The Past, Present, and Future of Intelligence:

From Artificial Intelligence to Autonomous Intelligence (AI 2.0)

(from Black Box to White Box and from Open Loop to Closed Loop)

Professor Yi Ma

IDS & CS, University of Hong Kong

EECS, University of California, Berkeley

"Everything should be made as simple as possible, but not any simpler."

「一切都应该尽可能简单，但不能更简单。」

-- Albert Einstein

HKU Musketeers Foundation Institute of Data Science 香港大學同心基金數據科學研究院

Berkeley EECS ELECTRICAL ENGINEERING & COMPUTER SCIENCES

80 Years of Evolution of (Artificial) Neural Networks

人工神经网络近 80 年的演变

Beginnings

- 1943 — Thresholded Logic Unit — S. McCulloch - W. Pitts
- 1957 — Perceptron — R. Rosenblatt
- 1960 — Adaline — B. Widrow - M. Hoff

1st Neural Winter

- 1969 — XOR Problem — M. Minsky - S. Papert

- 1982 — Multilayer Backprop — P. Werbos
- 1986 — CNNs — D. Rumelhart - G. Hinton - R. Williams
- 1989 — LSTMs — Y. Lecun J. Schmidhuber

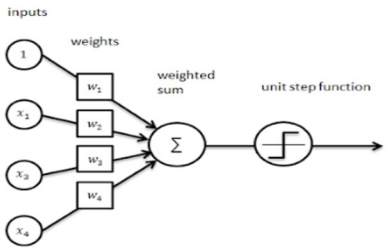
2nd Neural Winter

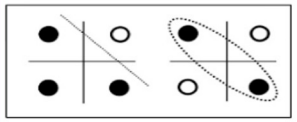
- 1995 — SVMs — C. Cortes - V. Vapnik

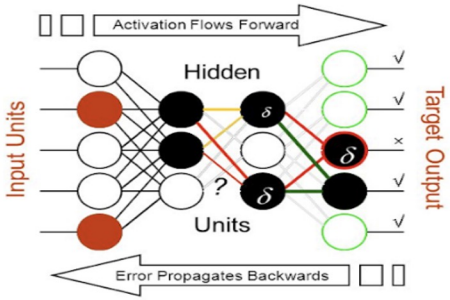
GPU Era

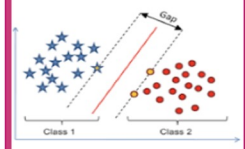
- 2006 — Deep Nets — R. Salakhutdinov - J. Hinton - A. Krizhevsky - I. Sutskever
- 2012 — Alex Net

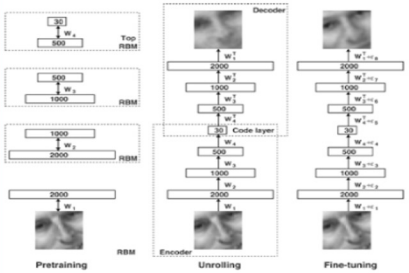
1940 | 1950 | 1960 | 1970 | 1980 | 1990 | 2000 | 2010











Slide courtesy of Prof. Rene Vidal

A Magical Decade for AI Since 2012: Revolution of Deep Network/Learning

2012年以来人工智能的神奇十年:深度网络/学习的革命

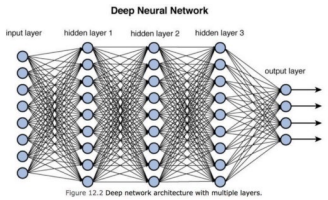
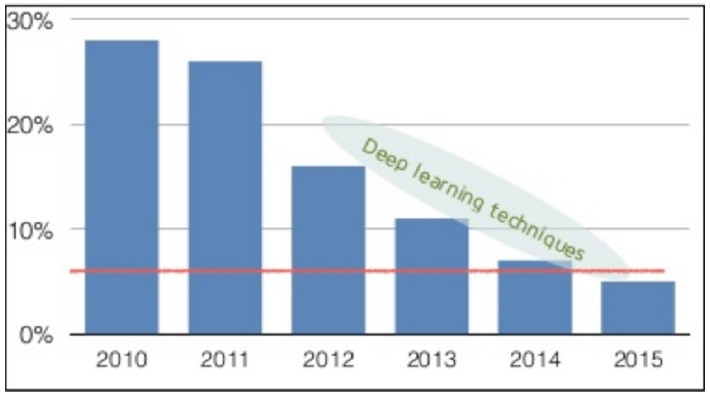
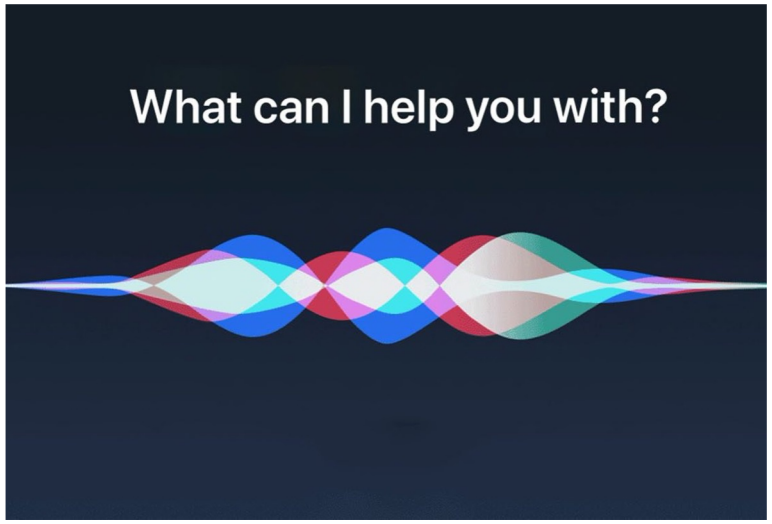
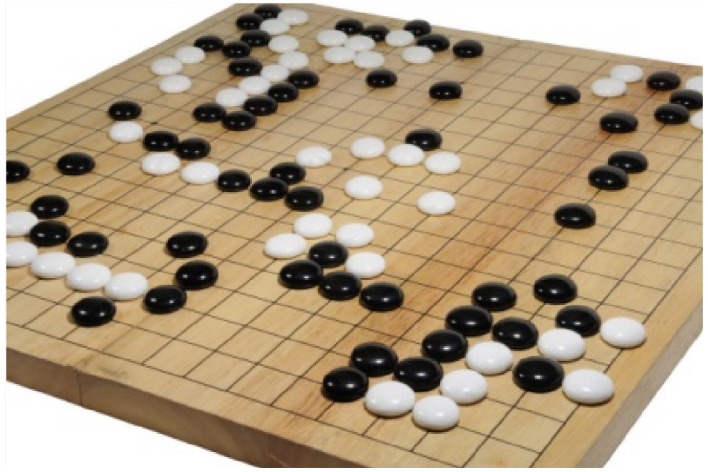
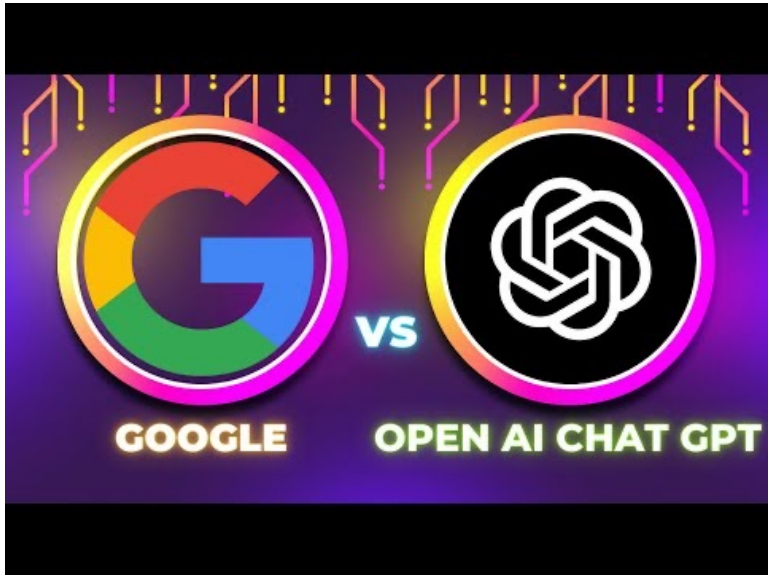


Figure 12.2 Deep network architecture with multiple layers.









This Artificial Intelligence Is Not That "Artificial" Intelligence
对人工智能 (AI)的误解

The Dartmouth Project in 1956?

Organized by **John McCarthy**, Marvin Minsky, Nathaniel Rochester, Claude Shannon with attendees such as Herbert Simon and Allen Newell etc.

That "artificial" intelligence program mainly focuses on **higher-level** intelligent functions:
symbolic, causal, logical deduction, deep inference, or problem solving…

AI in the past 10 years: learning models of the world for perception and prediction.

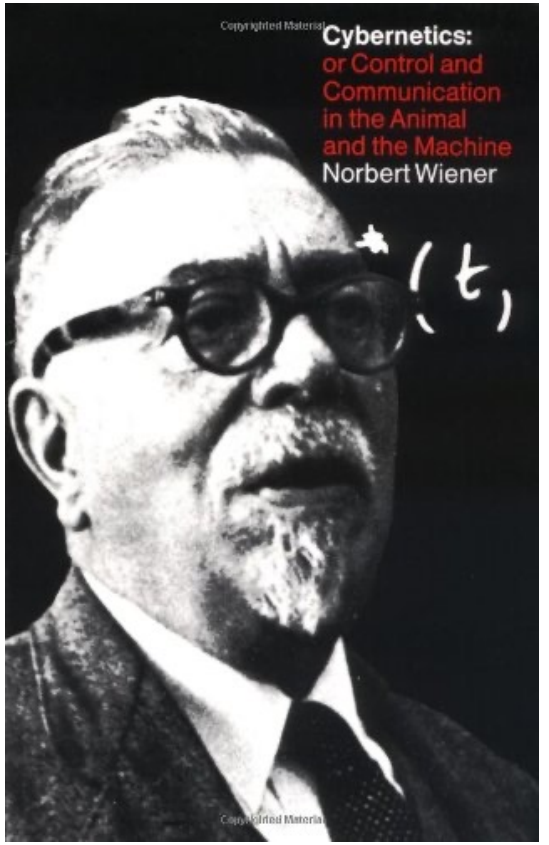
Perceive（感知）

Predict（预测）

Learn/Memory
学习与记忆

A More Magical Decade: the True Origin of Intelligence Study
Intelligence真正神奇的十年：智能研究的真正起源

- 1943, **Artificial Neural Networks**, Warren McCulloch and Walter Pitts
  （1943年，**人工神经网络**，沃伦·麦卡洛克和沃尔特·皮茨）
- 1948, **Information Theory**, Claude Shannon
  （1948年，**信息论**，克劳德·香农）
- 1948, **Feedback Control & Cybernetics**, **Nobert Wiener**
  （1948年，**反馈控制与控制论**，**诺伯特·维纳**）
- 1944, **Game Theory**, John von Neumann
  （1944年，**博弈论**，约翰·冯·诺依曼）
- 1940's, **Turing Machine and Turing Test**, Alan Turing etc.
  （1940年代，**图灵机与图灵测试**，艾伦·图灵等。）

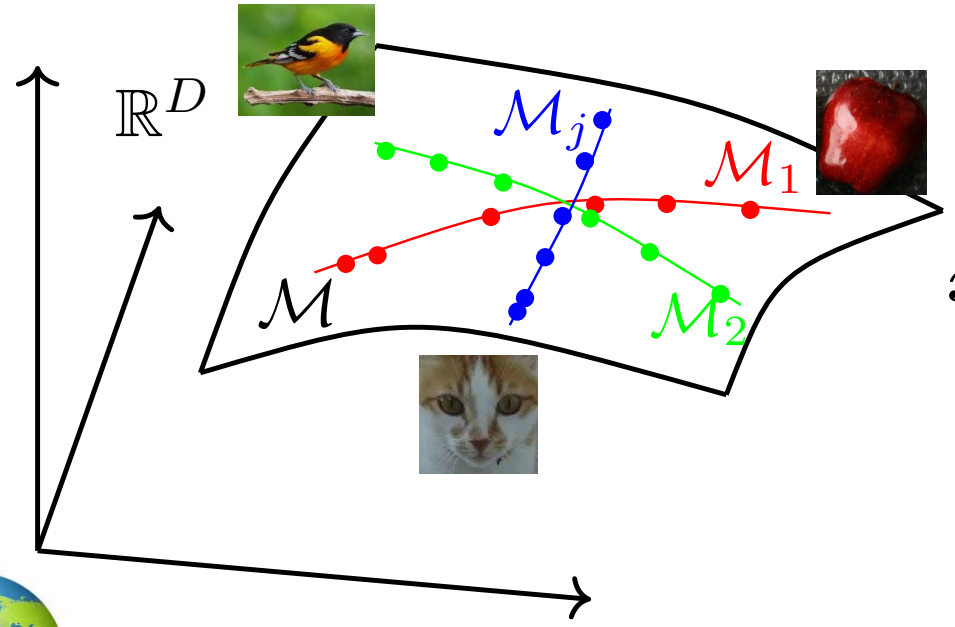


Perceive（感知）

Predict（预测）

Learn/memory
（学习与记忆）

The Magical Decade Since 2012
First Phase: Training Perception/Discriminative Models



$f(\boldsymbol{x}, \theta)$

$\boldsymbol{x}$ → $\boldsymbol{y}$

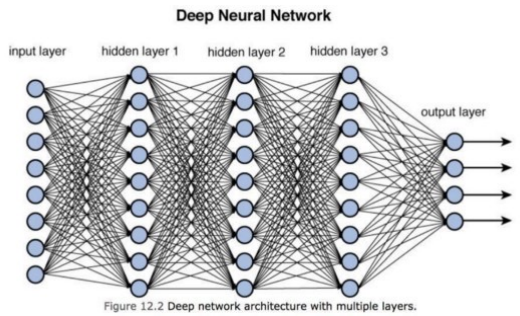


Figure 12.2 Deep network architecture with multiple layers.

$\mathbb{R}^k$

$$\begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix} \quad \begin{bmatrix}0\\1\\\vdots\\0\end{bmatrix} \quad \cdots \quad \begin{bmatrix}0\\0\\\vdots\\1\end{bmatrix}$$

apple bird cat

Perception and Recognition（感知和识别）：

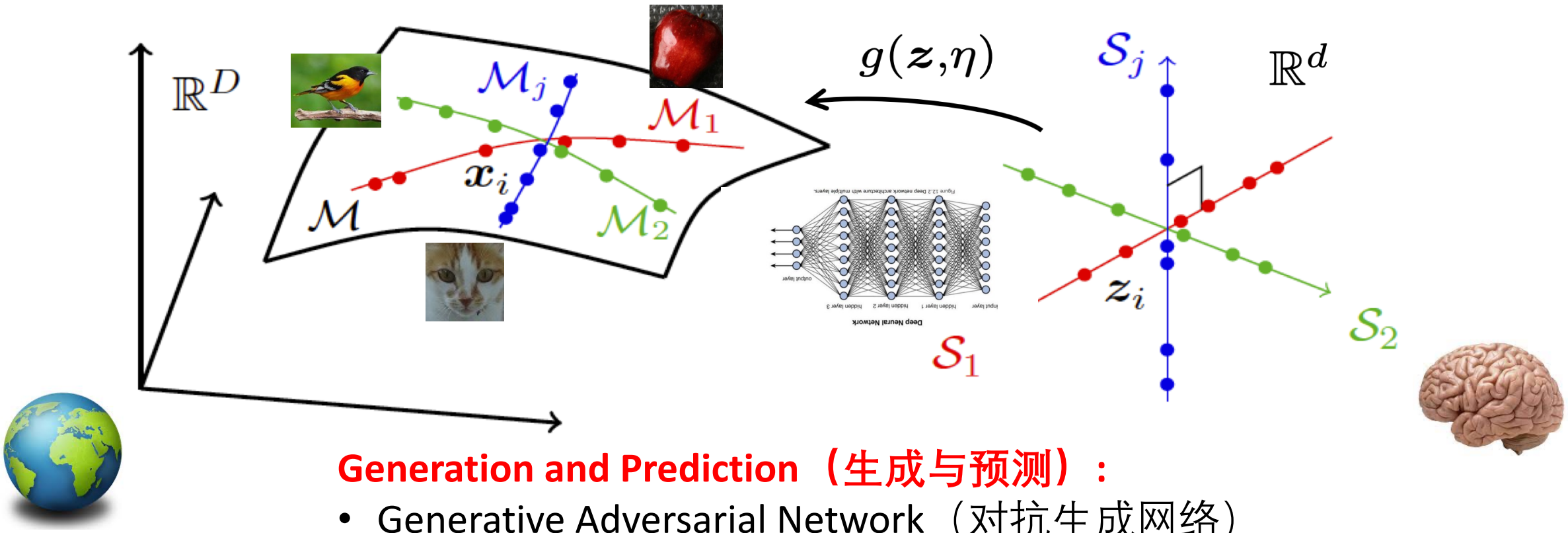
- Speech Recognition（语音识别）
- Object Recognition（物体识别）
- Image Segmentation and Parsing（图像分割与剖析）
- … …

