \mathcal{L}(G_{\mathcal{Z}}, G_{\mathcal{Y}}, F_{\mathcal{M}}, D_{\mathcal{X}}, D_{\mathcal{Y}}).$$

- Manifold estimators (sample-based):

  $\widetilde{\mathcal{M}} = \widehat{G^*_{\mathcal{Z}}}(\mathcal{Z})$ or $\widehat{\mathcal{M}} = F_{\mathcal{M}}(\widetilde{\mathcal{M}})$ estimates $\mathcal{M}$.

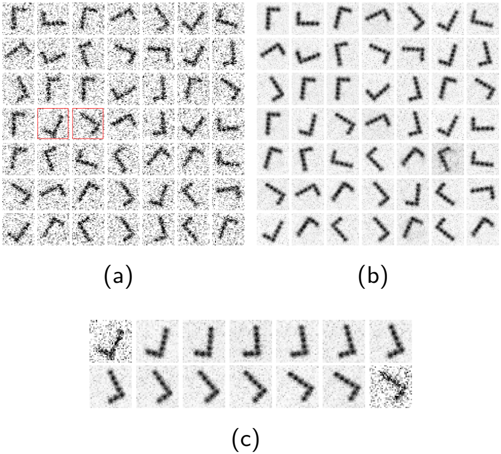
- Noise canceling:

  $\widehat{G^*_{\mathcal{Z}}} \circ \widehat{G^*_{\mathcal{Y}}} : \ y_i \mapsto \widehat{y}_i \in \widetilde{\mathcal{M}}$.

- Nonlinear interpolating:

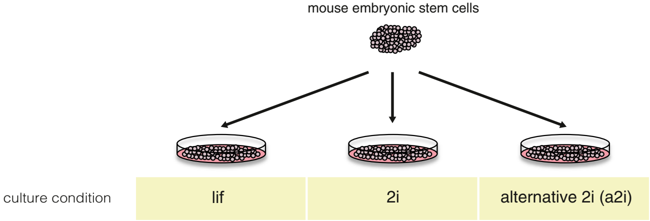
  $\widehat{G^*_{\mathcal{Z}}} \left(t\widehat{G^*_{\mathcal{Y}}}(y_i) + (1-t)\widehat{G^*_{\mathcal{Y}}}(y_j) \right)$ nonlinear interpolates between $\widehat{y}_i$ and $\widehat{y}_j$.

A Toy Example – Fully Control on Data Manifold



- (a) Images of a rotating simple shape, with ambient space noise.
- (b) Denoised version of (a) with CycleGAN/Manifold Fitting model.
- (c) Nonlinear interpolation of two examples with red boxes in (a).

Fitting in Biology Data – scRNA (single-cell) Example



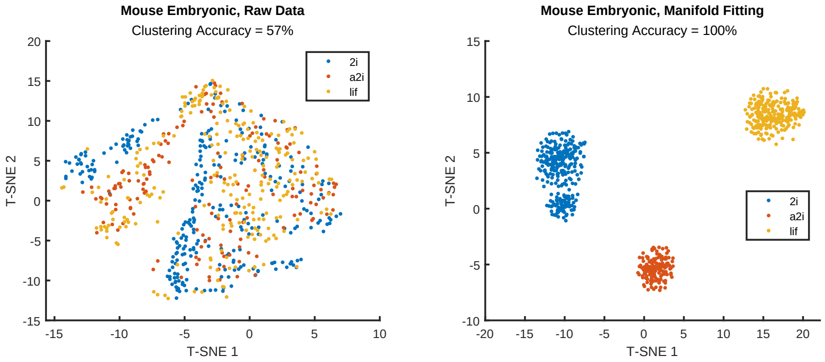
The Kolodziejczyk dataset[‡] : 704 cells with 13,473 features in 3 classes

Focuses:

- Utilizing the potential molecular mechanisms governing cell differentiation and maintenance.
- Keeping the three classes of Mouse embryo stem cells.
- Improving other unsupervised clustering methods with the help of fitting.

[‡] Kolodziejczyk A A, et al. *Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation*. Cell Stem Cell, 2015.

Manifold Fitting in scRNA



Manifold fitting methods improve the spatial distribution of the data and the unsupervised clustering score for this data after fitting (right), significantly higher than the other methods without using fitting (left).

Manifold Fitting in scRNA

Accuracy (measured by ARI[§]) in clustering 25 scRNA datasets.

