统计学习理论与方法

博士生课程

Lecturer: Liqing Zhang

Dept. Computer Science & Engineering,

Shanghai Jiao Tong University

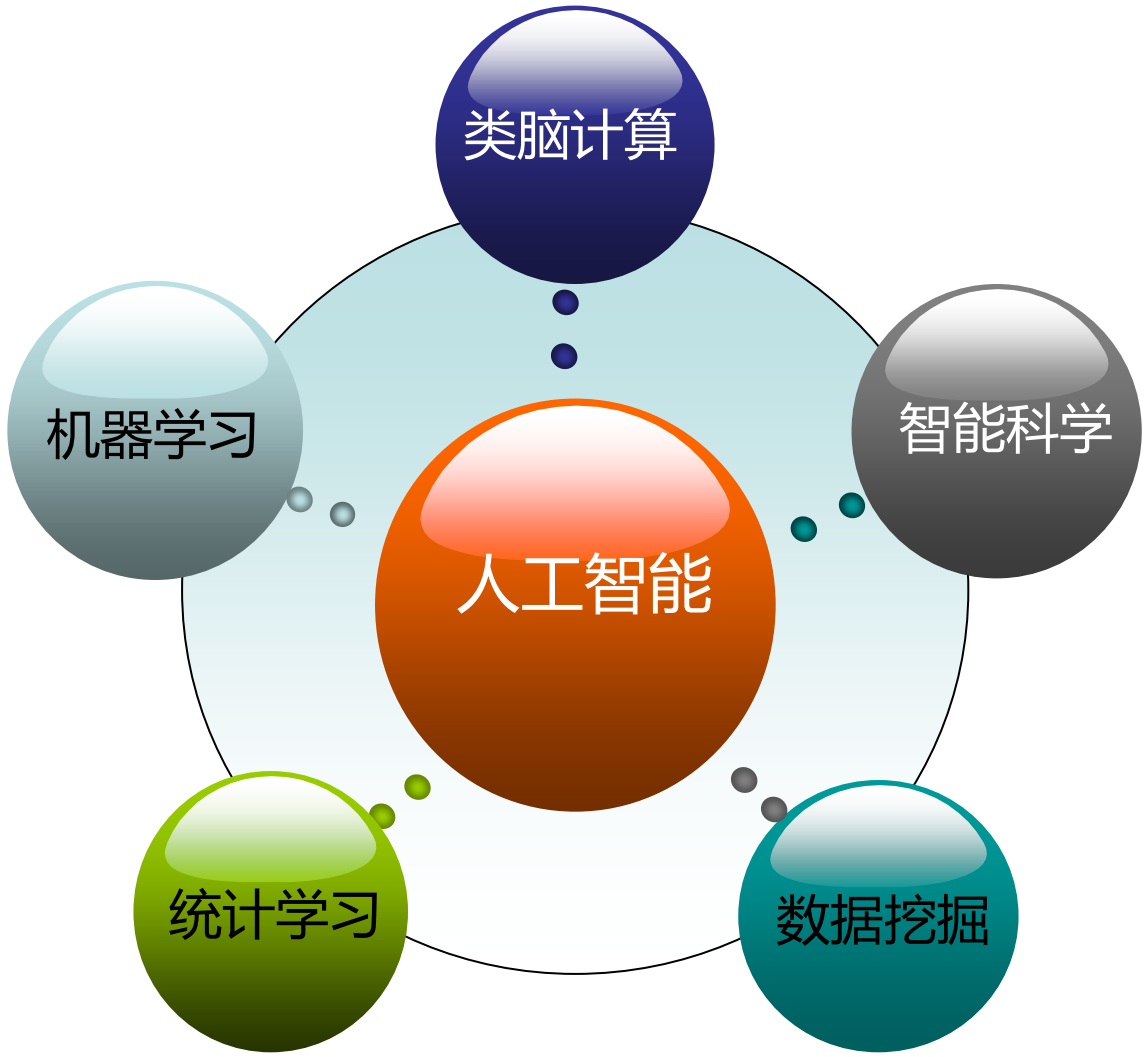
Textbooks and References

- Trevor Hastie, Robert Tibshirani, Jerome Friedman, **The Elements of Statistical Learning:** Data Mining, Inference, and Prediction, 2017, Springer-Verlag