First Phase: Training Perception/Discriminative Models
第一个阶段：训练感知/判别模型

Speech, recognition, object detection, segmentation, and parsing…
语音、识别、对象检测、分割和剖析…

Hey Siri

The Magical Decade Since 2012
Second Phase: Training Generative/Predictive Models



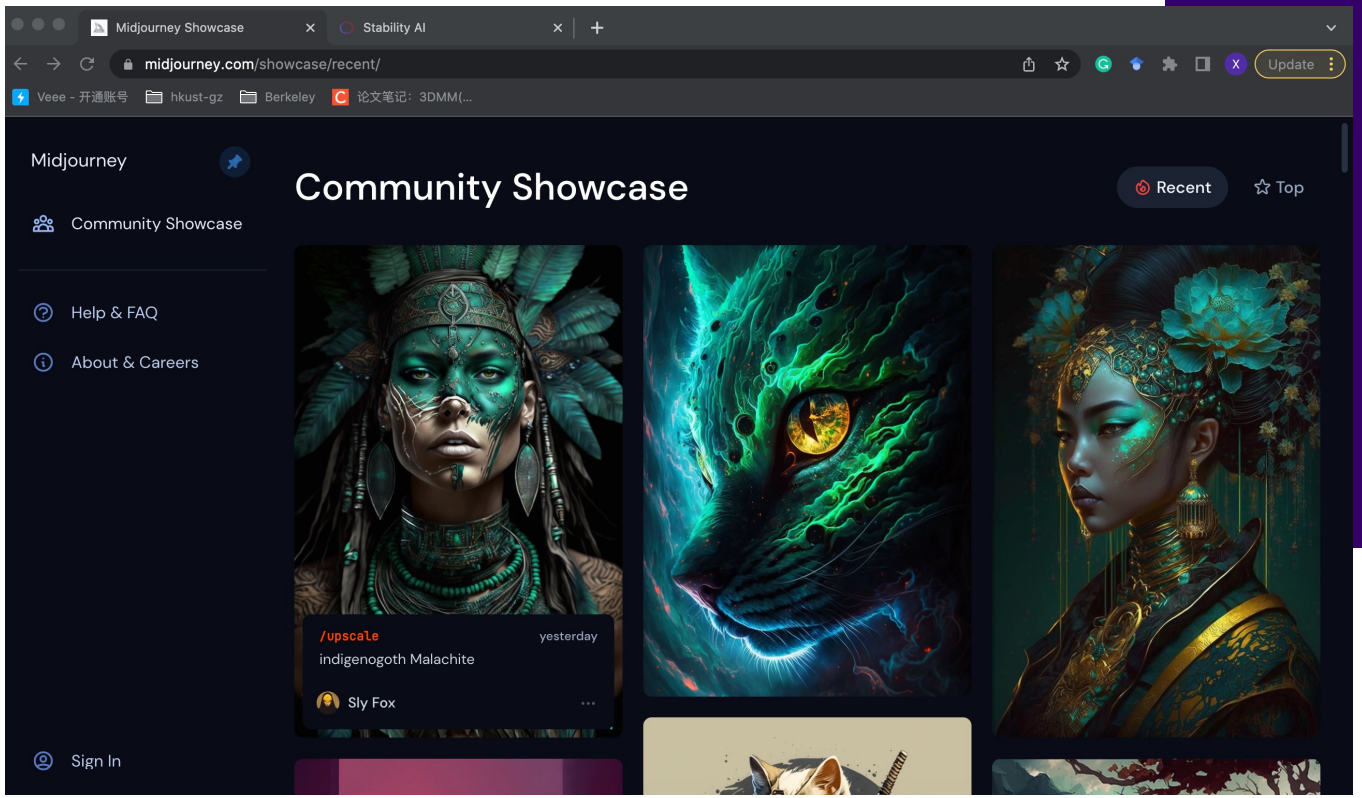
Generation and Prediction（生成与预测）：

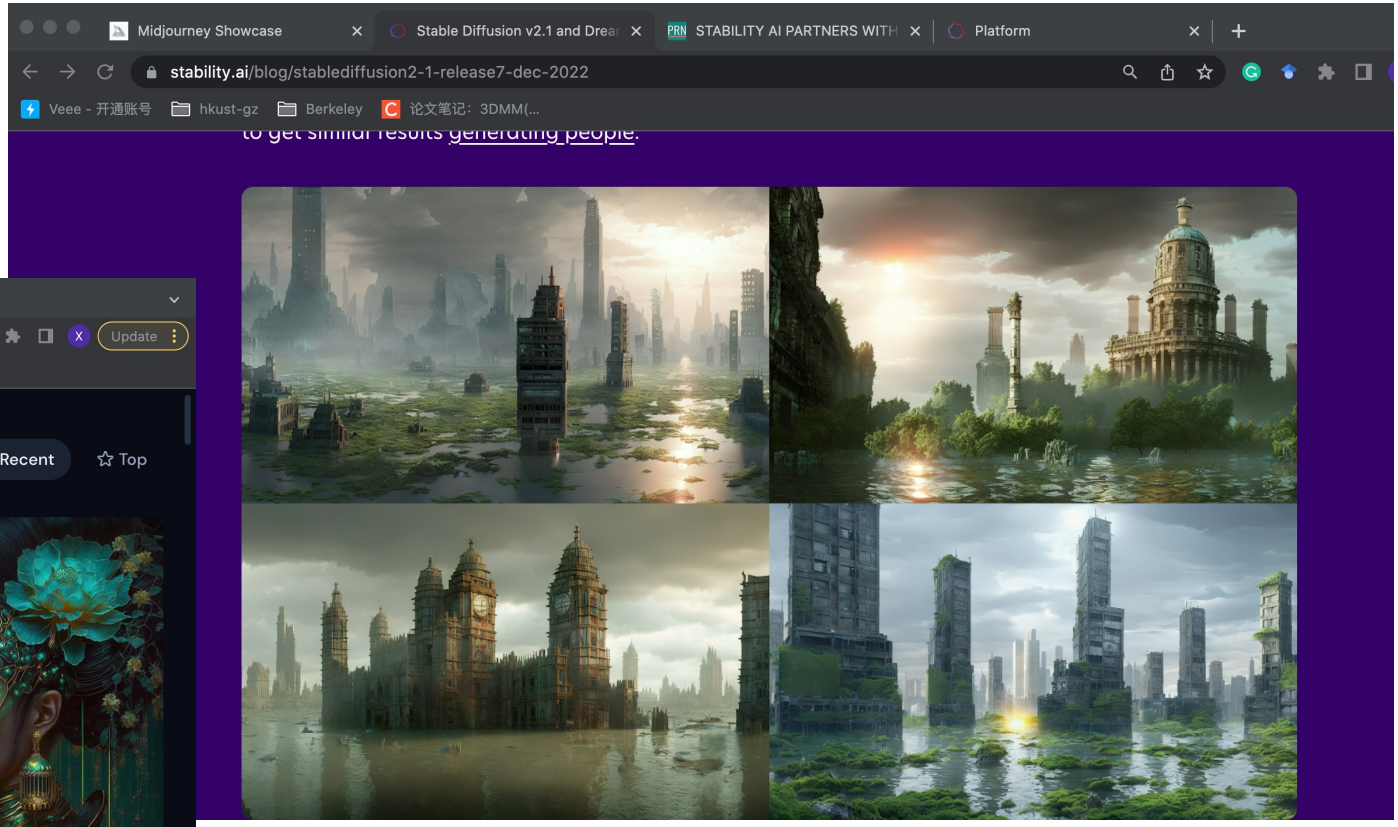
- Generative Adversarial Network（对抗生成网络）
- Diffusion and Denoise Process（扩散与去噪过程）
- Autoencoding and Autoregression（自编码自回归）
- … …

Generating Natural Images: Midjourney or Stability.ai

The Diffusion and Denoising method:
Based on Laplace method or Langevin dynamics
(over 250 years or 100 years ago, respectively)

基于拉普拉斯方法或朗之万动力学
（**分别是250多年前或100多年前**）





Trained with at least billions of images paired with texts, on thousands of GPUs.

在数千个 GPU 上训练至少数十亿配对的图像文本。

Generating Natural Language: BERT or ChatGPT

Method:

1. **Pre-train a large language model (LLM)**

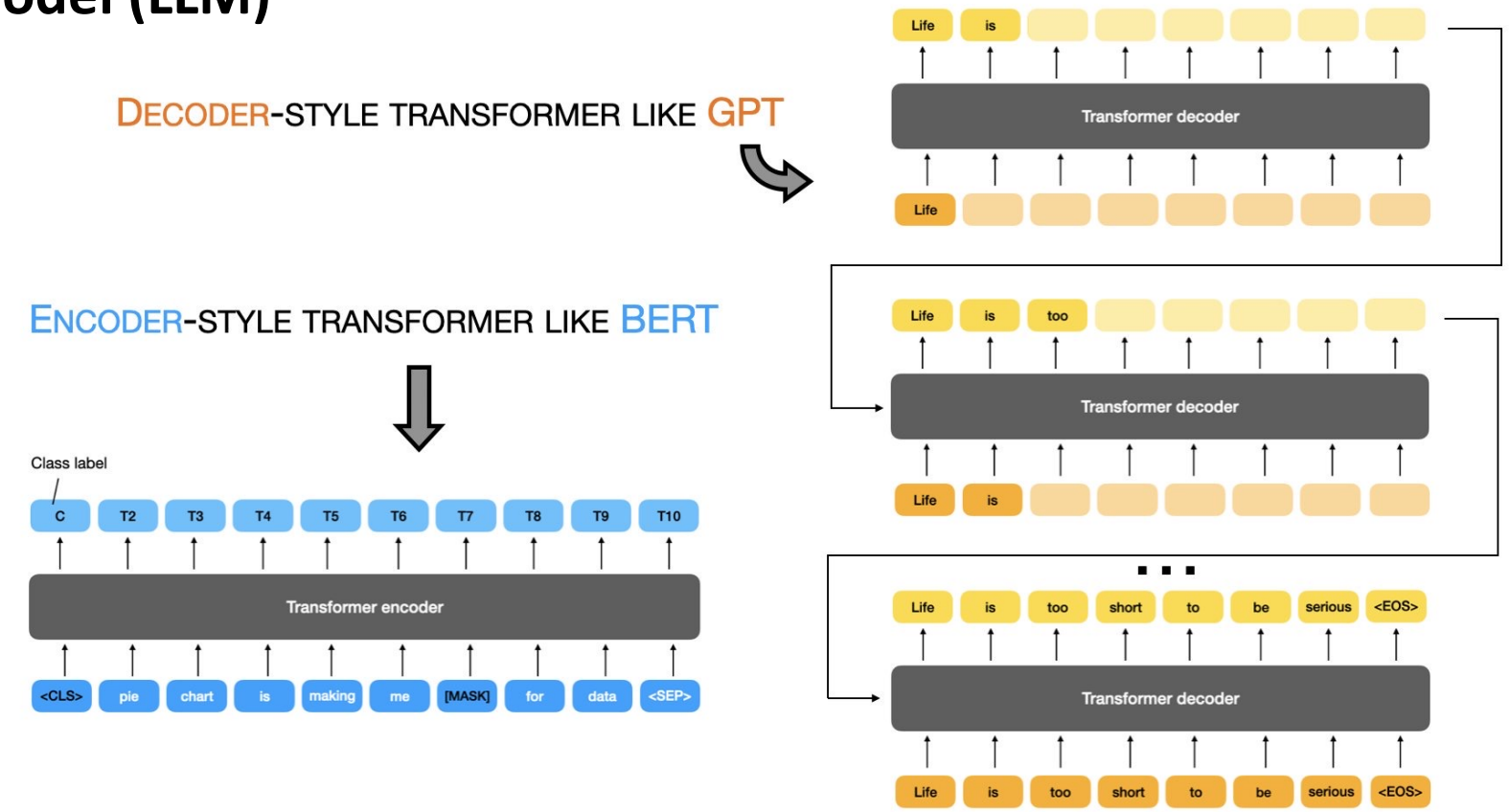
based on “auto-regression”. Two main approaches:

Filling the “blanks” from context (BERT, Google)
完形填空

versus

Predicting the “next word” from Pretext (GPT, OpenAI)
预测下一词

1. 预训练大模型（死记硬背）



Sampling from learned distribution
先做到能一本正经胡说八道

Generating Natural Language: GPT Family

Method:

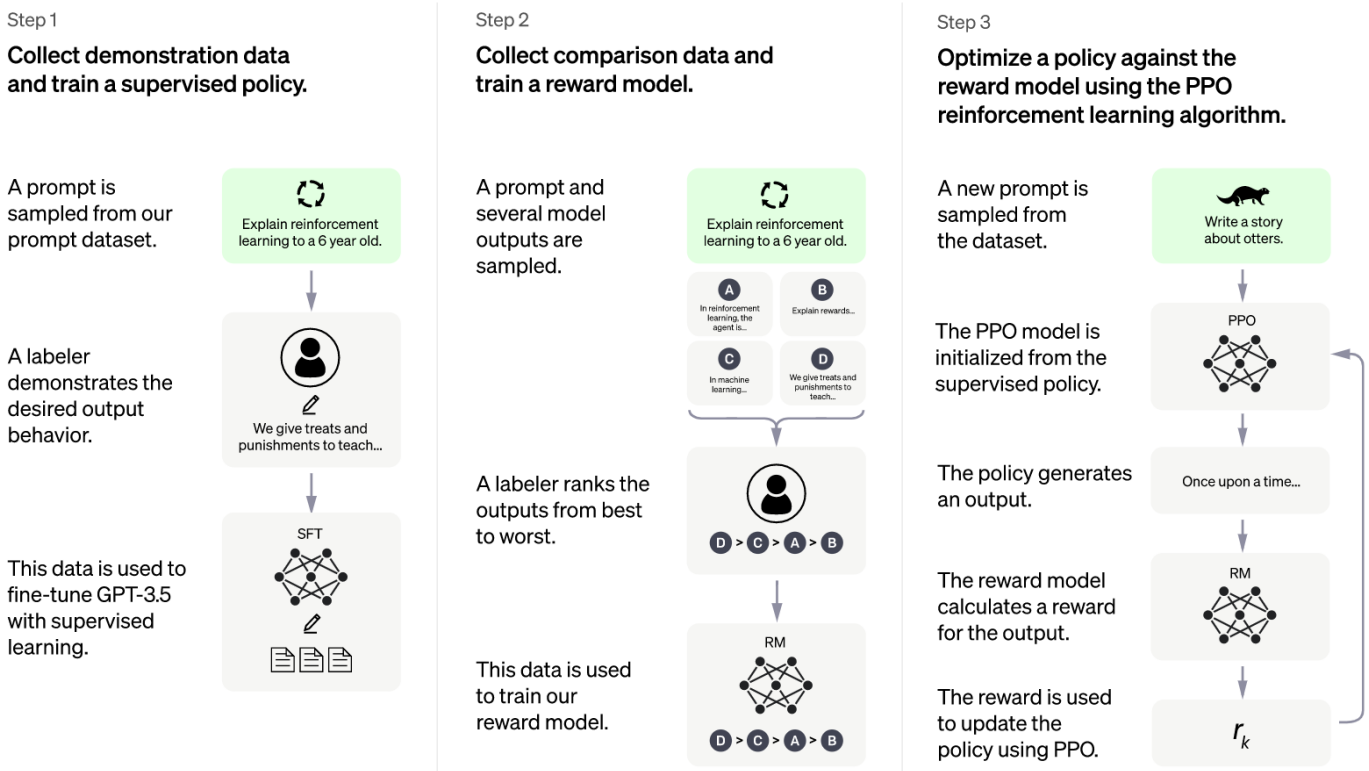
1. Pretrain a large language model (LLM)

2. Finetune or to align with human via supervision and reinforcement

（通过监督与强化学习反馈纠错）

- Supervision with demonstration data
- Learn reward with human ranked data
- Reinforce with learned reward model

2. 模型纠错与人对齐（老师指导）



Enforcing correctness via human feedback

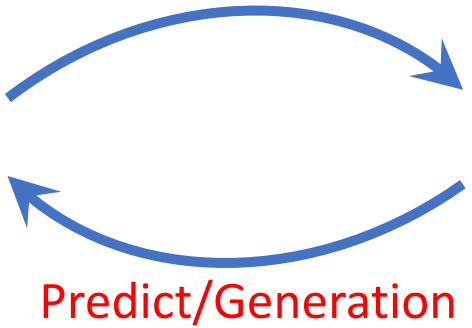
再做到基本符合常识的回答以及推理

What are Missing? Or What are the Next?

- Theory for Deep Networks: from Black Boxes to **White Boxes?**
  深度网络的理论 ：从黑盒到**白盒?**
- Integrated Functions: from Open-Loop to **Closed-Loop?**
  学习模型的功能 ：从开环到**闭环？**
- Science of Intelligence: from Artificial to **Autonomous/Natural?**
  智能的科学 ：从人工到**自主/自然？**

Towards a universal computational mechanism for intelligence at all scales?