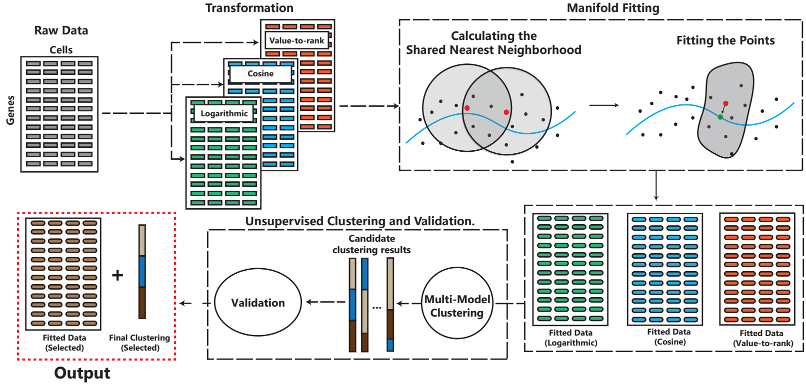
Dataset	scDHA[¶] Raw	Manifold Fitting	Dataset	scDHA Raw	Manifold Fitting
EMTAB2600	1.00	1.00	SRP041736	0.85	0.91
EMTAB3321	0.86	0.91	EGAD00001010074	0.46	0.55
GSE36552	0.78	0.86	GSE202352	0.42	0.73
GSE59739	0.64	0.88	16-WM8C	0.09	0.80
GSE60361	0.82	0.87	GSE132042	0.60	0.84
GSE67835	0.72	0.75	GSE132042-Liver	0.46	0.67
GSE81252	0.37	0.41	Midbrain	0.51	0.93
GSE81608	0.53	0.82	GSE81547	0.42	0.46
GSE83139	0.7	0.83	E-MTAB-11265	0.65	0.75
GSE84133-M	0.47	0.67	MAC-Bladder	0.51	0.57
GSE85241	0.92	0.87	Local11	0.47	0.82
GSE103322	0.59	0.59	MAC-Brain	0.13	0.83
GSE108097	0.26	0.39	Average	0.57	0.75

Manifold fitting can significantly improve the clustering result.

[§] Adjusted Rand Index: a measure of the similarity between clustering results and the ground truth.

[¶] A leading scRNA clustering method: Tran, Duc, et al. (2021). *Fast and precise single-cell data analysis using a hierarchical autoencoder*. Nature communications.

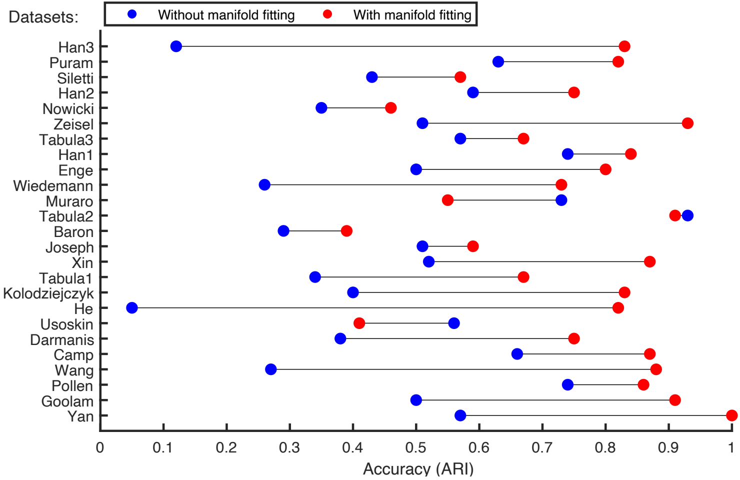
Single-Cell Analysis via Manifold Fitting (scAMF)‖



The scAMF pipeline processes raw data through transformations and denoising via manifold fitting, followed by unsupervised clustering. It concludes with a validation step, outputting the final clustering results.

‖Yao Z., Li B., Lu Y., and Yau S.T., *Single-Cell Analysis via Manifold Fitting: A New Framework for RNA Clustering and Beyond*, revision at PNAS.

Clustering with and without Manifold Fitting



Manifold fitting helps in data analysis!

Outlook – Fields Without 'Billion-scale' Data

Data Manifolds + Data Sciences

- New Thinking:
  - Metric learning on data manifolds;
  - Generalized dimension reduction methods on manifolds;
  - ETC. ...
- Application:
  - CellScape: Enhanced cell atlas builder;
  - Health: Enhanced precision medicine and sub-type analysis;
  - ETC. ...
- Collaborator's Group: Zhigang Yao, National U of Singapore

Part II: Geometric View of Optimal Transport

Major Tasks of DL

According to manifold distribution principle, the major tasks of deep learning are to learn

1. Data manifold structure by *manifold fitting*;
2. Data distribution by *transportation maps or plans*;

Transport Maps

The data distribution (Ω, μ) is transported to the Gaussian or uniform distributions (Ω^*, ν) via various algorithms, mostly *diffusion maps* or *optimal transport maps.*

Generative Model

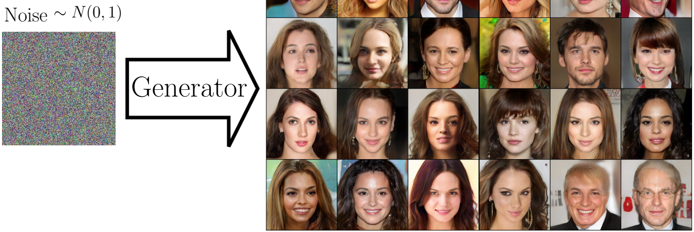


Figure: Generative Model: the input is a Gaussian noise image, the output is a realistic photo (or video), such as Sora.