- **Other References**
- Simon J.D. Prince, **Understanding Deep Learning, 2023， MIT Press**
- Roman Vershynin, **High-Dimensional Probability：** An Introduction with Applications in Data Science， **2018， Cambridge University Press**

人工智能与统计学习

- **Problem: Insuffient Number of Samples**
  - Uncertainty / Ergodicity
- **Model Complexity and Generalization**
  - Measure and Regularization
- **Model Learning and Model Selection**
  - Error bound Estimation
  - Convergence Rate
- **Modern Topics**
  - Common Knowledge
  - Concept and Attributes
  - Causal Inference
  - Data driven ➔ Data generation
  - … …

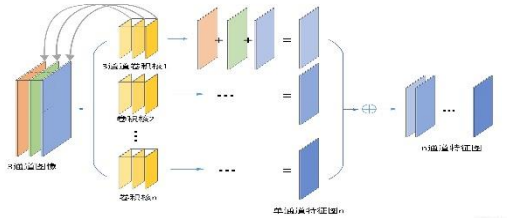


Contents of the Course

- Introduction
- Overview of Supervised Learning
- Linear Method for Regression and Classification
- Basis Expansions and Regularization
- Kernel Methods
- Model Selections and Inference
- Support Vector Machine
- Latent Variable Model and Variational Approximation
- Unsupervised Learning
- Deep Learning and Universal Approximation
- Generative Model and Diffusion Model

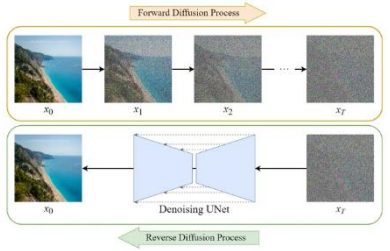
研究现状与发展趋势

- **未来趋势：** **Data-Driven Deep learning Approach** ⟹ **LLM based Generative AI**



- Face Recognition
- Object Recognition
- Medical Diagnosis
- ….

核心模型：CNN Core



- Image / Text Generation
- Chat Robots
- Software Development
- ….

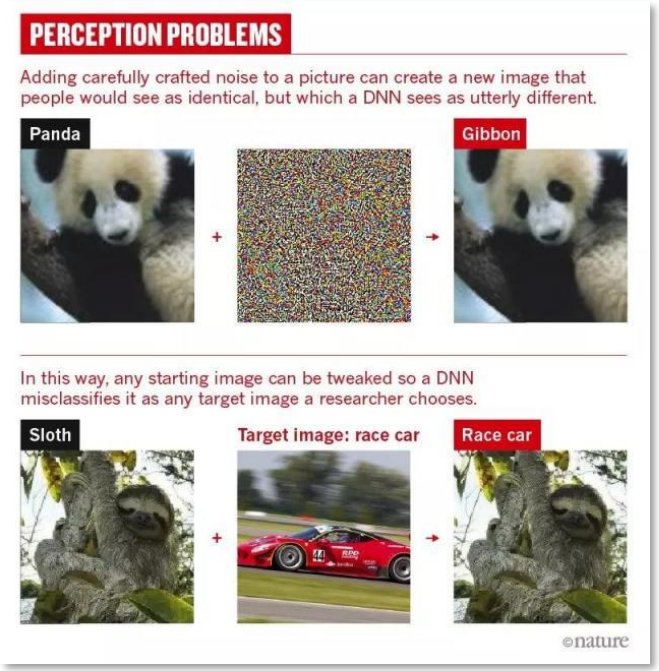
核心模型：Transformer / Diffusion Model

当前AI挑战与局限性： AI Systems lack of solid verification (unexplainability), Weak OOD generalization（long-tailed distribution）；AI generative System: Hard-controlability 、‘Hallucinations’