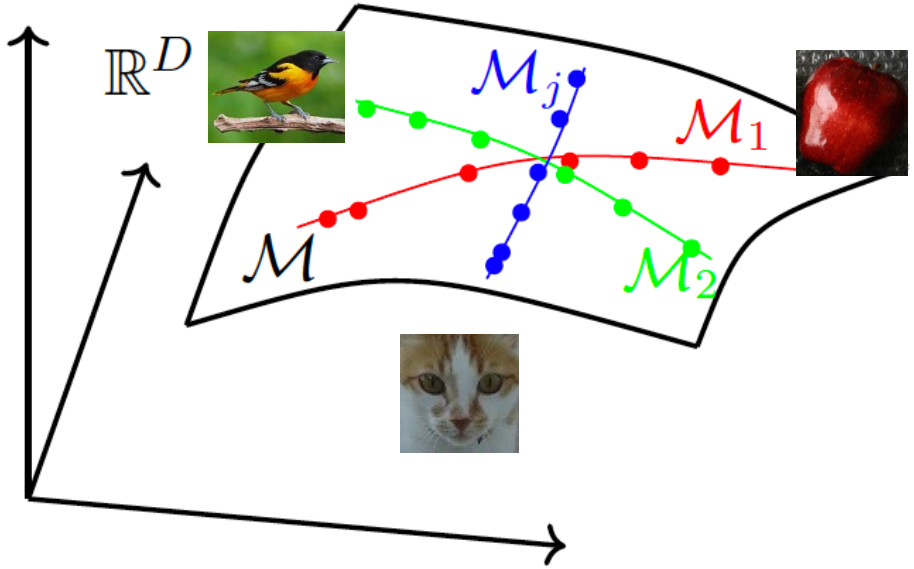
迈向适合所有规模智能系统的通用计算机制？



Learn/Memory

Objective of Learning from High Dimensional Data

Figure: **High-dimensional Real-World Data**: data samples $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m]$ in $\mathbb{R}^D$ lying on a mixture of low-dimensional submanifolds $\boldsymbol{X} \subset \cup_{j=1}^{k} \mathcal{M}_j \subset \mathbb{R}^D$.



The main objective of learning from (samples of) such real-world data:

seek a most compact and structured representation of the data.

为从外界感知的数据寻找一种紧凑和结构化的表示。

Objective of Learning from High Dimensional Data

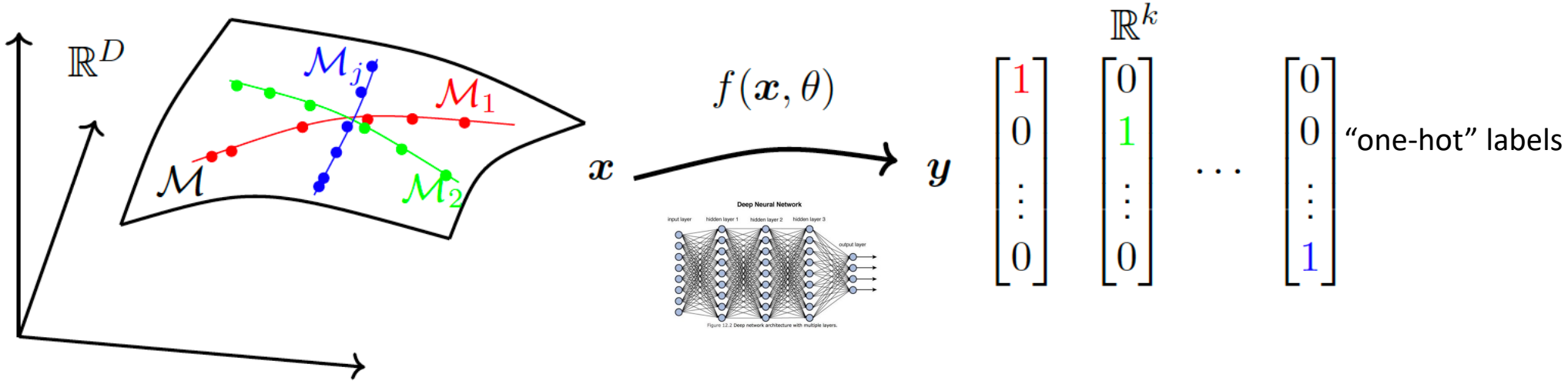
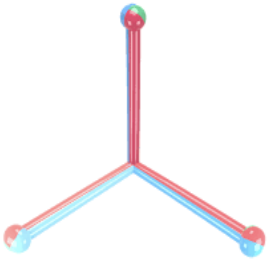


Figure: **Black Box DNN for Classification:** $\boldsymbol{y}$ is the class label of $\boldsymbol{x}$ represented as a “one-hot” vector in $\mathbb{R}^k$. To learn a nonlinear mapping $f(\cdot, \theta) : \boldsymbol{x} \mapsto \boldsymbol{y}$, say modeled by a deep network, using cross-entropy (CE) loss.

$$\min_{\theta \in \Theta} \mathrm{CE}(\theta, \boldsymbol{x}, \boldsymbol{y}) \doteq -\mathbb{E}[\langle \boldsymbol{y}, \log[f(\boldsymbol{x}, \theta)] \rangle] \approx -\frac{1}{m} \sum_{i=1}^{m} \langle \boldsymbol{y}_i, \log[f(\boldsymbol{x}_i, \theta)] \rangle. \qquad (1)$$

Prevalence of **neural collapse** *during the terminal phase of deep learning training*, Papyan, Han, and Donoho, 2020.

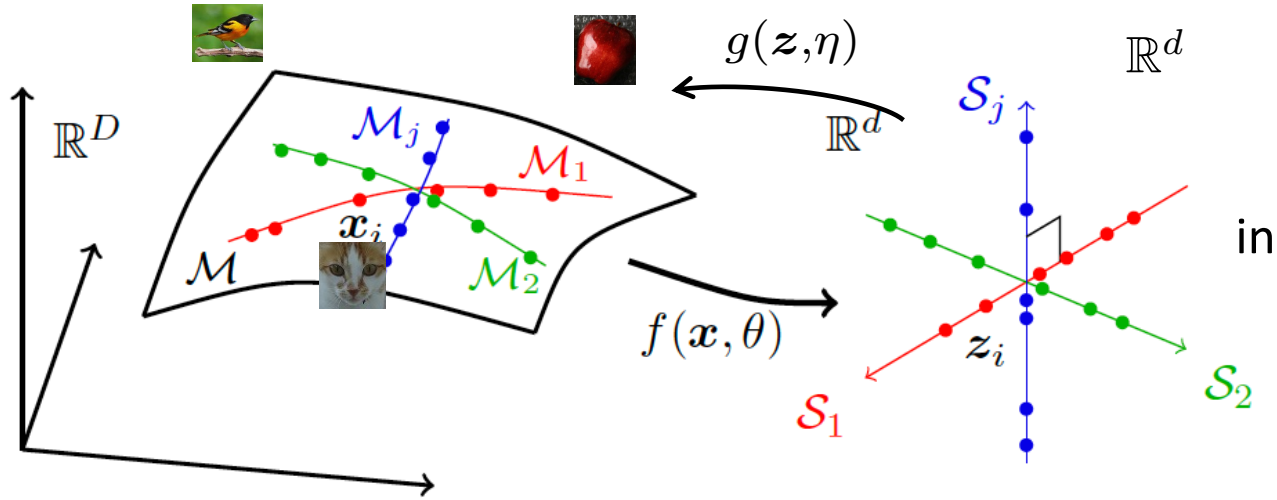


Parsimony: What to Learn from High-Dimensional Data?

简约：要从高维数据中学到什么？

Assumption: the data $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m] \subset \mathbb{R}^D$ lie on one or multiple low-dim submanifolds: $\boldsymbol{X} \subset \cup_{j=1}^{k} \mathcal{M}_j$ in a high-dim space $\in \mathbb{R}^D$:

A mixture of low-dim submanifolds



A mixture of incoherent subspaces

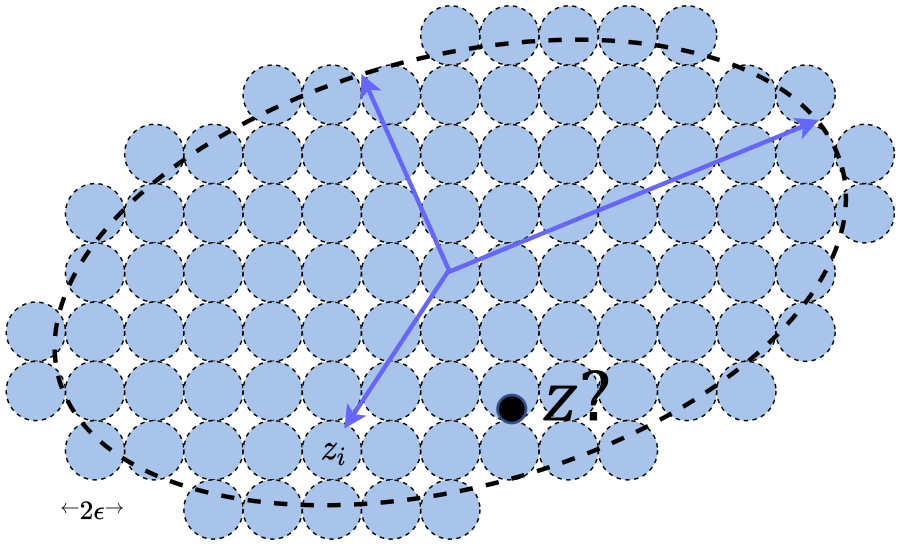
Goal: a linear discriminative representation (LDR) $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m] \subset \mathbb{R}^d$ $(d \ll D)$ for the data $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m] \subset \mathbb{R}^D$ such that:

$$\boldsymbol{X} \subset \mathbb{R}^D \xrightarrow{f(\boldsymbol{x},\theta)} \boldsymbol{Z} \subset \mathbb{R}^d \xrightarrow{g(\boldsymbol{z},\eta)} \hat{\boldsymbol{X}} \approx \boldsymbol{X} \in \mathbb{R}^D.$$

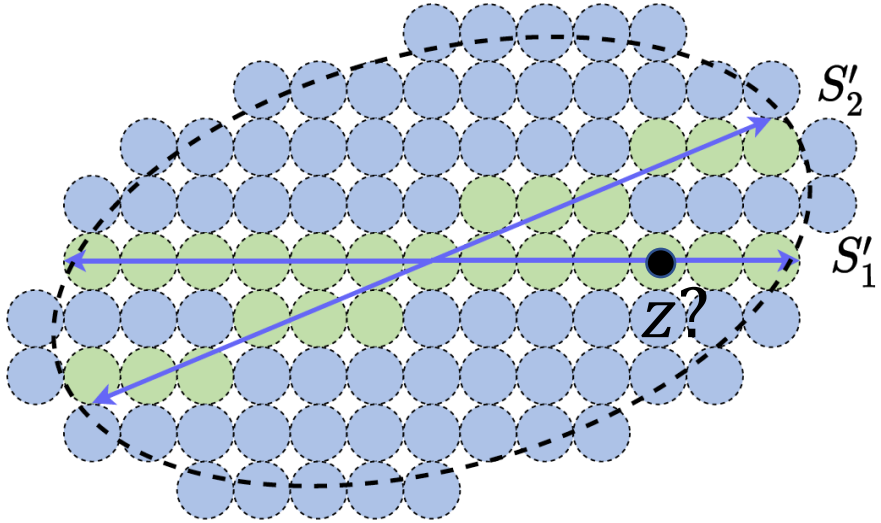
Parsimony: Coding and Organizing Information

Information Gain: How to measure the goodness of a learned representation?

Volume: # spheres packing; **Information**: # bits to specify a sphere



$\mathrm{vol}(Z)$



$\mathrm{vol}(Z')$

rate reduction ΔR =
rate of whole - rate sum of parts

"The whole is more than the sum of its parts."
「整体比部份的和多。」
-- Aristotle, 320BC

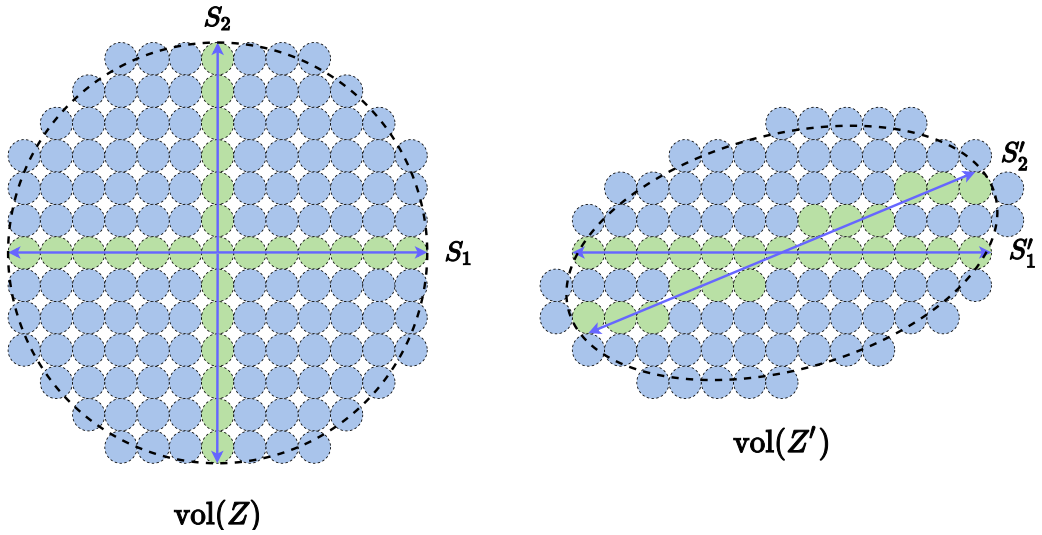
Parsimony: Compact Coding and Structured Representation

Difference in rate distortion between the whole and the parts:

$$\Delta R(\boldsymbol{Z}, \boldsymbol{\Pi}, \epsilon) = \underbrace{\frac{1}{2}\log\det\left(\boldsymbol{I} + \frac{d}{m\epsilon^2}\boldsymbol{Z}\boldsymbol{Z}^\top\right)}_{R(\boldsymbol{Z})} - \underbrace{\sum_{j=1}^{k}\frac{\mathrm{tr}(\boldsymbol{\Pi}_j)}{2m}\log\det\left(\boldsymbol{I} + \frac{d}{\mathrm{tr}(\boldsymbol{\Pi}_j)\epsilon^2}\boldsymbol{Z}\boldsymbol{\Pi}_j\boldsymbol{Z}^\top\right)}_{R^c(\boldsymbol{Z}|\boldsymbol{\Pi},\epsilon)}$$

The whole is **to be maximally** greater than the sum of the parts!

整体要最大限度地大于部分的总和!