Generative Adversarial Networks (GAN)

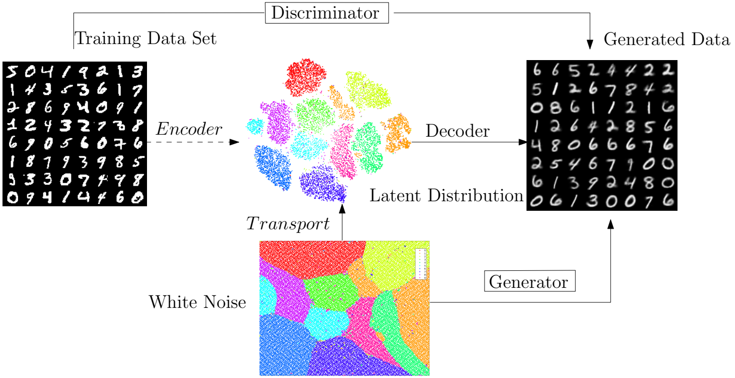
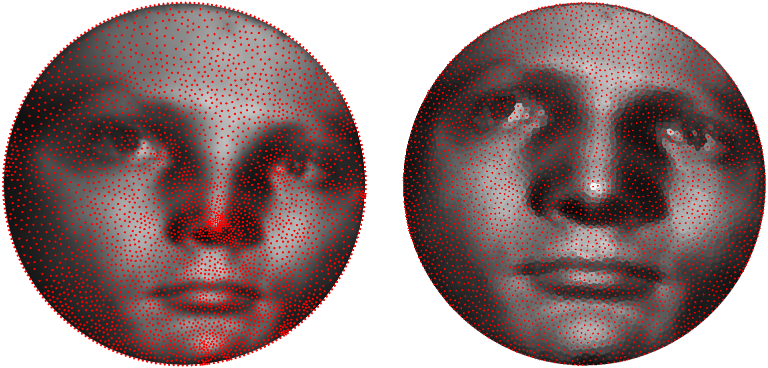


Figure: Generative Adversarial Networks. The uniform distribution (white noise) is transported to the latent distribution.

Transport Map



source data distribution μ target uniform distribution ν

Figure: Transport map $T : \mathbb{D}^2 \to \mathbb{D}^2$ is an automorphism, $\mu \mapsto \nu = T_{\#}\mu$ maps the data distribution μ (red dots)to the uniform distribution.

Optimal Transport Map

Problem (Monge)

Given compact metric spaces Ω and Ω with measures $d\mu(x) = f(x)dx$ and $d\nu(y) = g(y)dy$, where $\mu(\Omega) = \nu(\Omega^)$, and given a transport map $T : \Omega \to \Omega^*$, satisfying

$$\int_{T^{-1}(E)} d\mu = \int_E d\nu, \quad \forall \text{ Borel } E \subset \Omega^*,$$

Then T is measure-preserving, denoted as $T_{\mu=\nu}$. Given a transport cost $c : \Omega \times \Omega^* \to \mathbb{R}$, we seek the optimal transport map.

$$\min_{T_\# \mu = \nu} \int_\Omega c(x, T(x))\, d\mu(x).$$

Optimal Transport map

Theorem (Brenier)

If the cost function is the square of the Euclidean distance, then the optimal transport map T is the gradient map of the Brenier potential function $u : \Omega \to \Omega^*$, $T = \nabla u$, where u satisfies the Monge-Ampère equation:

$$\det\left(\frac{\partial^2 u}{\partial x_i \partial x_j}\right) = \frac{f(x)}{g \circ \nabla u(x)}.$$

The Monge-Ampère equation is a strongly nonlinear elliptic equation.

S.-Y. Cheng, S.-T. Yau, On the regularity of the monge-ampère equation $det(\partial^2 u/\partial x_i \partial x_j) = f(x, u)$, Communications on Pure and Applied Mathematics, 1977.