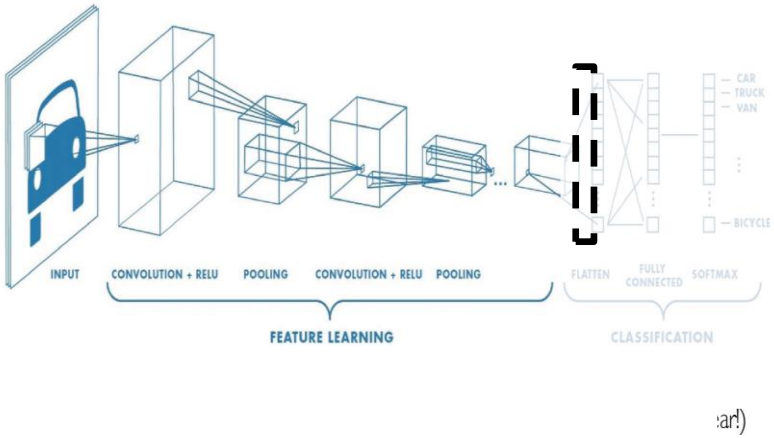
- **科学问题:**
  - Semantic Representation for Multi-modal Information
  - Controllable AI generative systems
  - How to embed common knowledge to help AI Systems predict rationally?

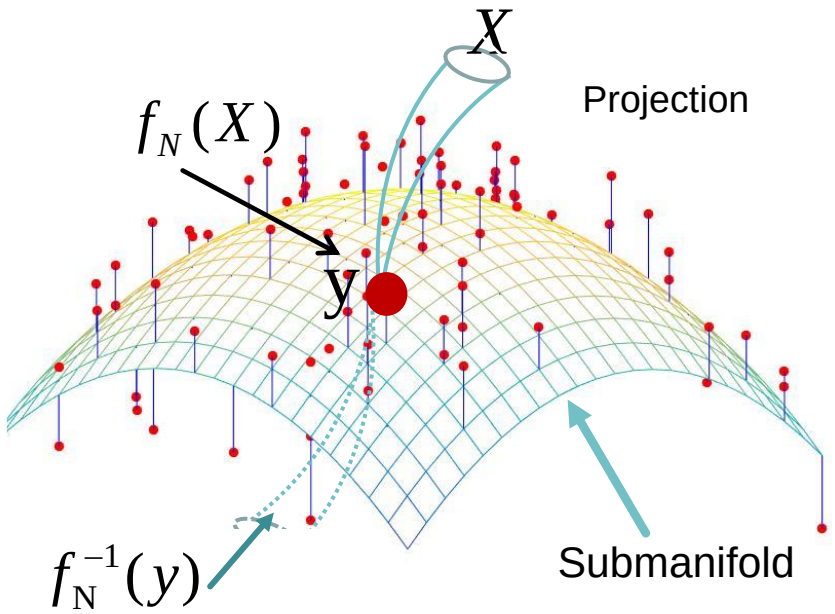
AI 挑战: 可验证性 / 可解释性

- **维数灾难问题:** CNN mapping images into embedding vectors, without expressing any semantic structures