The optimal representation **m**aximizes the **c**oding **r**ate **r**eduction (**MCR**^{2}):

$$\max_{\theta}\Delta R\big(\boldsymbol{Z}(\theta), \boldsymbol{\Pi}, \epsilon\big) = R(\boldsymbol{Z}(\theta)) - R^c(\boldsymbol{Z}(\theta)\mid\boldsymbol{\Pi}, \epsilon), \quad \text{s.t.} \quad \boldsymbol{Z}\subset\mathbb{S}^{d-1}.$$

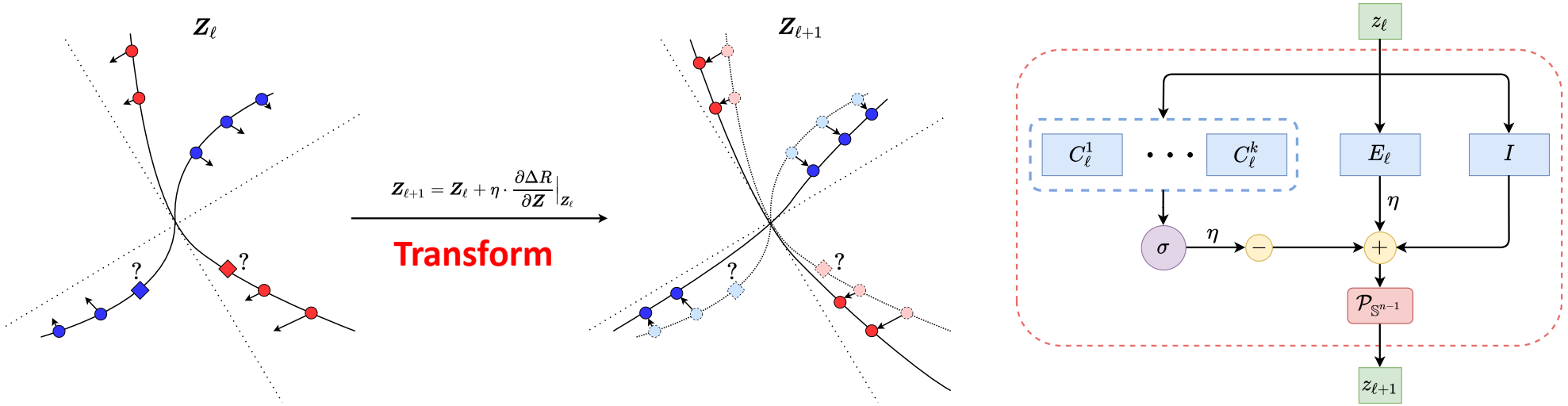
Parsimony: Deep Neural Networks are Nature's Optimization

简约：深度神经网络是自然界对优化算法的实现

A white-box, forward-constructed, multi-channel (convolution) deep neural network from maximizing the rate reduction via projected gradient flow:

$$\boldsymbol{Z}_{\ell+1} \;\propto\; \boldsymbol{Z}_\ell + \eta \cdot \frac{\partial \Delta R(\boldsymbol{Z}, \boldsymbol{\Pi}, \epsilon)}{\partial \boldsymbol{Z}}\bigg|_{\boldsymbol{Z}_\ell} \quad \text{s.t.} \quad \boldsymbol{Z}_\ell \subset \mathbb{S}^{d-1}.$$

$$\frac{\partial R(\boldsymbol{Z})}{\partial \boldsymbol{Z}}\bigg|_{\boldsymbol{Z}_\ell} = \underbrace{\alpha(\boldsymbol{I} + \alpha \boldsymbol{Z}_\ell \boldsymbol{Z}_\ell^*)^{-1}\boldsymbol{Z}_\ell}_{\text{auto-regression residual}} \doteq \boldsymbol{E}_\ell \boldsymbol{Z}_\ell \approx \underbrace{\alpha\big[\boldsymbol{Z}_\ell - \alpha \boldsymbol{Z}_\ell(\boldsymbol{Z}_\ell^* \boldsymbol{Z}_\ell)\big]}_{\text{self-attention head}}.$$

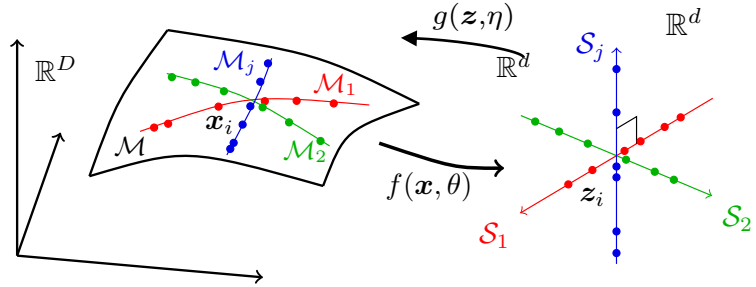


ReduNet: *A Whitebox Deep Network from Rate Reduction* (JMLR, 2022):

Parsimony: White-box Objective, Architecture & Representation

简约：白盒目标、架构和表示

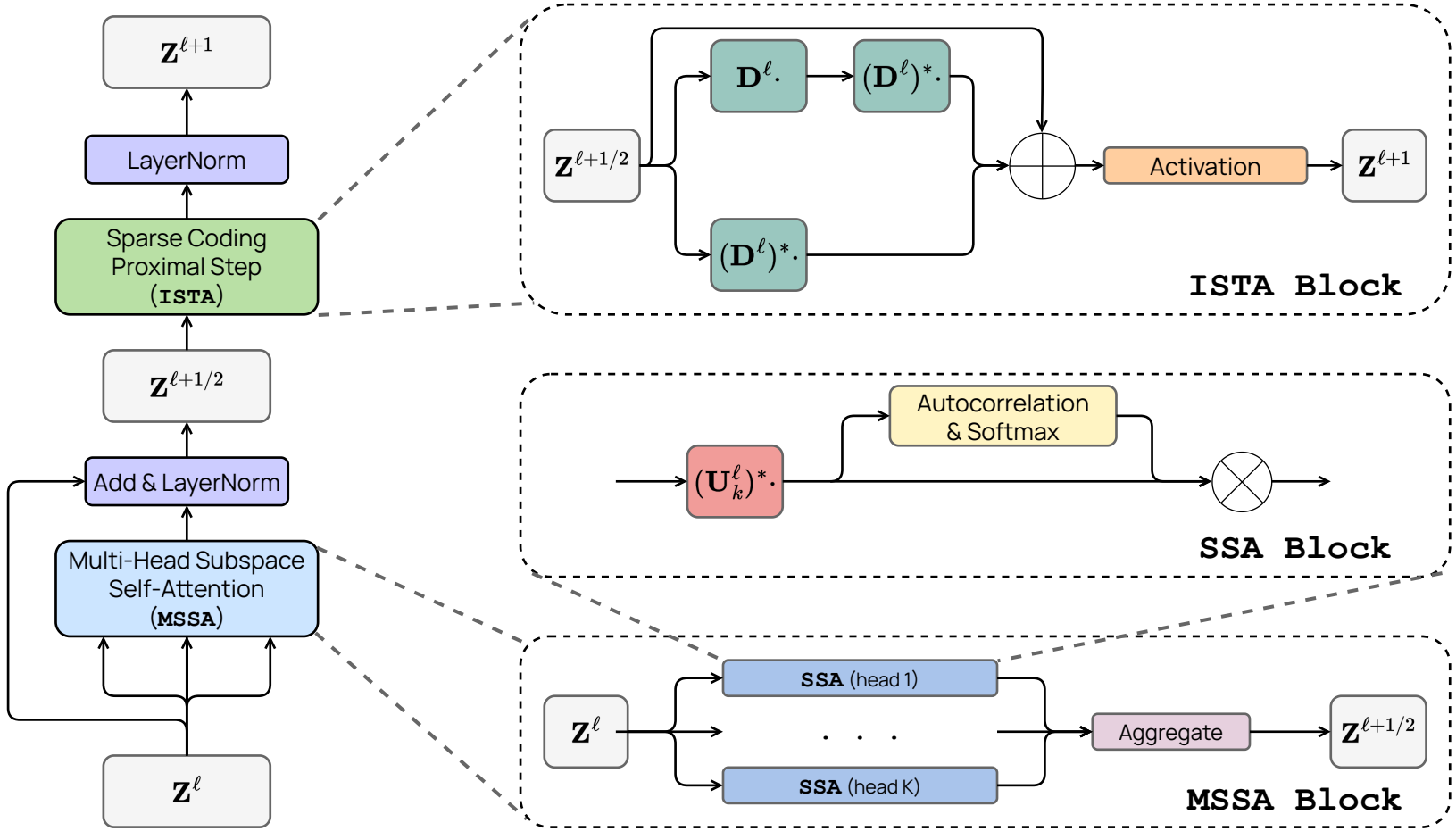
Comparison with conventional practice of NNs (since McCulloch-Pitts'1943).



	Conventional DNNs	ReduNets
Objectives	input/output fitting	information gain
Deep architectures	trial & error	iterative optimization
Layer operators	empirical	projected gradient
Shift invariance	CNNs+augmentation	invariant ReduNets
Initializations	random/pre-design	forward unrolled[1]
Training/fine-tuning	back prop	forward/back prop
Interpretability	black box	white box
Representations	hidden/latent	incoherent subspaces

Parsimony: White-Box Transformer Network

$$\max_{f \in \mathcal{F}} \mathbb{E}_{\boldsymbol{Z}} \left[\Delta R(\boldsymbol{Z}; \boldsymbol{U}_{[K]}) - \lambda \|\boldsymbol{Z}\|_0 \right] \qquad f : \boldsymbol{X} \xrightarrow{f^0} \boldsymbol{Z}^0 \to \cdots \to \boldsymbol{Z}^\ell \xrightarrow{f^\ell} \boldsymbol{Z}^{\ell+1} \to \cdots \to \boldsymbol{Z}^L = \boldsymbol{Z}$$



github.com/Ma-Lab-Berkeley/CRATE

Parsimony: White-Box Transformer Network

$$\max_{f \in \mathcal{F}} \mathbb{E}_{\boldsymbol{Z}} \left[\Delta R(\boldsymbol{Z}; \boldsymbol{U}_{[K]}) - \lambda \|\boldsymbol{Z}\|_0 \right] \qquad f: \boldsymbol{X} \xrightarrow{f^0} \boldsymbol{Z}^0 \to \cdots \to \boldsymbol{Z}^\ell \xrightarrow{f^\ell} \boldsymbol{Z}^{\ell+1} \to \cdots \to \boldsymbol{Z}^L = \boldsymbol{Z}$$

Table 1: Top 1 accuracy of CRATE on various datasets with different model scales when pre-trained on ImageNet. For ImageNet/ImageNetReaL, we directly evaluate the top-1 accuracy. For other datasets, we use models that are pre-trained on ImageNet as initialization and the evaluate the transfer learning performance via fine-tuning.

Datasets	CRATE-T	CRATE-S	CRATE-B	CRATE-L	ViT-T	ViT-S
# parameters	6.09M	13.12M	22.80M	77.64M	5.72M	22.05M
ImageNet	66.7	69.2	70.8	71.3	71.5	72.4
ImageNet ReaL	74.0	76.0	76.5	77.4	78.3	78.4
CIFAR10	95.5	96.0	96.8	97.2	96.6	97.2
CIFAR100	78.9	81.0	82.7	83.6	81.8	83.2
Oxford Flowers-102	84.6	87.1	88.7	88.3	85.1	88.5
Oxford-IIIT-Pets	81.4	84.9	85.3	87.4	88.5	88.6

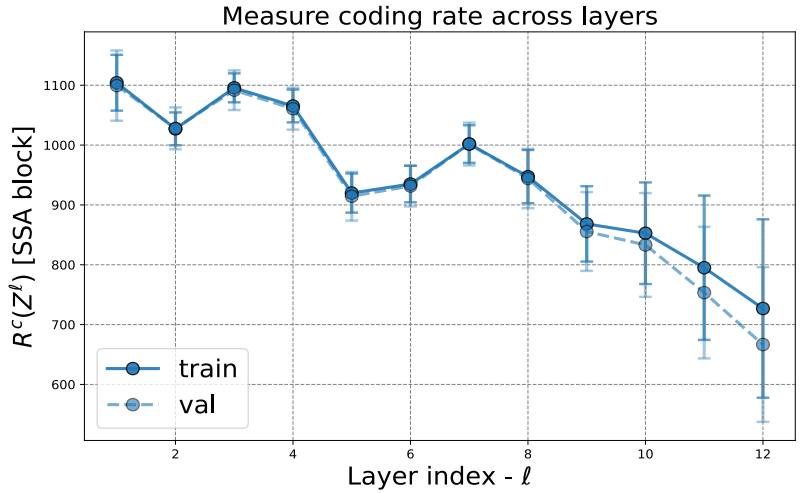
White-box Transformers via Sparse Rate Reduction, under submission.

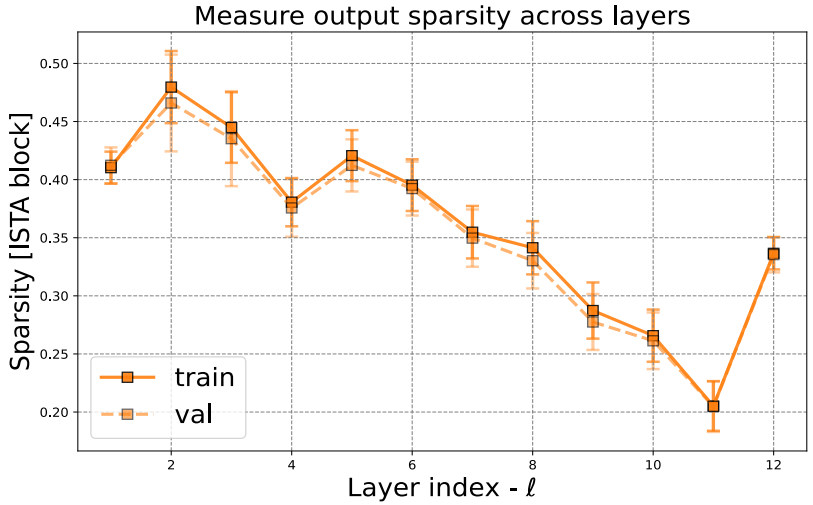
github.com/Ma-Lab-Berkeley/CRATE

Parsimony: White-Box Transformer Network

$$\max_{f \in \mathcal{F}} \mathbb{E}_{\boldsymbol{Z}} \left[\Delta R(\boldsymbol{Z}; \boldsymbol{U}_{[K]}) - \lambda \|\boldsymbol{Z}\|_0 \right] \qquad f : \boldsymbol{X} \xrightarrow{f^0} \boldsymbol{Z}^0 \to \cdots \to \boldsymbol{Z}^\ell \xrightarrow{f^\ell} \boldsymbol{Z}^{\ell+1} \to \cdots \to \boldsymbol{Z}^L = \boldsymbol{Z}$$

Given a learned CRATE model, we measure the compression term of $\boldsymbol{Z}^{\ell+1/2}$ (*left*, $R^c(\boldsymbol{Z}^{\ell+1/2})$) and the sparsification term of $\boldsymbol{Z}^{\ell+1}$ (*right*, $\|\boldsymbol{Z}^{\ell+1}\|_0$) on train/validation samples at **each layer**.





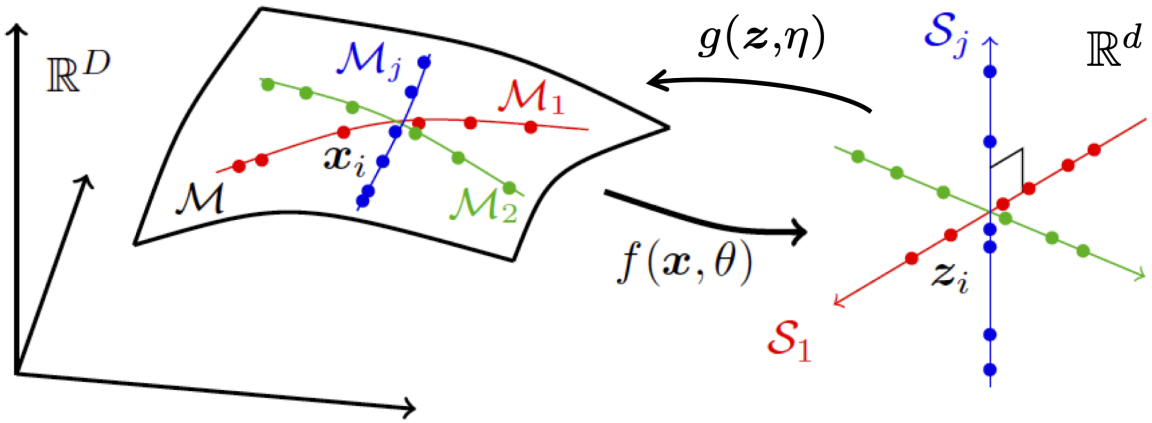
- The learned CRATE model indeed performs its design objective – each layer incrementally optimizes the compression term and the sparsification term.

github.com/Ma-Lab-Berkeley/CRATE

Self-Consistency: How to Learn Correctly & Autonomously?

自洽：如何自主地学习到正确的模型？

A mixture of low-dim submanifolds



A mixture of incoherent subspaces

Goal: Transcribe the data $\boldsymbol{X} \subset \cup_{j=1}^{k} \mathcal{M}_j$ onto an LDR $\boldsymbol{Z} \subset \cup_{j=1}^{k} \mathcal{S}_j$:

$$\underbrace{f(\mathcal{M}_j) = \mathcal{S}_j}_{\text{linear}} \quad \text{with} \quad \underbrace{\mathcal{S}_i \perp \mathcal{S}_j}_{\text{discriminative}} \quad \text{and} \quad \underbrace{g(f(\mathcal{M}_j)) = \mathcal{M}_j}_{\text{auto-embedding}}.$$

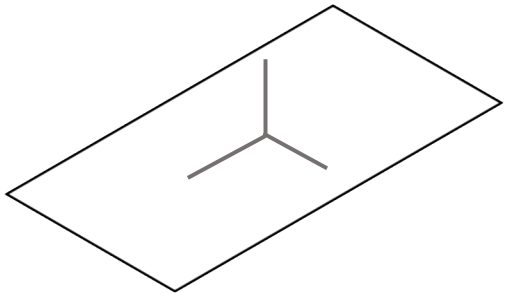
Autoencoding of multiple low-dim nonlinear submanifolds:

$$\boldsymbol{X} \subset \cup_{j=1}^{k} \mathcal{M}_j \xrightarrow{f(\boldsymbol{x},\theta)} \cup_{j=1}^{k} \boldsymbol{Z}_j \subset \mathcal{S}_j \xrightarrow{g(\boldsymbol{z},\eta)} \hat{\boldsymbol{X}} \subset \cup_{j=1}^{k} \mathcal{M}_j.$$

Self-Consistency: How to Learn Correctly & Autonomously?