Minkowski Theorem

Theorem (Minkowsi)

Given $\{(A_i, \mathbf{n}_i)\}_{i=1}^k$, *such that* $\mathbf{n}_i$*'s can not be contained in any hemi-sphere and*

$$\sum_{i=1}^{k} A_i \mathbf{n}_i = \mathbf{0}, \quad A_i > 0,$$

then the convex polytope P *with face normals* $\mathbf{n}_i$*'s and face areas* A_i*'s exists and is unique upto a translation.*

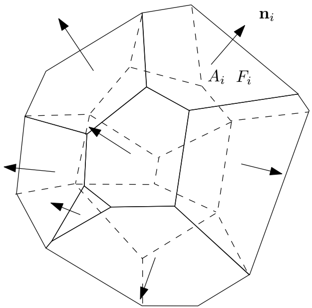


Figure: Minkowski Theorem.

S.-Y. Cheng, S.-T. Yau, On the regularity of the Solution of the n-Dimensional Minkowski Problem, Communications on Pure and Applied Mathematics, 1976.

Alexandrov Theorem

Theorem (Alexadnrov 1950)

Ω is a compact convex domain in $\mathbb{R}^n$, for each face F_i, the normal $\mathbf{n}_i$ and the projected area A_i is given,

$$\sum_i A_i = vol(\Omega),$$

then the convex polytope exists and is unique upto a vertical translation.

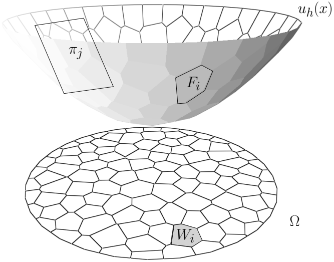


Figure: Alexandrov theorem.

Alexandrov Theorem

- By taking to the limits, the Minkowski theorem and the Alexandrov theorem determine the shape of a convex surface by prescribing the Gaussian curvatures.
- Both Brenier theorem and the Alexandrov theorem are reduced to the Monge-Ampère equation, therefore equivalent.
- Alexandrov's original proof is based on algebraic topology, not constructive.
- My students and I developed a convex variational approach to solve the Alexandrov problem.

X.Gu,F.Luo, J.Sun and S.-T. Yau, "Variational Princples for Minkowski type Problems, Discrete Optimal Transport, and Discrete Monge-Ampere Equations", AJM 20(2), 383–398, April 2016.

Geometric Variational Method for OT Map

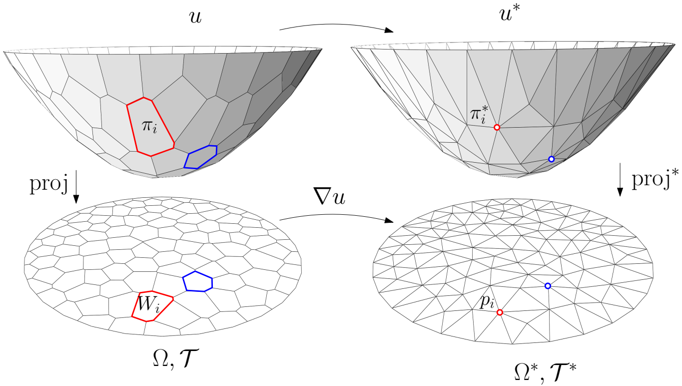
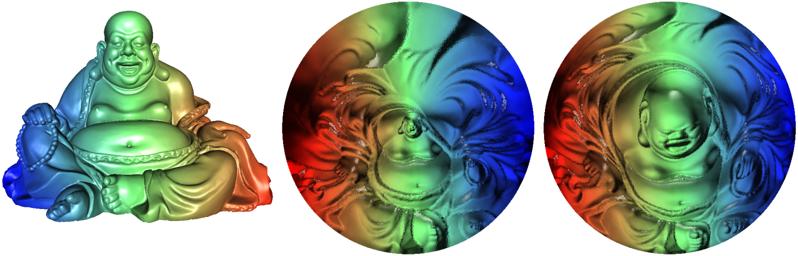


Figure: Discrete OT map (left to right): maps W_i to p_i; Discrete Monge-Ampère equation (right to left): $\mu(W_i)$ is the Hessian at p_i. The method can be implemented using the power diagram and the weighted Delaunay algorithms with Newton's optimization method.

Optimal Transport map



$(S, \mathbf{g})$ area element $dA_{\mathbf{g}}$ $\quad(\mathbb{D}^2, \mu)\quad$ $(\mathbb{D}^2, \mathcal{L})$

Figure: The Riemann mapping $\varphi : (S, \mathbf{g}) \to (\mathbb{D}^2, \mu)$ pushes forward the surface element $dS_{\mathbf{g}}$ to the measure μ on the planar disk, such that $\varphi_{\#} dS_{\mathbf{g}} = \mu$; the optimal transport map $T : \mathbb{D}^2 \to \mathbb{D}^2$ pushes forward μ to the Lebesgue measure $\mathcal{L}$.