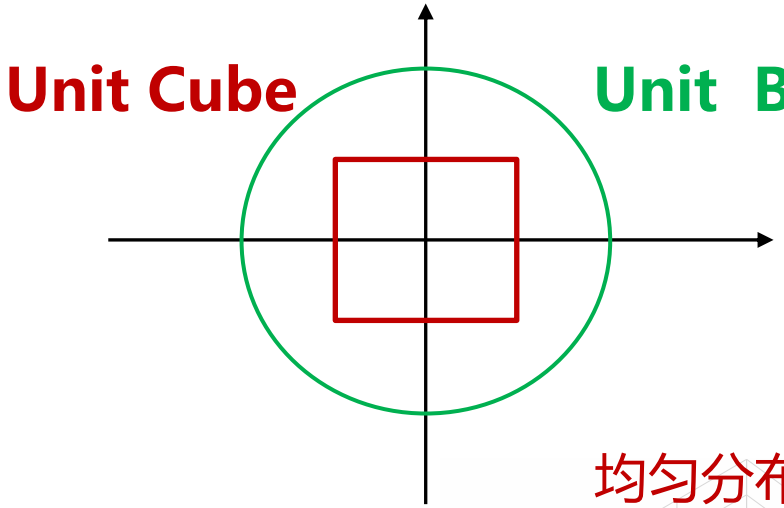
$X \in R^{M}$ $f_N(X)$ $y \in R^{N}$

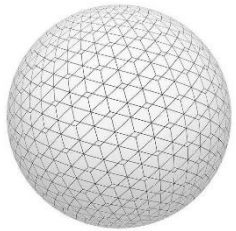




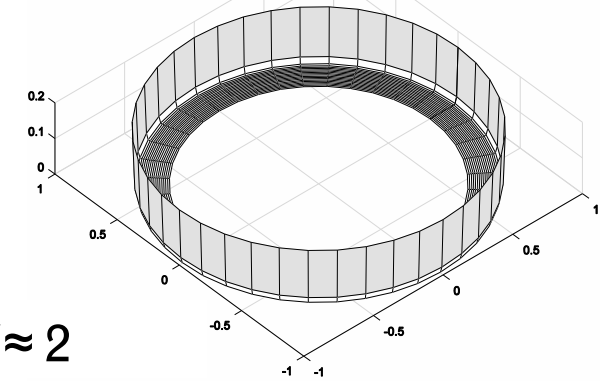
高维空间数据分布特性

- 问题 1：高位空间中单位球和单位立方体哪个体积大？



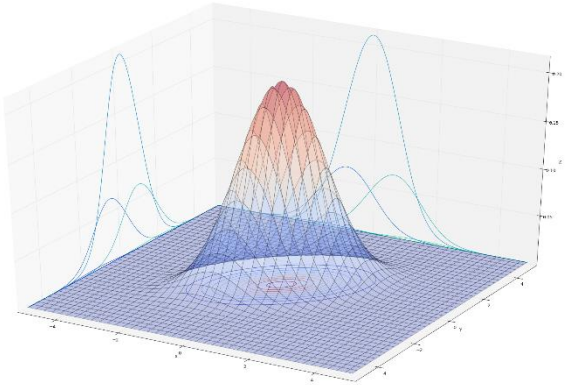


均匀分布

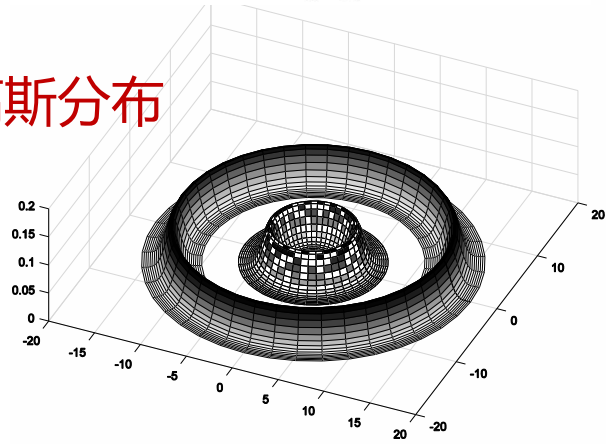


单位球内任意两点距离≈2

- 问题 2：高维空间中高斯分布的数据主要集中在均值附近吗？