One low-dim linear subspace: principal component analysis (PCA)

$$\boldsymbol{X} \subset \mathcal{S}^D \xrightarrow{\boldsymbol{V}^T} \boldsymbol{Z} \subset \mathcal{S}^d \xrightarrow{\boldsymbol{V}} \hat{\boldsymbol{X}} \subset \mathcal{S}^D.$$

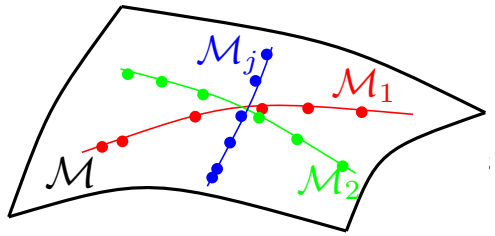


Solve the following optimization problem:

$$\min_{\boldsymbol{V}} \|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_2^2 \quad \text{s.t.} \quad \hat{\boldsymbol{X}} = \boldsymbol{V}\boldsymbol{V}^T\boldsymbol{X}, \quad \boldsymbol{V} \in \mathsf{O}(D,d).$$

One low-dim nonlinear submanifold: Nonlinear PCA

$$\boldsymbol{X} \subset \mathcal{M}^D \xrightarrow{f(\boldsymbol{x},\theta)} \boldsymbol{Z} \subset \mathcal{S}^d \xrightarrow{g(\boldsymbol{z},\eta)} \hat{\boldsymbol{X}} \subset \mathcal{M}^D.$$



Solve the following optimization problem:

$$\min_{\theta,\eta} \underbrace{\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_2^2}_{d(\boldsymbol{X},\hat{\boldsymbol{X}})^2} \quad \text{s.t.} \quad \hat{\boldsymbol{X}} = g(f(\boldsymbol{X},\eta),\theta).$$

What is the right distance $d(\boldsymbol{X},\hat{\boldsymbol{X}})$, say for images?

Self-Consistency: How to Learn Correctly & Autonomously?

Learning generative models via **discriminative** approaches? (Tu'2007)

Generative Adversarial Nets (GAN) (Goodfellow'2014):

$$\boldsymbol{Z} \xrightarrow{g(\boldsymbol{z},\eta)} \hat{\boldsymbol{X}}, \boldsymbol{X} \xrightarrow{d(\boldsymbol{x},\theta)} \mathbf{0}, \mathbf{1}.$$

A minimax game between generator and discriminator:

$$\min_{\eta} \max_{\theta} \mathbb{E}_{p(\boldsymbol{x})}\big[\log d(\boldsymbol{x},\theta)\big] + \mathbb{E}_{p(\boldsymbol{z})}\big[1 - \log d(\underbrace{g(\boldsymbol{z},\eta)}_{\hat{\boldsymbol{x}} \sim p_g},\theta)\big].$$

This is equivalent to minimize the *Jensen-Shannon divergence*:

$$\mathcal{D}_{JS}(p, p_g) = \mathcal{D}_{KL}\big(p \| (p+p_g)/2\big) + \mathcal{D}_{KL}\big(p_g \| (p+p_g)/2\big).$$

But the J-S divergence is extremely difficult, if not impossible, to compute and optimize.

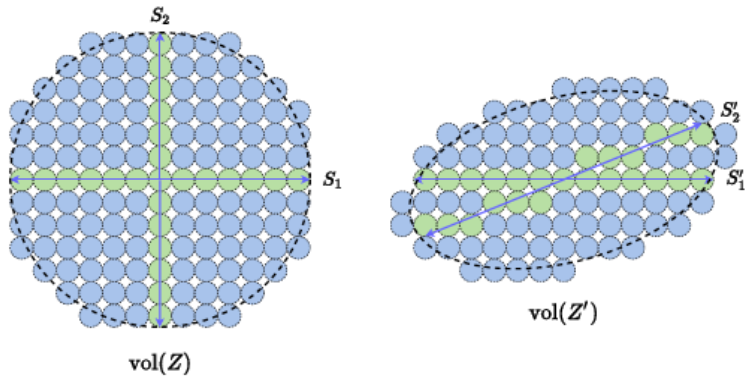
Self-Consistency: How to Learn Correctly & Autonomously?

Rate reduction ΔR gives a closed-form distance for (non-overlapping) mixture of subspaces/Gaussians!

$$\Delta R(\boldsymbol{Z}, \boldsymbol{\Pi}, \epsilon) = \underbrace{\frac{1}{2}\log\det\left(\boldsymbol{I} + \frac{d}{m\epsilon^2}\boldsymbol{Z}\boldsymbol{Z}^\top\right)}_{R} - \underbrace{\sum_{j=1}^{k}\frac{\mathrm{tr}(\boldsymbol{\Pi}_j)}{2m}\log\det\left(\boldsymbol{I} + \frac{d}{\mathrm{tr}(\boldsymbol{\Pi}_j)\epsilon^2}\boldsymbol{Z}\boldsymbol{\Pi}_j\boldsymbol{Z}^\top\right)}_{R^c}.$$

A good measure for the (LDR-like) features $\boldsymbol{Z}$, but what about $d(\boldsymbol{X}, \hat{\boldsymbol{X}})$?

$$\boldsymbol{X} \xrightarrow{f(\boldsymbol{x},\theta)} \boldsymbol{Z} \xrightarrow{g(\boldsymbol{z},\eta)} \hat{\boldsymbol{X}}.$$



一个大问题：我们是否需要在数据 x 空间中进行测量？

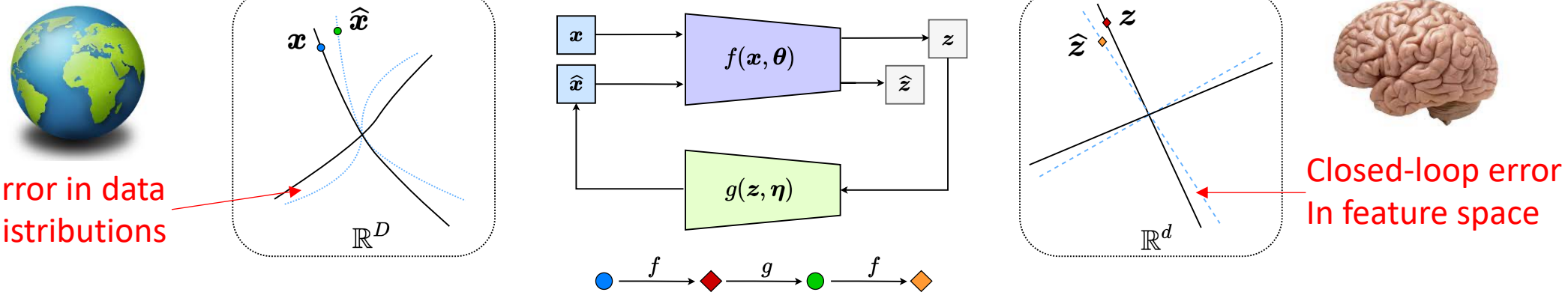
A BIG question: do we ever need to measure in the data x space?

Self-Consistency: Closed-Loop Error Feedback

自洽：闭环反馈纠错

Is it possible to measure everything only in the feature z space?

$$\boldsymbol{X} \xrightarrow{f(\boldsymbol{x},\theta)} \boldsymbol{Z} \xrightarrow{g(\boldsymbol{z},\eta)} \hat{\boldsymbol{X}} \xrightarrow{f(\boldsymbol{x},\theta)} \hat{\boldsymbol{Z}}.$$



Yes! Measure difference in $\boldsymbol{X}_j$ and $\hat{\boldsymbol{X}}_j$ through their features $\boldsymbol{Z}_j$ and $\hat{\boldsymbol{Z}}_j$:

$$\boldsymbol{X}_j \xrightarrow{f(\boldsymbol{x},\theta)} \boldsymbol{Z}_j \xrightarrow{g(\boldsymbol{z},\eta)} \hat{\boldsymbol{X}}_j \xrightarrow{f(\boldsymbol{x},\theta)} \hat{\boldsymbol{Z}}_j, \quad j = 1,\ldots,k.$$

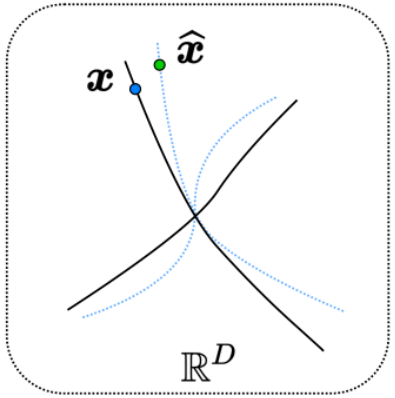
with "their distance" measured by the **rate reduction**:

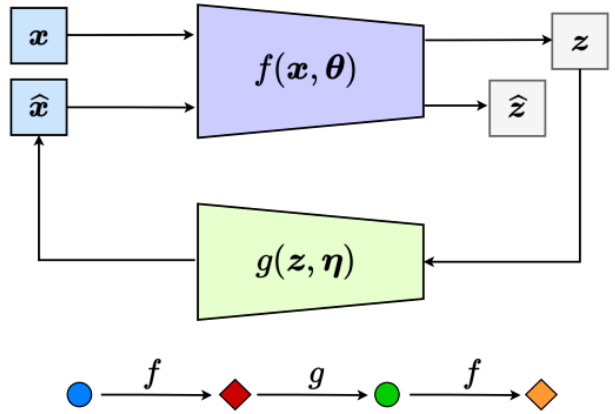
$$\Delta R(\boldsymbol{Z}_j, \hat{\boldsymbol{Z}}_j) \doteq R(\boldsymbol{Z}_j \cup \hat{\boldsymbol{Z}}_j) - \frac{1}{2}\big(R(\boldsymbol{Z}_j) + R(\hat{\boldsymbol{Z}}_j)\big), \quad j = 1,\ldots,k.$$

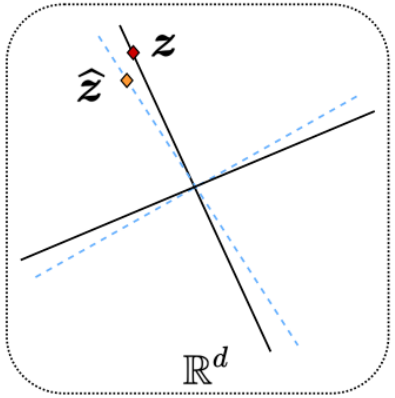
Self-Consistency: Closed-Loop Self-Critiquing Game

自洽：闭环博弈纠错

Just close the loop and minimize the error is not enough!







$f(x,\theta)$ is compressive

Decoder/controller g minimizes the difference between $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$:

$$d(\boldsymbol{X},\hat{\boldsymbol{X}}) \doteq \min_{\eta} \sum_{j=1}^{k} \Delta R\big(\boldsymbol{Z}_j,\hat{\boldsymbol{Z}}_j\big) = \min_{\eta} \sum_{j=1}^{k} \Delta R\big(\boldsymbol{Z}_j, f(g(\boldsymbol{Z}_j,\eta),\theta)\big).$$

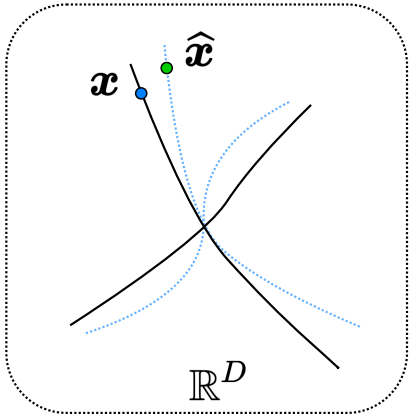
Encoder/sensor f amplifies any difference between $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$:

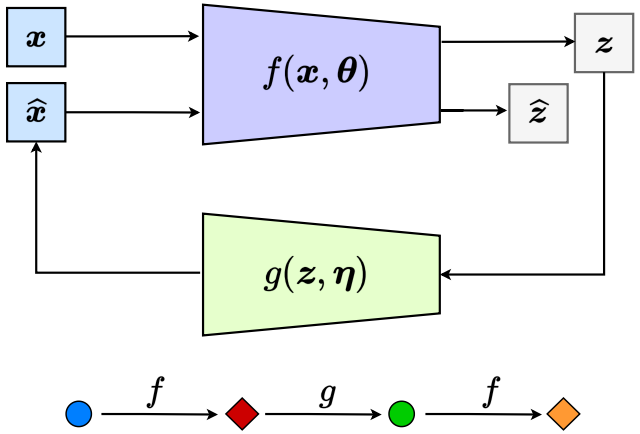
$$d(\boldsymbol{X},\hat{\boldsymbol{X}}) \doteq \max_{\theta} \sum_{j=1}^{k} \Delta R\big(\boldsymbol{Z}_j,\hat{\boldsymbol{Z}}_j\big) = \max_{\theta} \sum_{j=1}^{k} \Delta R\big(f(\boldsymbol{X}_j,\theta), f(\hat{\boldsymbol{X}}_j,\theta)\big).$$

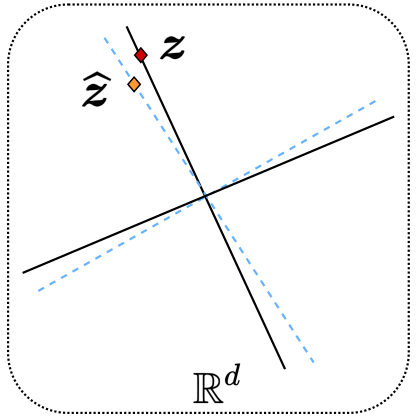
Self-Consistency: Closed-Loop Feedback and Game

自洽：闭环反馈与博弈纠错

f is both an encoder and sensor; and g is both a decoder and controller. They form a closed-loop system for feedback and game:







A closed-loop notion of "**self-consistency**" between $\boldsymbol{Z}$ and $\hat{\boldsymbol{Z}}$ is achieved by a self-critiquing game between the sensor f and the generator g:

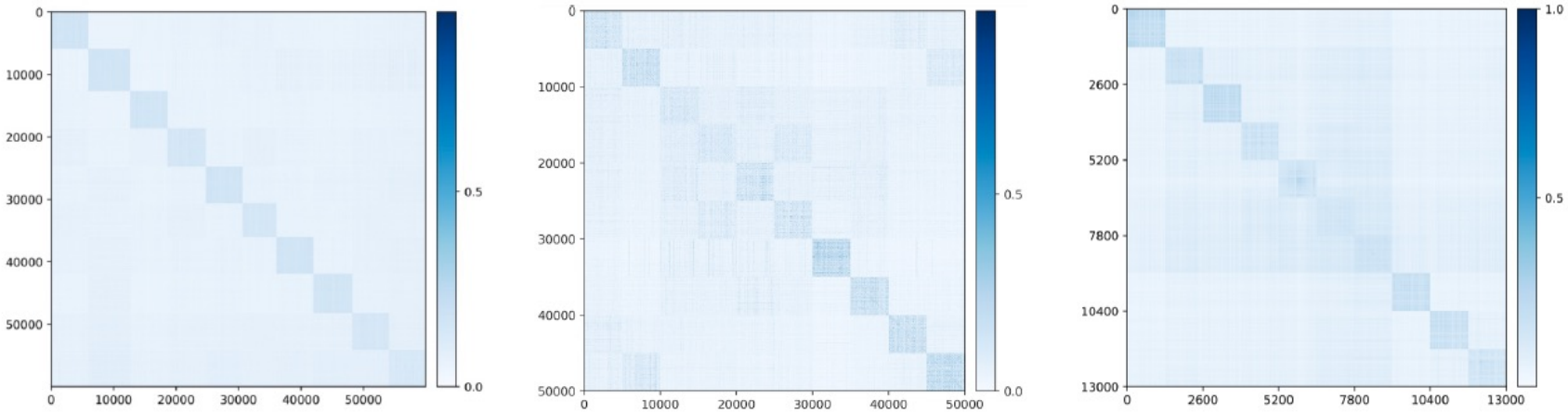
$$\mathcal{D}(\boldsymbol{X}, \hat{\boldsymbol{X}}) \doteq \max_{\theta}\min_{\eta} \sum_{j=1}^{k} \Delta R\big(\underbrace{f(\boldsymbol{X}_j,\theta)}_{\boldsymbol{Z}_j(\theta)}, \underbrace{f(g(f(\boldsymbol{X}_j,\theta),\eta),\theta)}_{\hat{\boldsymbol{Z}}_j(\theta,\eta)}\big).$$

Self-Consistency: Closed-Loop Feedback and Game

$$\max_{\theta}\min_{\eta}\Delta R\big(\boldsymbol{Z}(\theta)\big)+\Delta R\big(\hat{\boldsymbol{Z}}(\theta,\eta)\big)+\sum_{j=1}^{k}\Delta R\big(\boldsymbol{Z}_j(\theta),\hat{\boldsymbol{Z}}_j(\theta,\eta)\big).$$

- **Simplicity:** all terms are closed-form rate reduction on features.
- **Self-consistency:** enforced by closed-loop encoding and decoding.
- **Structured:** distribution of learned features $\boldsymbol{Z}$ is an LDR.
- **No** need to specify a prior or a surrogate target distribution.
- **No** need of any direct explicit distance between $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$.
- **No** more approximations or bounds for (KL-, JS-, W-) "distances".
- **No** heuristics or regularizing terms in the objective.