Optimal Transport map

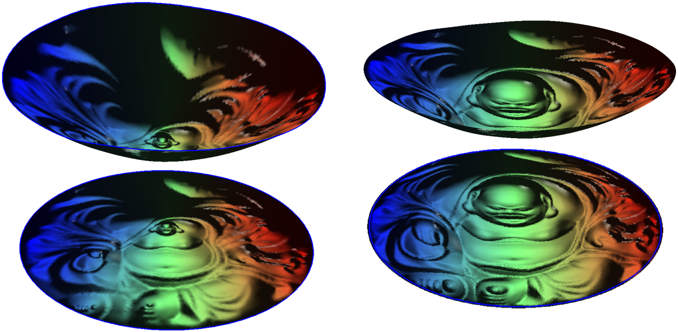


Figure: The optimal transport map $T : (\mathbb{D}^2, \mu) \to (\mathbb{D}^2, \mathcal{L})$ is equal to the gradient of the Brenier potential function.

X.Gu,F.Luo, J.Sun and S.-T. Yau, "Variational Princples for Minkowski type Problems, Discrete Optimal Transport, and Discrete Monge-Ampere Equations", AJM 20(2), 383–398, April 2016.

Singularity Set of OT Maps

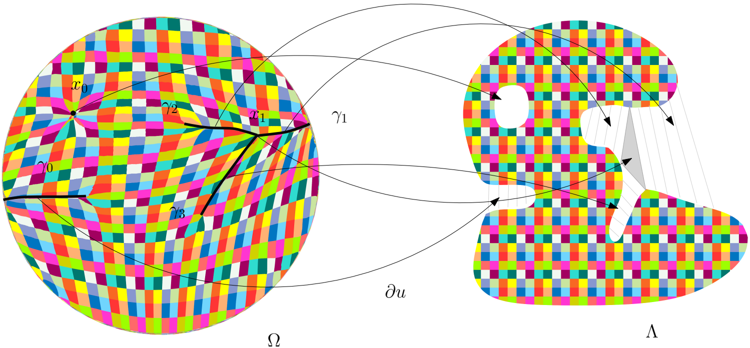


Figure: Figalli's example: Singularity structure of an optimal transportation map.

A.Figalli, Regularity properties of optimal maps between nonconvex domains in the plane, Comm. Partial Differential Eequations, 35(3):465-479, 2010.

Singularity Set of OT Map

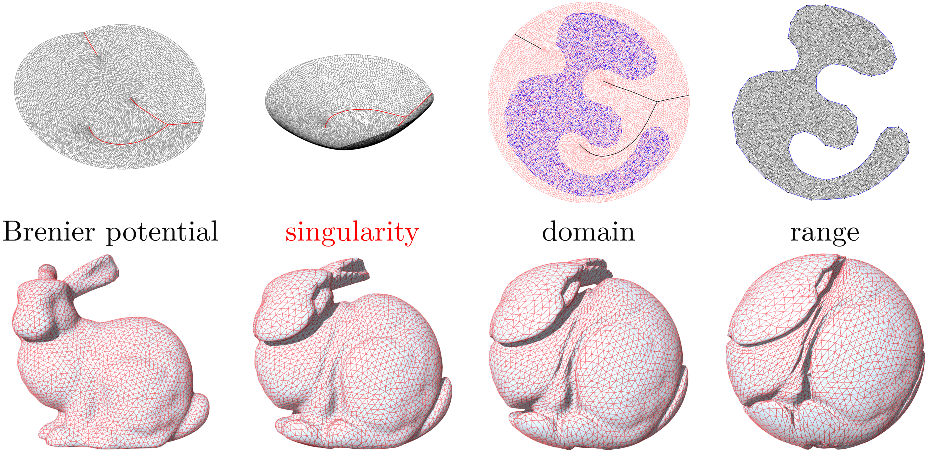


Figure: The singularity sets of the optimal transport maps.

Transport Map - Heat Diffusion

The entropy of a distribution $(\Omega, \rho dx)$ is $H(\rho) = \int_\Omega \rho \log \rho dx$. The entropy flow $(\mathbf{v}(x,t), \rho(x,t))$ maximizes the entropy

$$\partial_t \rho(x,t) = -\Delta_x \rho(x,t). \quad \boxed{\mathbf{v}(x,t) = \nabla \log \rho(x,t)}.$$

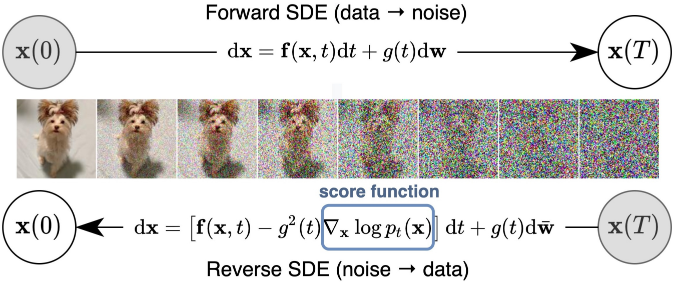
From the Langevin dynamics, the equivalent stochastic differential equation (SDE) can be defined as:

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\, dt + g(t)\, d\mathbf{w},$$

Here, $\mathbf{f}(\cdot, t), \mathbf{g}(t)$ are the drift and the diffusion coefficients, and $\mathbf{w}$ the standard Brownian motion. The discrete approximation is:

$$\mathbf{x}_{t+\Delta t} = \mathbf{x}_t + \mathbf{f}(\mathbf{x}_t, t)\Delta t + g(t)\sqrt{\Delta t}\varepsilon, \varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

The SDE has a corresponding backward SDE.



Intuitively, the stochastic differential equation method is equivalent to adding white noise to an image multiple times until the image becomes white noise; the reverse process is equivalent to reducing white noise from an image until the original image is restored.

OT Map vs. Heat Diffusion

- Optimal transport map can transport any source measure to the target measure; Heat diffusion maps source measure to *Gaussian distribution* only.
- Geometric optimal transport map can discovers the *singularity set*, which gives the *support boundaries* of the target measure precisely; Heat diffusion obscures the support boundaries of of the target measure;
- The *support boundaries* of the data distribution have special physical/logical importance, ignoring or crossing them will cause *physical mistakes* for generative models.

Singularity Set Experiment

Generative Model for human facial images

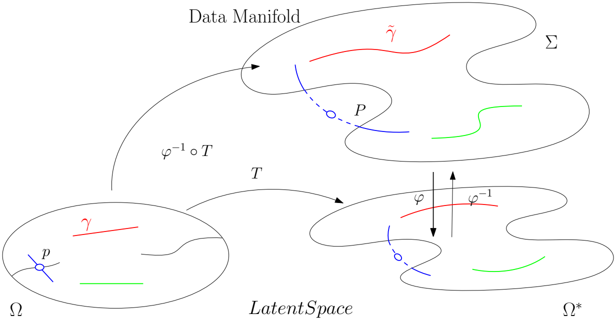
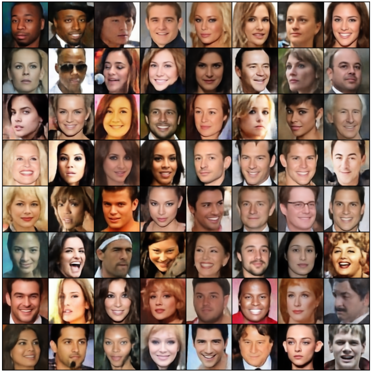


Figure: The blue line in Ω crosses the singularity set of the OT map, the blue curve on Σ crosses the boundary of the manifold.

Yau et al. "A geometric view of optimal transportation and generative model", arXiv:1710.05488.

The boundary of the data manifold



(a) generated samples

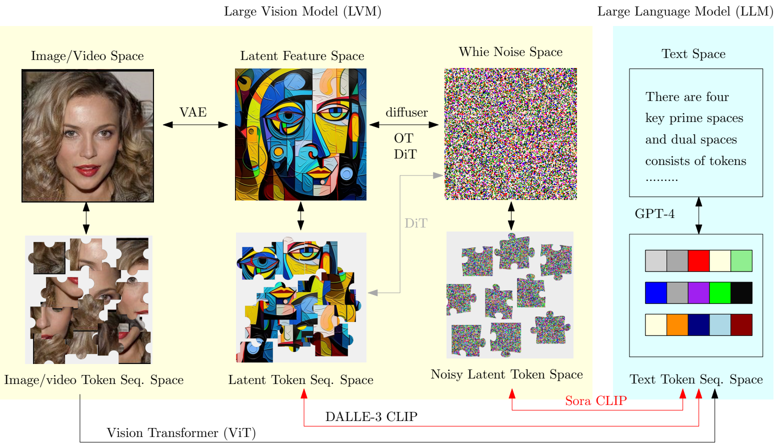
(b) curve crossing singularity

Figure: The heterochromia faces are on the boundary of the human face image manifold, which are detected by the singularity set of the geometric OT map.

Text-to-Video Generative Model: Sora

Figure: The Sora video generation system from OpenAI emerged unexpectedly, stunning the world.

Fundamental Framework



The top left 3 frames show a generative model in Sora, the diffuser ignores the support boundaries of the latent distribution, thus induces mistakes.

Figure: The video "Red Wine Breaking the Glass" generated by Sora.

Absence of Critical States

- The vast majority of physical processes in nature involve an alternation between steady states and critical states;
- In steady states, system parameters change slowly, making observational data easy to obtain. In critical states, the system undergoes abrupt changes, catching observers off guard, making it difficult to capture observational data.
- Critical state samples in physical processes are often distributed at the boundary of the data manifold, and during the generation process, Sora tends to skip over critical states.
- Indeed, in human cognition, the most crucial observations often pertain to the critical states with nearly zero probability.