Experiments: Supervised Learning on Real-World Datasets

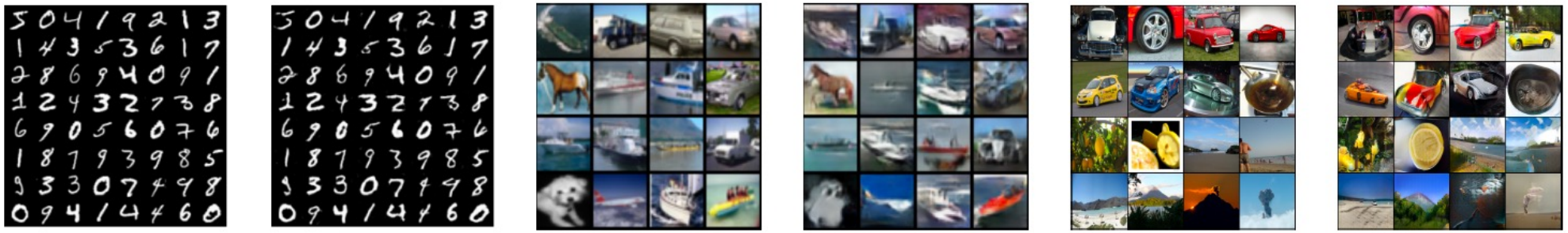


(a) MNIST (b) CIFAR10 (c) ImageNet

Figure: Visualizing the alignment between Z and $\hat{Z}$: $|Z^\top \hat{Z}|$.

Structured Representations
有结构的表示

No neural collapse,
No mode collapse!
无神经塌陷，
没有模式崩溃！

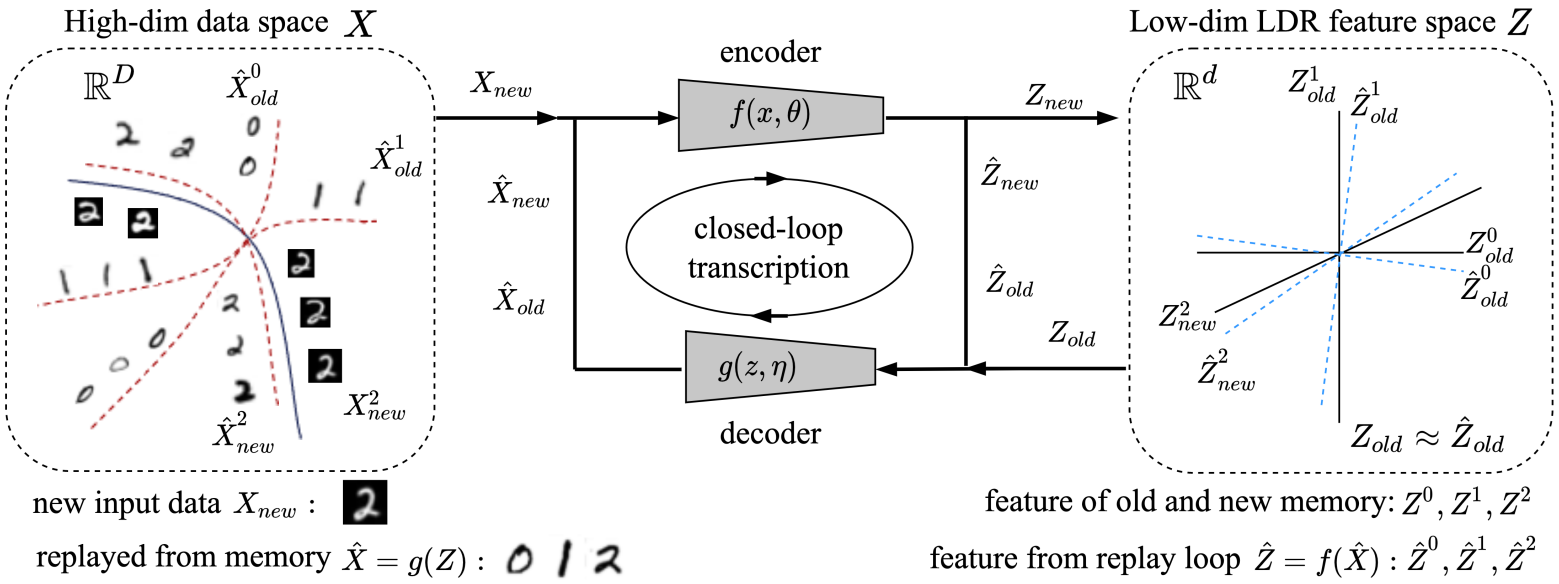


(a) MNIST x (b) MNIST $\hat{x}$ (c) CIFAR10 x (d) CIFAR10 $\hat{x}$ (e) ImageNet x (f) ImageNet $\hat{x}$

Figure: Visualizing the auto-encoding property: $x \approx \hat{x} = g \circ f(x)$.

Experiments: Incremental and Continuous Learning

Incremental Learning of Structured Memory: one class at a time.[9]



No catastrophic forgetting!
没有记忆丧失

$$\max_{\theta}\min_{\eta} \quad \Delta R(\boldsymbol{Z}) + \Delta R(\hat{\boldsymbol{Z}}) + \Delta R(\boldsymbol{Z}_{new}, \hat{\boldsymbol{Z}}_{new})$$

$$\text{subject to} \quad \Delta R(\boldsymbol{Z}_{old}, \hat{\boldsymbol{Z}}_{old}) = 0.$$

[9] Incremental Learning of Structured Memory via Closed-Loop Transcription, S. Tong and Yi Ma et. al., ICLR 2023. (arXiv:2202.05411)

Experiments: Incremental and Continuous Learning

Incremental Learning of Structured Memory: one class at a time.[10]

Method	MNIST	CIFAR10
INFORS (Sun et. al., ICLR 2022)	0.814	0.526
CLS-ER (Arani et. al. ICLR 2022)	0.895	0.662
i-LDR(ours)	**0.990**	**0.723**

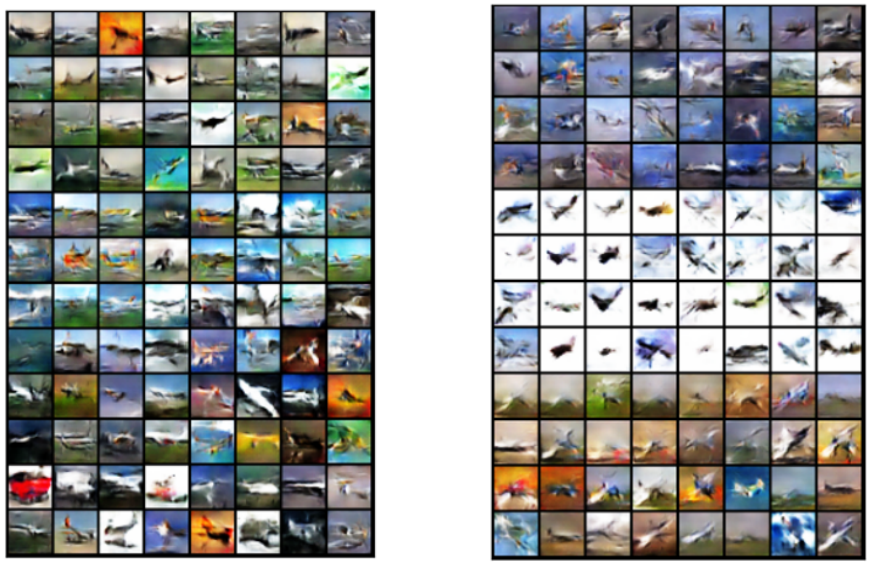
Table: Comparison with latest SOTA on MNIST and CIFAR-10.

iCaRL-S	EEIL-S	DGMw	EEC	EECS	i-LDR
0.290	0.118	0.178	0.352	0.309	**0.523**

Table: Comparison on ImageNet-50. The results of other methods are as in the EEC paper.

No catastrophic forgetting!

Memory consolidation via review (ICLR 2023)



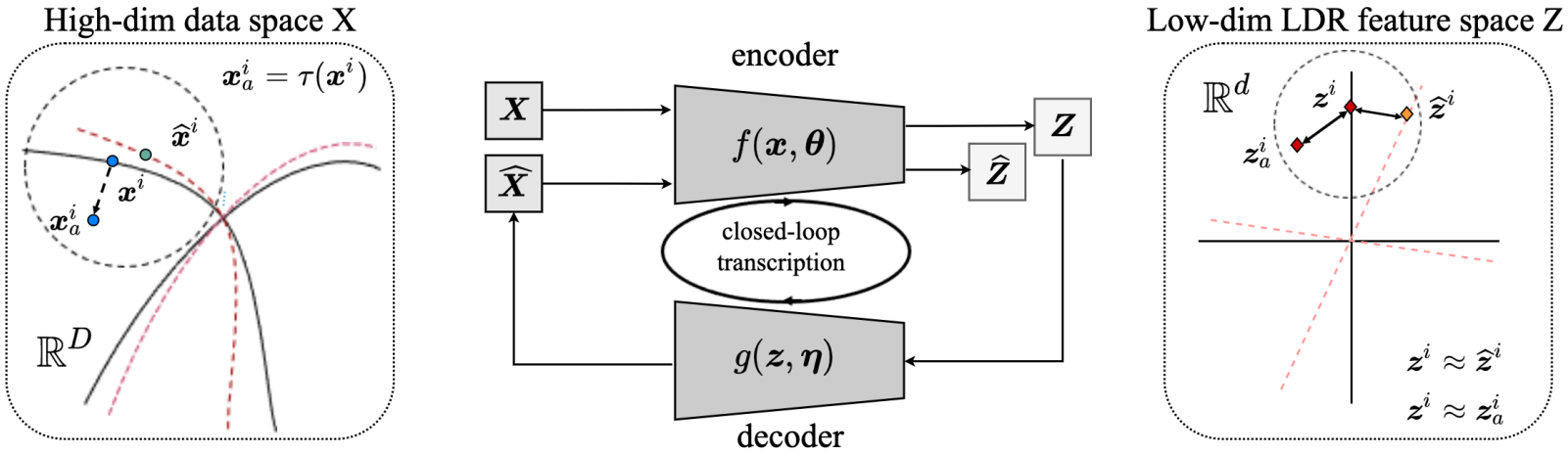
(a) $\hat{\boldsymbol{x}}_{old}$ before review (b) $\hat{\boldsymbol{x}}_{old}$ after review

Figure: Visualization of replayed images $\hat{\boldsymbol{x}}_{old}$ of class 1-'airplane' in CIFAR10.

而是复习使记忆增强

Experiments: Unsupervised or Self-Supervised Learning

Unsupervised Learning of Structured Memory: one sample at a time[11]



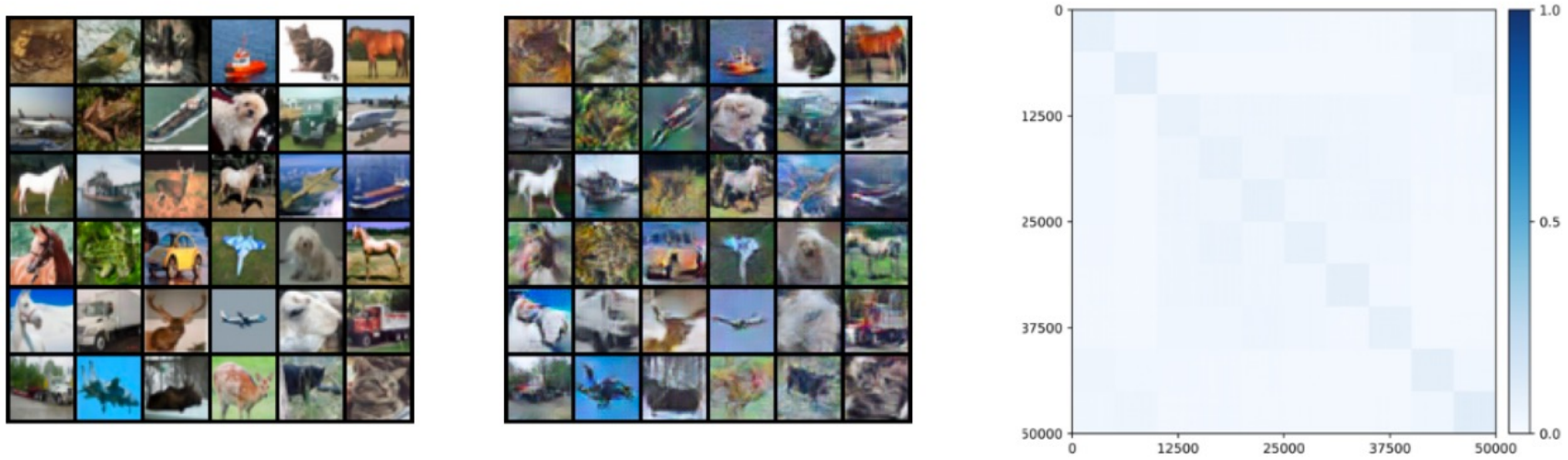
No catastrophic forgetting!
没有记忆丧失

$$\max_{\theta} \min_{\eta} \quad R(\boldsymbol{Z}) + \Delta R(\boldsymbol{Z}, \hat{\boldsymbol{Z}})$$

$$\text{subject to} \quad \sum_{i \in N} \Delta R(\boldsymbol{z}^i, \hat{\boldsymbol{z}}^i) = 0, \quad \text{and} \quad \sum_{i \in N} \Delta R(\boldsymbol{z}^i, \boldsymbol{z}_a^i) = 0.$$

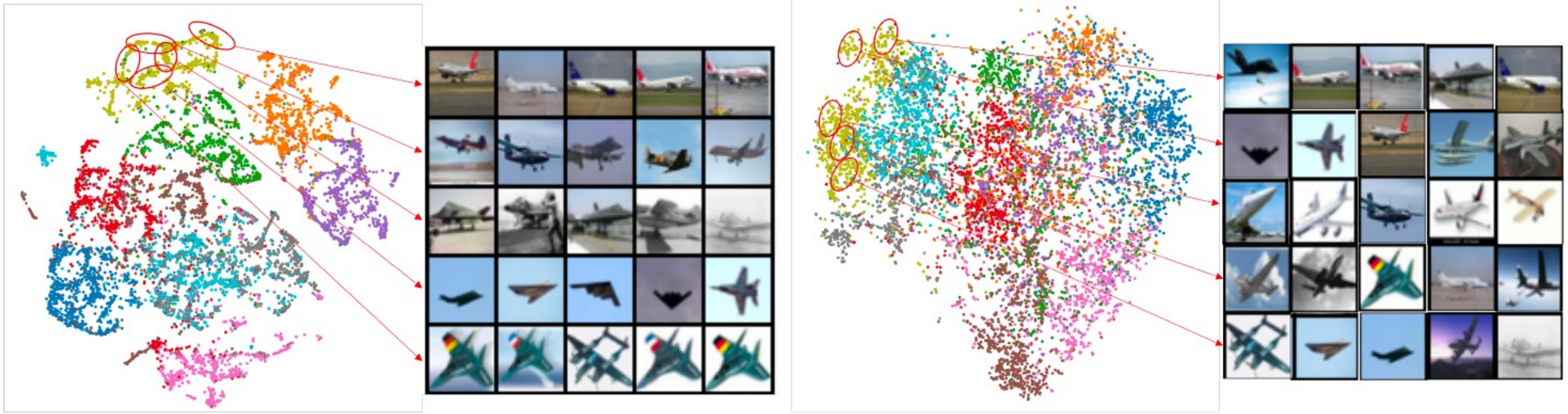
[11] Unsupervised Learning of Structured Representations via Closed-Loop Transcription, S. Tong, Yann LeCun, and Yi Ma, arXiv:2210.16782, 2022.

Experiments: Unsupervised or Self-Supervised Learning



Structured Representations
有结构的表示

Figure: Sample-wise self-consistency and block-diagonal structures.



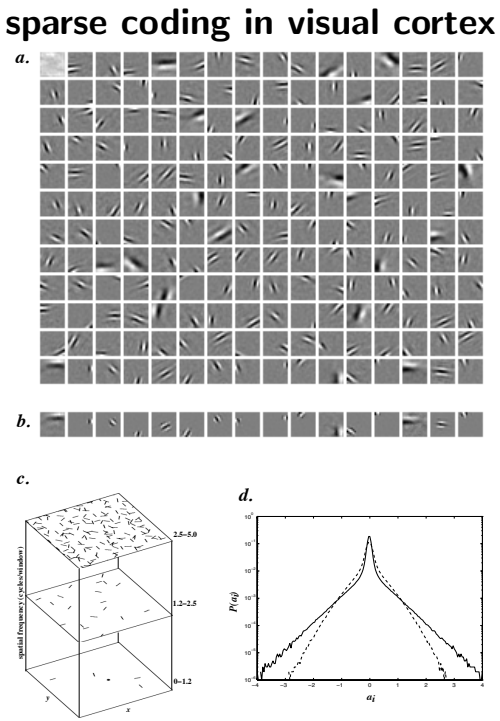
From 1000 epochs towards one epoch!