Figure: Sora generated the video ”Clone of puppies”

Across Critical States

- Sora generates a video of a group of puppies playing and frolicking, sometimes hiding behind each other, sometimes dispersing. In one moment of the video, the three puppies on the screen suddenly become four.
- The images of the 4 (or 3) puppies form a connected component on the manifold. At the boundary of the 4 puppy branch, there is a critical event: the 4 puppies are partially occluded by each other, and only 3 puppies are visible.
- Sora crosses the boundary of the two branches, 4 puppies suddenly become 3.
- Geometric OT is able to detect the branch boundary and reduce the physical mistakes.

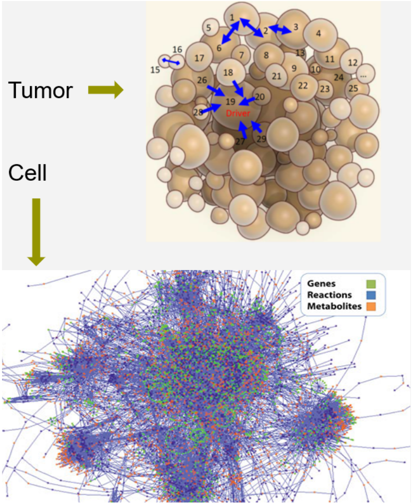
Conclusion

- By manifold distribution principle, the fundamental tasks for generative models are to learn the manifold structure and learn the distribution;
- The manifold fitting method learns the manifold structure; and the transport maps learn the distribution;
- Due the regularity theory of Monge-Ampère equation, the transport maps may have singularity set;
- Some of the physical mistakes made by Sora are due to the absence of critical states, because diffusion model obscures the boundaries of the branches;
- Geometric variational method for optimal transport can detect the boundaries precisely and has the potential to correct mistakes of current generative models.

Part III: Dynamic Systems Modeling of Artificial Intelligence via GLMY Theory

Nature is a complex system

- Complex systems are "complex" in terms of the number of components and interactions.
- A tumor contains many distinct cell types that interact with each other in a complicated way to determine tumor behavior.
- Even for a cell, there are thousands of genes, reactions and metabolites inside, which function as a whole.
- In nature, complex systems occur everywhere.

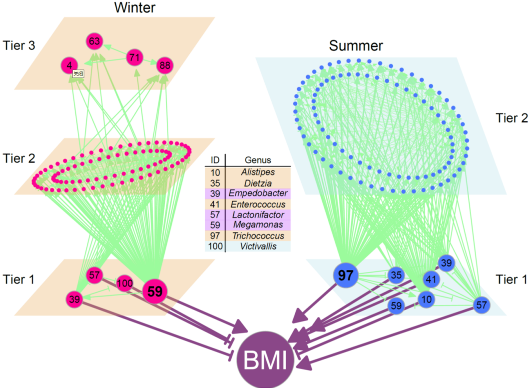


Single-cell analysis (Komarova 2016: Nature)

Dynamic Systems and Its Role in AI

- Networks are fundamental to complex systems.
- Networks can capture how each component interacts with every other component to shape systems dynamics.
- Microbial interactions in the gut microbiota affect human health and diseases.
- Some microbes, such as #59, directly affect BMI, whereas many others, like #4 and #63, are linked with BMI through complex indirect pathways.

The Gut Microbiota



Reanalysis of Davenport et al.’s (PLoS ONE, 2015) data.

A Unified Statistical Mechanics Framework for Network Reconstruction

We integrate evolutionary game theory, allometric scaling law, graph theory, and topology theory to build a unified framework for reconstructing informative, dynamic, omnidirectional, and personalized networks.

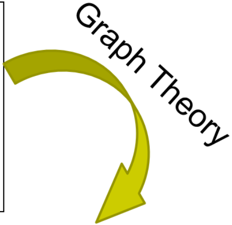
This framework can address the following issues:

- Nonlinear interdependence among components constituting complex systems
- Causal relationships of variables inside and outside complex systems
- Information manifolds from one variable to the next
- Coalescing of all components into a multilayer and multiplex network

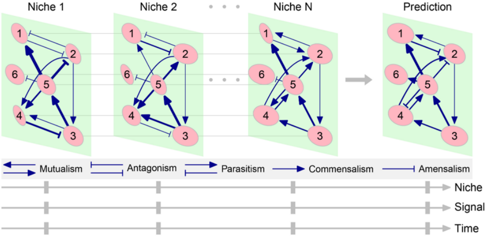
PCA is one commonly used approach for high-dimensional data analysis via linear combination and reduction of variables into orthogonal PC units.