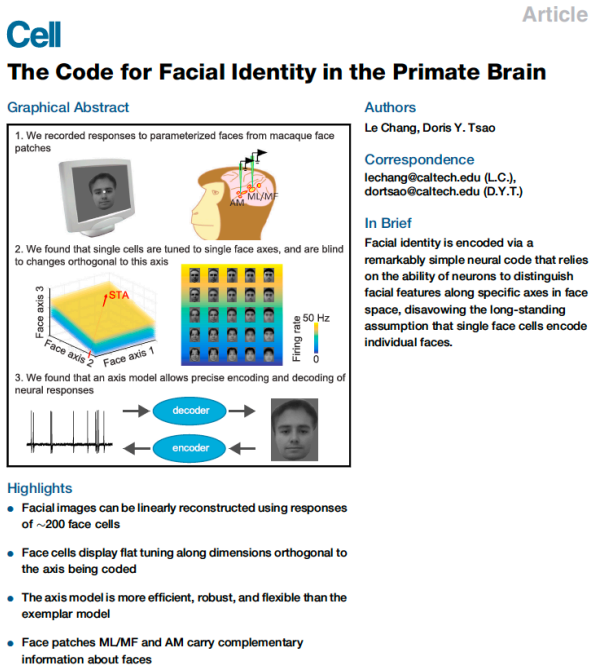
Figure: t-SNE of learned features . Left: U-CTRL and Right: MoCoV2.

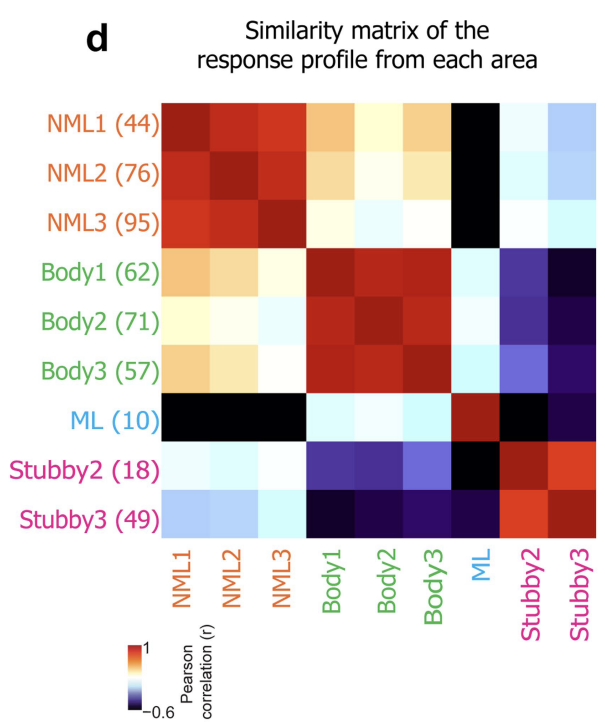
Connections to Neuroscience? 与神经科学的联系？

System demonstrated the same characteristics of (visual) memory in nature:

- Sparse coding in visual cortex (Olshausen, Nature 1996) [12].
- Subspace embedding (Tsao, Cell 2017, Nature 2020).[13]
- Predictive coding in visual cortex (Rao, Nature Neuroscience 1999).

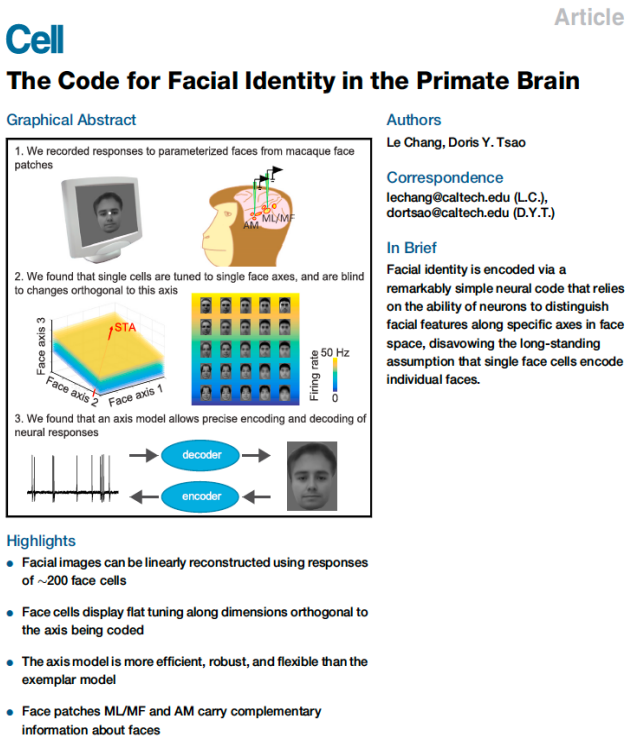






Connections to Neuroscience?与神经科学的联系？

- Parsimony: what's in neuroscience to verify this principle?
- 如何验证自然界遵从简约原理？
- Self-consistency: what's in neuroscience to verify this principle?
- 如何验证自然界遵从自洽原理？
- Forward optimization versus backward propagation?
- 前馈还是反馈？
- Open loop versus closed loop？
- 开环还是闭环？
- Self-critiquing or self-interrogating mechanisms？
- 自纠正与自省机制？
- From signals, to structures, to symbols, to semantics?
- 从信号到结构、到符号、到语义？



Article

Cell

The Code for Facial Identity in the Primate Brain

Graphical Abstract

1. We recorded responses to parameterized faces from macaque face patches

ML/MF

AM

2. We found that single cells are tuned to single face axes, and are blind to changes orthogonal to this axis

STA

Face axis 3

Face axis 2

Face axis 1

Firing rate

50 Hz

0

3. We found that an axis model allows precise encoding and decoding of neural responses

decoder

encoder

Authors

Le Chang, Doris Y. Tsao

Correspondence

lechang@caltech.edu (L.C.), dortsao@caltech.edu (D.Y.T.)

In Brief

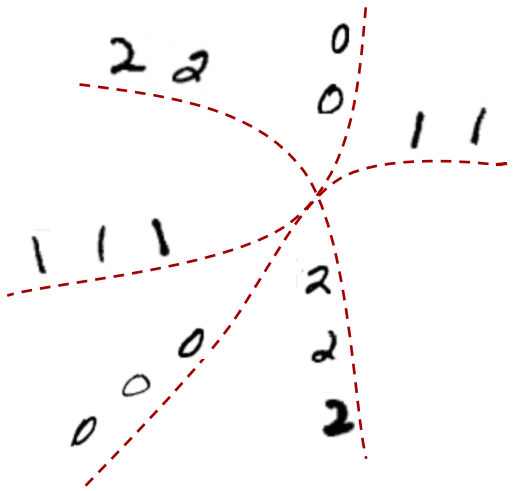
Facial identity is encoded via a remarkably simple neural code that relies on the ability of neurons to distinguish facial features along specific axes in face space, disavowing the long-standing assumption that single face cells encode individual faces.

Highlights

- Facial images can be linearly reconstructed using responses of ~200 face cells
- Face cells display flat tuning along dimensions orthogonal to the axis being coded
- The axis model is more efficient, robust, and flexible than the exemplar model
- Face patches ML/MF and AM carry complementary information about faces

A Universal Learning Engine: Compressive Closed-loop Transcription
通用学习引擎：压缩闭环转录

external multi-modal high-dim sensory data with low-dim structures



visual

tactile

auditory

olfactory

…

x

$\hat{x} = g(z)$

sensor/encoder f

$f(x, \theta)$

$z_{working} = f(x)$

$\hat{z} = f(\hat{x})$

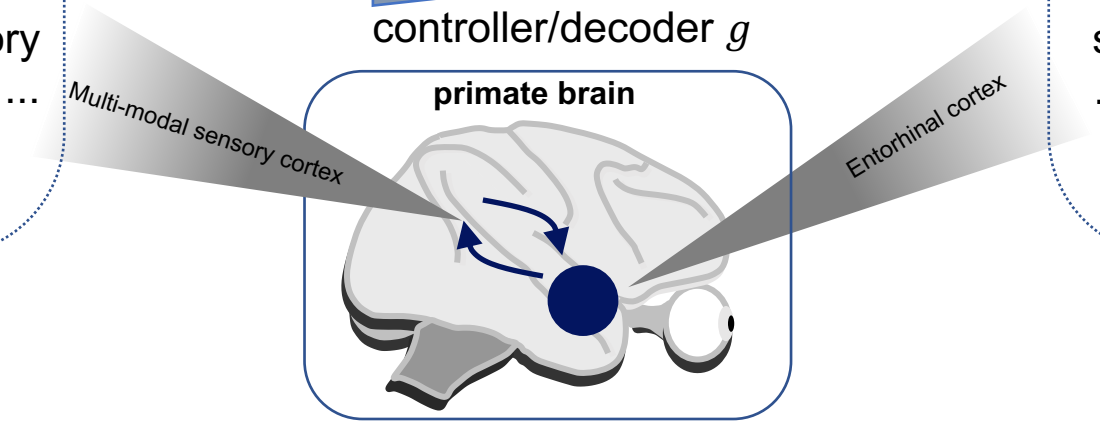
self-consistency
via closed-loop
$z \approx \hat{z} = f(g(z))$

脑回路？

z_{old}

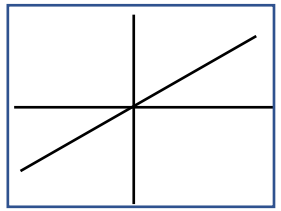
$g(z, \eta)$

controller/decoder g



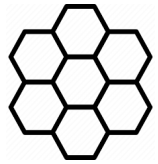
internal **compact** and **structured** representations for **parsimony**

sparse codes

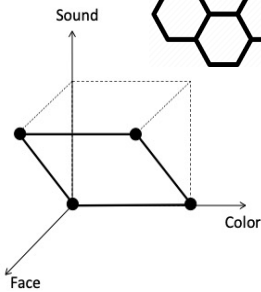


subspaces

grid/place cells

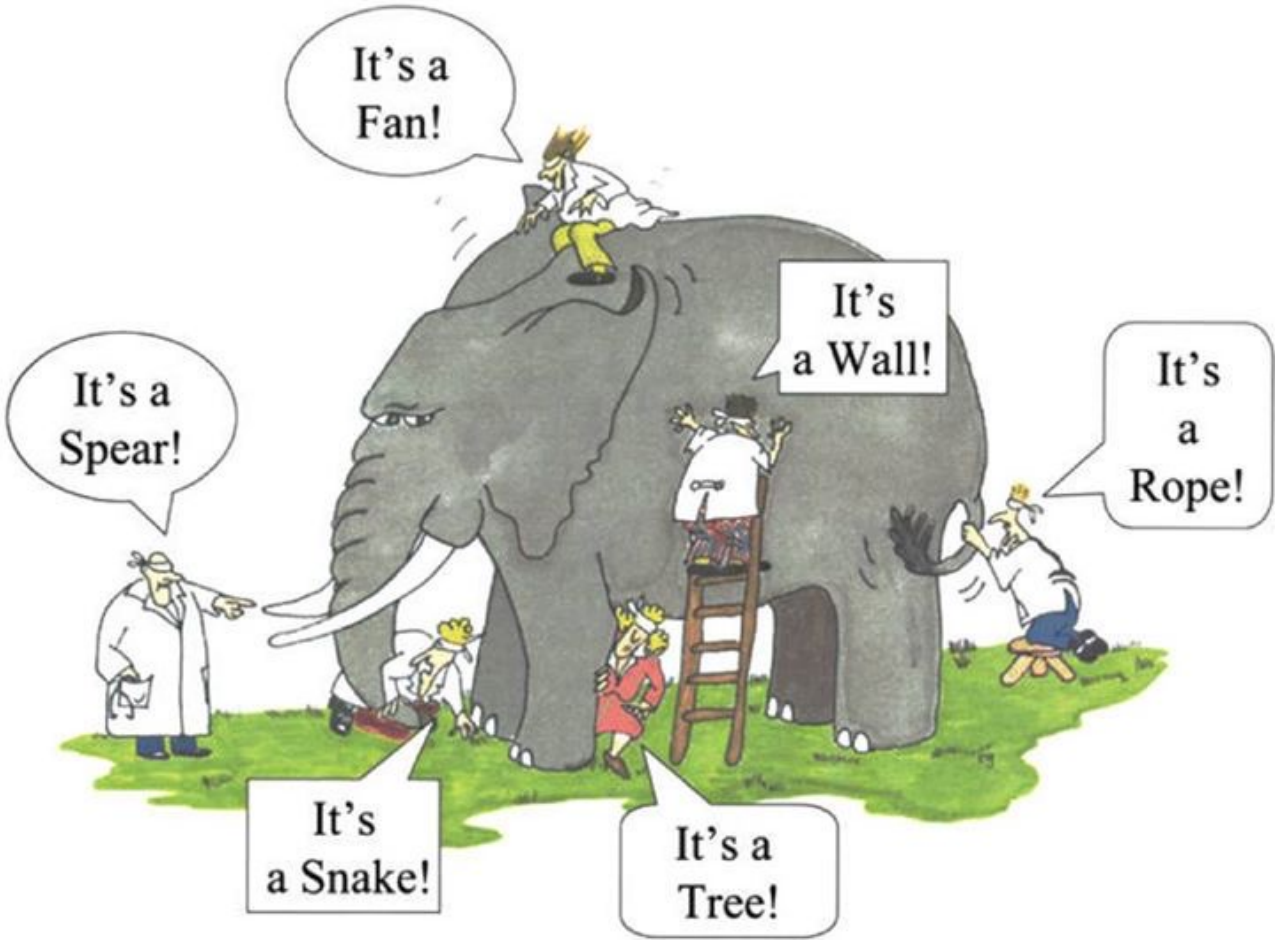


sparse graphs



…

A Universal Learning Engine: Compressive Closed-loop Transcription
通用学习引擎：压缩闭环转录



"Deep network is all you need."

"Reward is all you need."

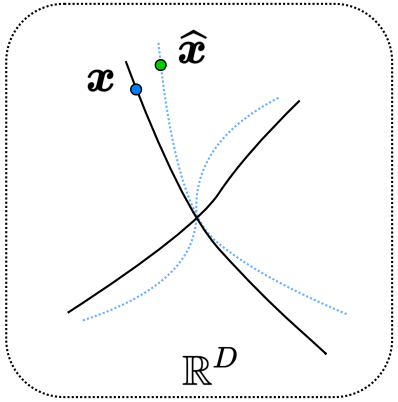
"Attention is all you need."

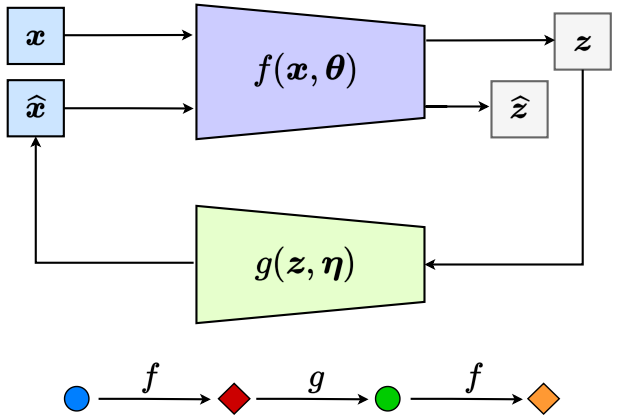
… …

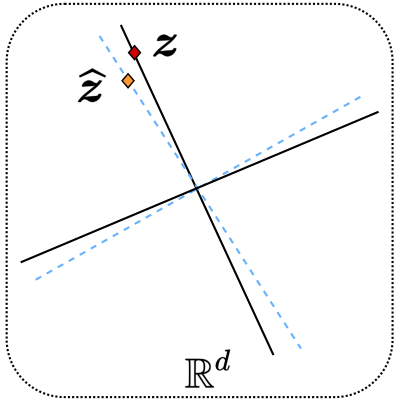
Intelligence is the characteristics of a highly integrated system!

A Universal Learning Engine: Compressive Closed-loop Transcription
通用学习引擎：压缩闭环转录

External world







Internal memory

- **a universal learning engine:** transform sensed data of external world to a compact and structured (LDR or sparse) internal representation.
- **parsimony:** optimization of the information gain (rate reduction) via a sensor and a generator.
- **self-consistency:** a self-critiquing game between the sensor and generator through a closed-loop feedback system.
- **a white-box system:** learning objectives, network architectures & operators, and learned representations.

The Past Magical Decade: the Emergence of Intelligence in GPT?
过去神奇的十年：OpenAI GPT智能的涌现？

March/2023

Video: https://youtu.be/ZZ0atq2yYJw[1]

Meeting: CONFERENCE JENSEN HUANG (NVIDIA) and ILYA SUTSKEVER (OPEN AI).AI TODAY AND VISION OF THE FUTURE (short article[2]*)*

Speaker: Dr Ilya Sutskever interviewed by NVIDIA's Jensen Huang

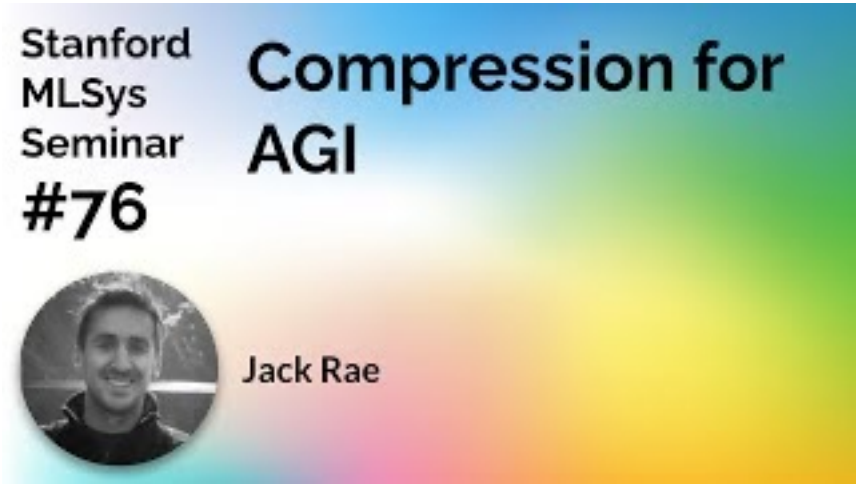
Transcribed by: OpenAI Whisper via SteveDigital's HF Space (https://huggingface.co/spaces/SteveDigital/free-fast-youtube-url-video-to-text-using-openai-whisper).

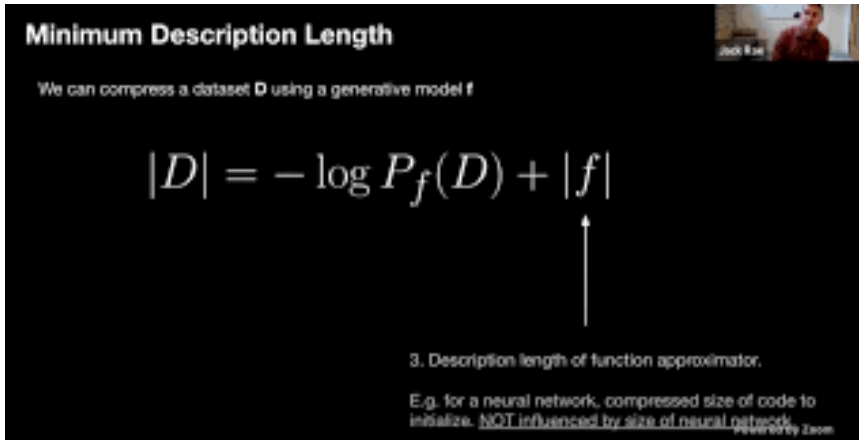
Edited by: Alan (without AI!)

Date: 15/Mar/2023 (day after GPT-4 release)

Highlights

– When we train a large neural network to accurately predict the next word in lots of different texts from the Internet, what we are doing is that we are learning a world model. It may look—on the surface—that we are just learning statistical correlations in text. But it turns out that to 'just learn' the statistical correlations in text, to compress them really well, what the neural network learns is some representation of the process that produced the text. This text is actually a projection of the world.





The Past Magical Decade: the Emergence of Intelligence in GPT?
过去神奇的十年：OpenAI GPT智能的涌现？

Consistency Models

Yang Song [1] Prafulla Dhariwal [1] Mark Chen [1] Ilya Sutskever [1]

Abstract

Diffusion models have made significant breakthroughs in image, audio, and video generation, but they depend on an iterative generation process that causes slow sampling speed and caps their potential for real-time applications. To overcome this limitation, we propose *consistency models*, a new family of generative models that achieve high sample quality without adversarial training. They support fast one-step generation by design, while still allowing for few-step sampling to trade compute for sample quality. They also support zero-shot data editing, like image inpainting, colorization, and super-resolution, without requiring explicit training on these tasks. Consistency models can be trained either as a way to distill pre-trained diffusion models, or as standalone generative models. Through extensive experiments,

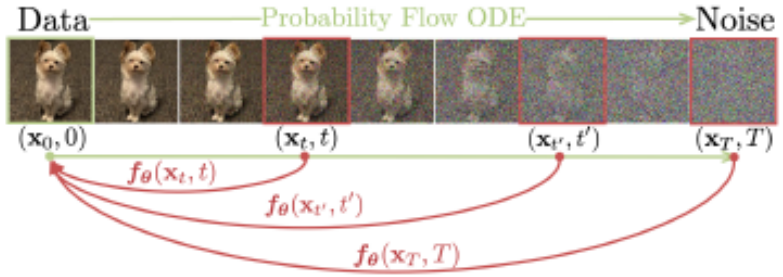


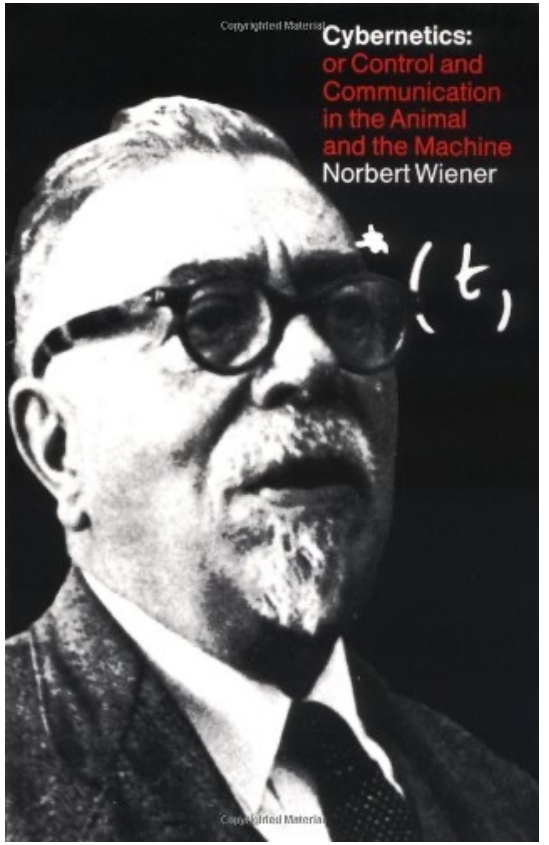
Figure 1: Given a Probability Flow (PF) ODE that smoothly converts data to noise, we learn to map any point (*e.g.*, $\mathbf{x}_t$, $\mathbf{x}_{t'}$, and $\mathbf{x}_T$) on the ODE trajectory to its origin (*e.g.*, $\mathbf{x}_0$) for generative modeling. Models of these mappings are called consistency models, as their outputs are trained to be consistent for points on the same trajectory.

Goodfellow et al. (2014)), these models do not rely on adversarial training and are thus less prone to issues such as unstable training and mode collapse. Additionally, diffusion

Recognition via Classification
or
Generation via Denoising are
not
Autoencoding like Memory!

A More Magical Decade: the True Origin of Intelligence Study
Intelligence真正神奇的十年：智能研究的真正起源

- 1943, **Artificial Neural Networks**, Warren McCulloch and Walter Pitts
  （1943年，**人工神经网络**，沃伦·麦卡洛克和沃尔特·皮茨）
- 1948, **Information Theory**, Claude Shannon
  （1948年，**信息论**，克劳德·香农）
- 1948, **Feedback Control & Cybernetics**, **Nobert Wiener**
  （1948年，**反馈控制与控制论**，**诺伯特·维纳**）
- 1944, **Game Theory**, John von Neumann
  （1944年，**博弈论**，约翰·冯·诺依曼）
- 1940's, **Turing Machine and Turing Test**, Alan Turing etc.
  （1940年代，**图灵机与图灵测试**，艾伦·图灵等。）



Perceive（感知）

Predict（预测）

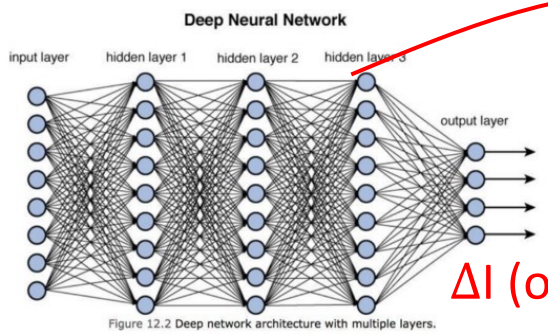
Learn/memory
（学习与记忆）

World Model: An Integrated System of Transcriptions?

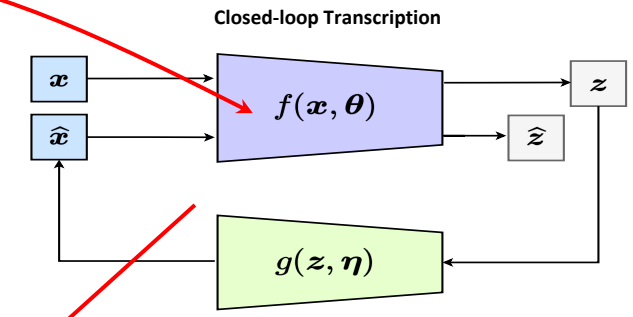
世界模型：大规模信息集成转录系统？

Neural networks are nature's optimization algorithms that maximize information gain. (one iteration per layer)

优化信息增量

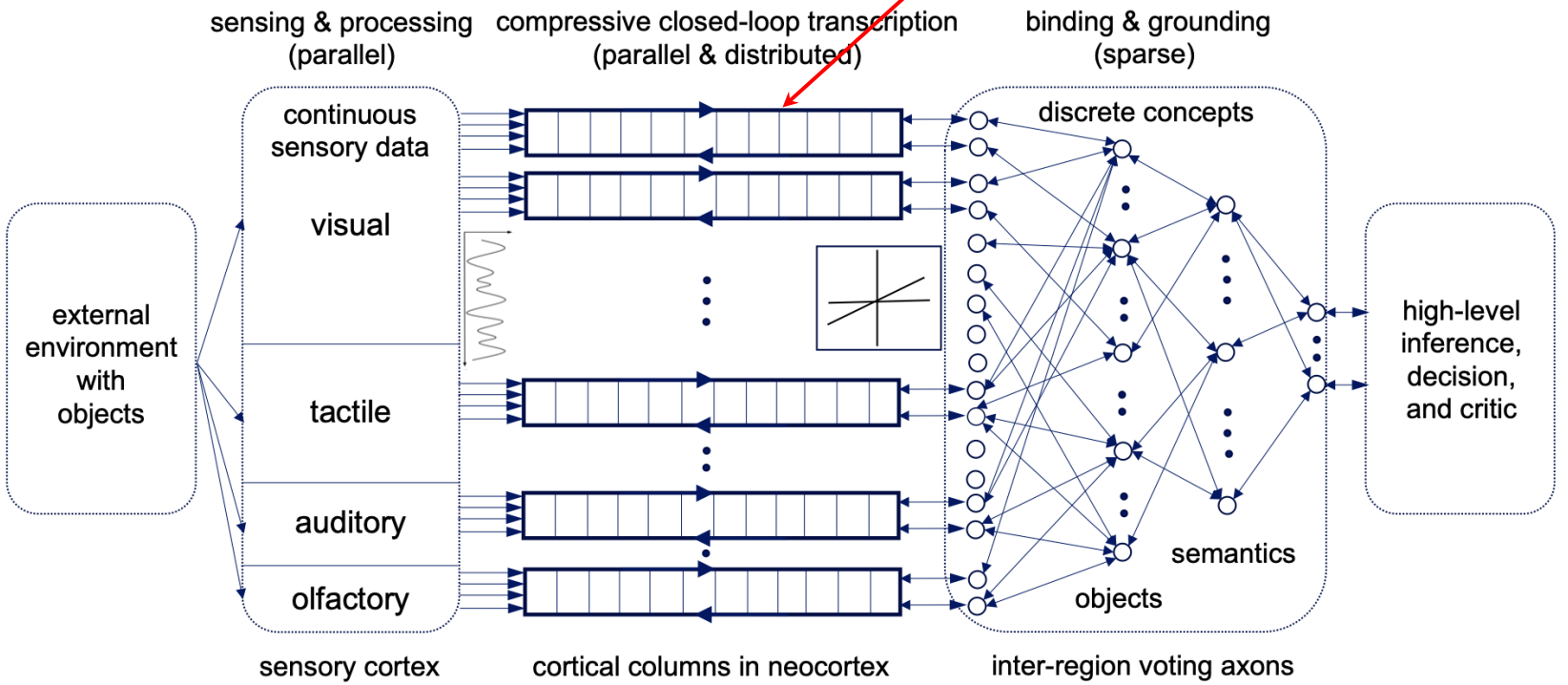


ΔI (or ΔR)

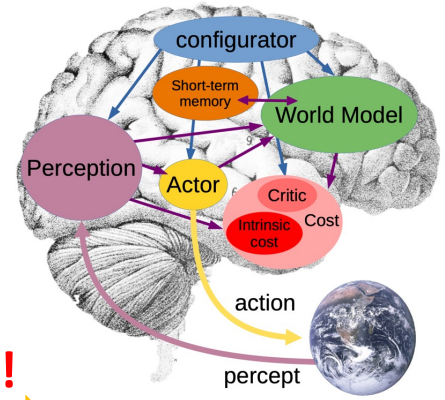


Closed-loop transcriptors are basic learning units for autonomous self-consistency. (error feedback & self-critique)

闭环自动学习



Robustly and efficiently learn compact structured representations of the world. (parallel & distributed)



A unified purpose: maximize "information gain" with every unit, at every stage!

References: Close the Loop between the Past and the Present

- On the Principles of Parsimony and Self-Consistency for the Emergence of Intelligence,
  Yi Ma, Doris Tsao and Heung-Yeung Shum, arXiv:2207.04630, FITEE, 2022.

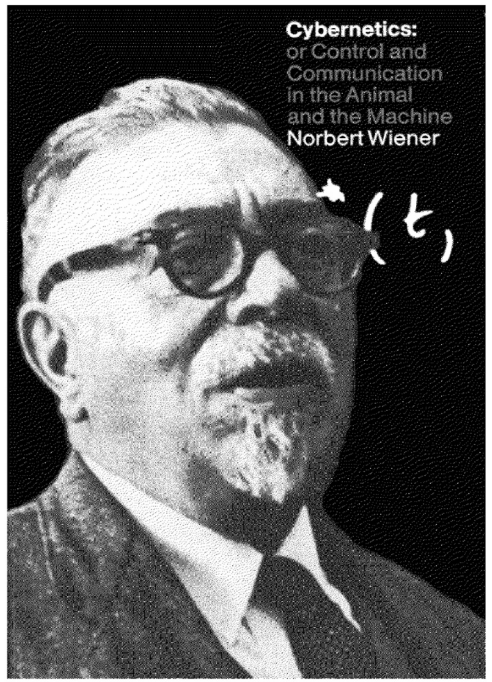
- CTRL: Closed-Loop Transcription to an LDR via Minimaxing Rate Reduction,
  Xili Dai, Shengbang Tong, Yi Ma et. al., arXiv:2111.06636, Entropy, March 2022.

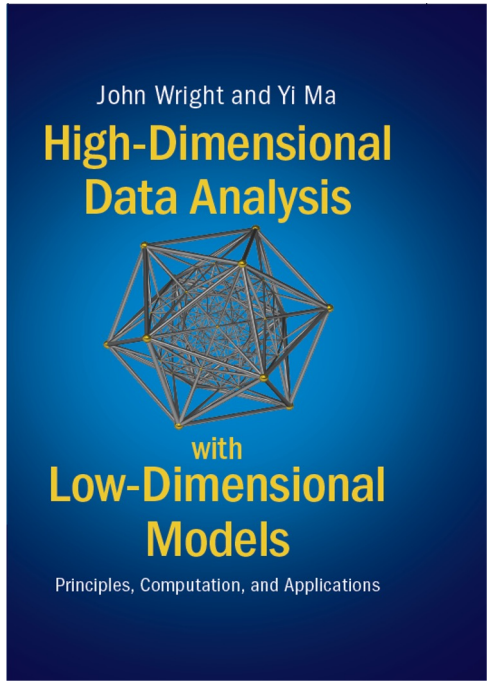
- ReduNet: A White-box Deep Network from the Principle of Maximizing Rate Reduction,
  Ryan Chan, Yaodong Yu, Yi Ma et. al., arXiv:2104.10446, JMLR, 2022.

"We compress to learn, we learn to compress!"

1948/1961

Compact coding
Closed-loop feedback
Learning via games
White-box modeling
Nonlinearity
Shift-invariance
Brain waves
...





2022

Sparse coding
Low-dimensional
Error correction
Optimization
Compression
Deep networks
...

Open the black box and close the loop for intelligence.

打开智能的黑盒子与实现闭环学习

Thanks!

"What I cannot create, I do not understand."

「对我自己无法创造出来的，我永远无法理解。」

-- Richard Feynman's last words

Office of Naval Research Science & Technology
NSF
SIMONS FOUNDATION
HKU Musketeers Foundation Institute of Data Science 香港大學同心基金數據科學研究院