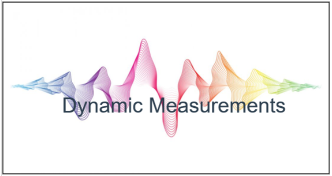
The new framework is more advantages over PCA by capturing all possible causal interrelationships among all variables without information loss.

- Differential equations:
  - Ordinary differential equations (ODE)–simplest
  - Delay differential equations (DDE)
  - Hybrid differential equations (HDE)
  - Partial differential equations (PDE)
  - Stochastic differential equations (SDE)
- Difference equations and state-space models
- Stochastic processes models: branching process etc.
- Agent-based models and cellular automata



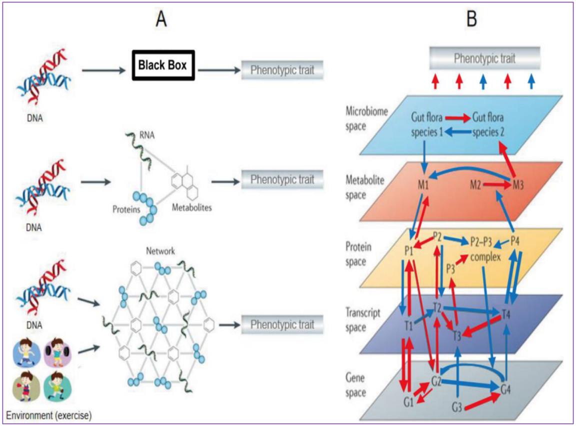
Dynamic Modeling





Informative, dynamic, omnidirectional, and personalized networks

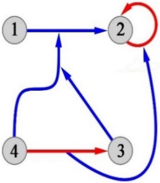
Reconstructing Tri-dimensional Geometry Networks of Biological Processes

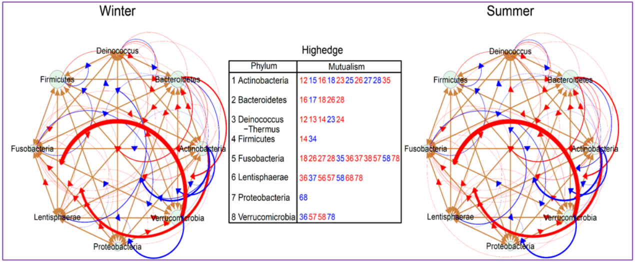


Networks across different spaces from genes to transcripts to proteins to metabolites to microbiomes to phenotypes

Reconstructing High-order Interaction Networks for the gut microbiota

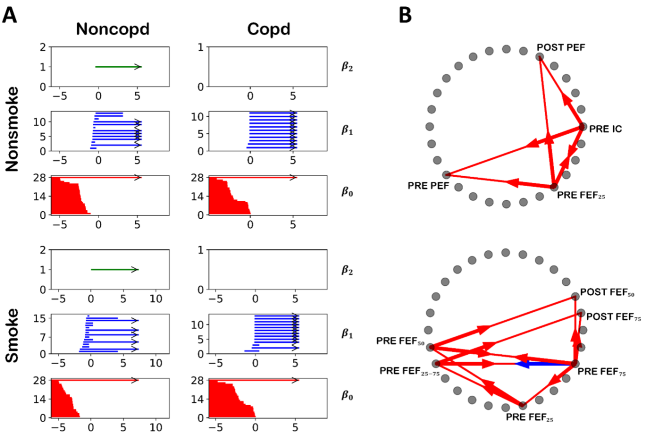
Davenport et al.'s (2015) data including 184 Amish samples from a founder, the Hutterites





Phylum	Mutualism
1 Actinobacteria	12 15 16 18 23 25 26 27 28 35
2 Bacteroidetes	16 17 18 26 28
3 Deinococcus −Thermus	12 13 14 23 24
4 Firmicutes	14 34
5 Fusobacteria	18 26 27 28 35 36 37 38 57 58 78
6 Lentisphaerae	36 37 56 57 58 68 78
7 Proteobacteria	68
8 Verrucomicrobia	36 57 58 78

Integrating GLMY Homology Theory*



GLMY theory can better reveal how healthy subjects (Non-COPD) differ from diseased groups (COPD).

They differ in the pattern of interdependence among lung function traits.

*Grigor'yan A, Lin Y, Muranov Y, Yau S-T (2020) *Path complexes and their homologies*. J Math Sci 248: 564-599.

Concluding Remarks

- Dynamic systems modeling is important for AI.
- Network modeling is a key approach for characterizing complex systems.
- An integrative approach shows its power to reconstruct informative, dynamic, omnidirectional, and personalized networks from any data domains.
- The integration of GLMY theory can leverage network models to reveal hidden patterns of complex systems from big data.