高斯分布



内圈 n=20；外圈 n=200，半径 $r=\sqrt{n-1}$

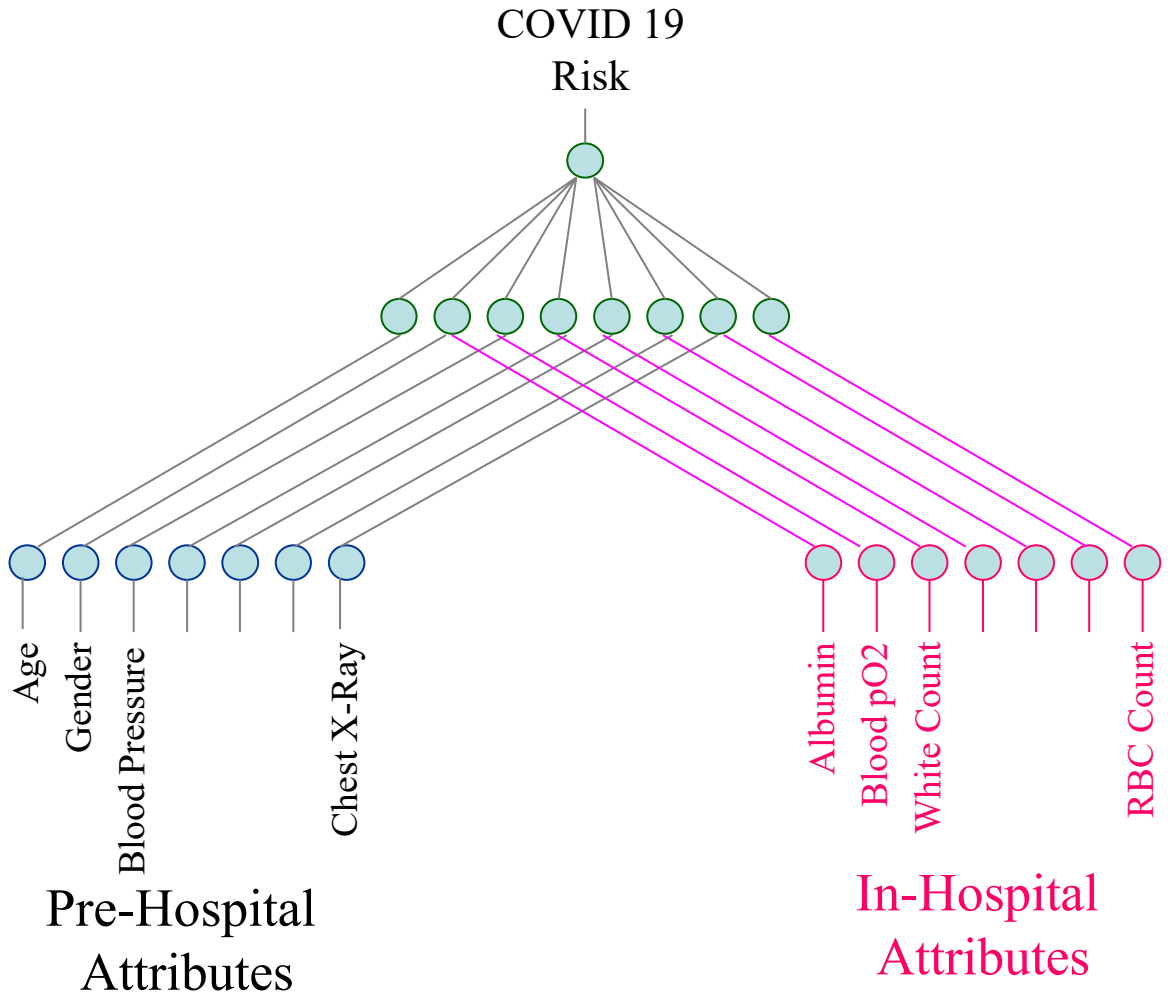
挑战问题: GPT models

- **The Vision Transformer**
  - Split an image into patches (fixed sizes)
  - Each patch is mapped to a lower dimensional vector via a learned linear transformation
  - Feed the sequence as an input to a SOTA transformer encoder

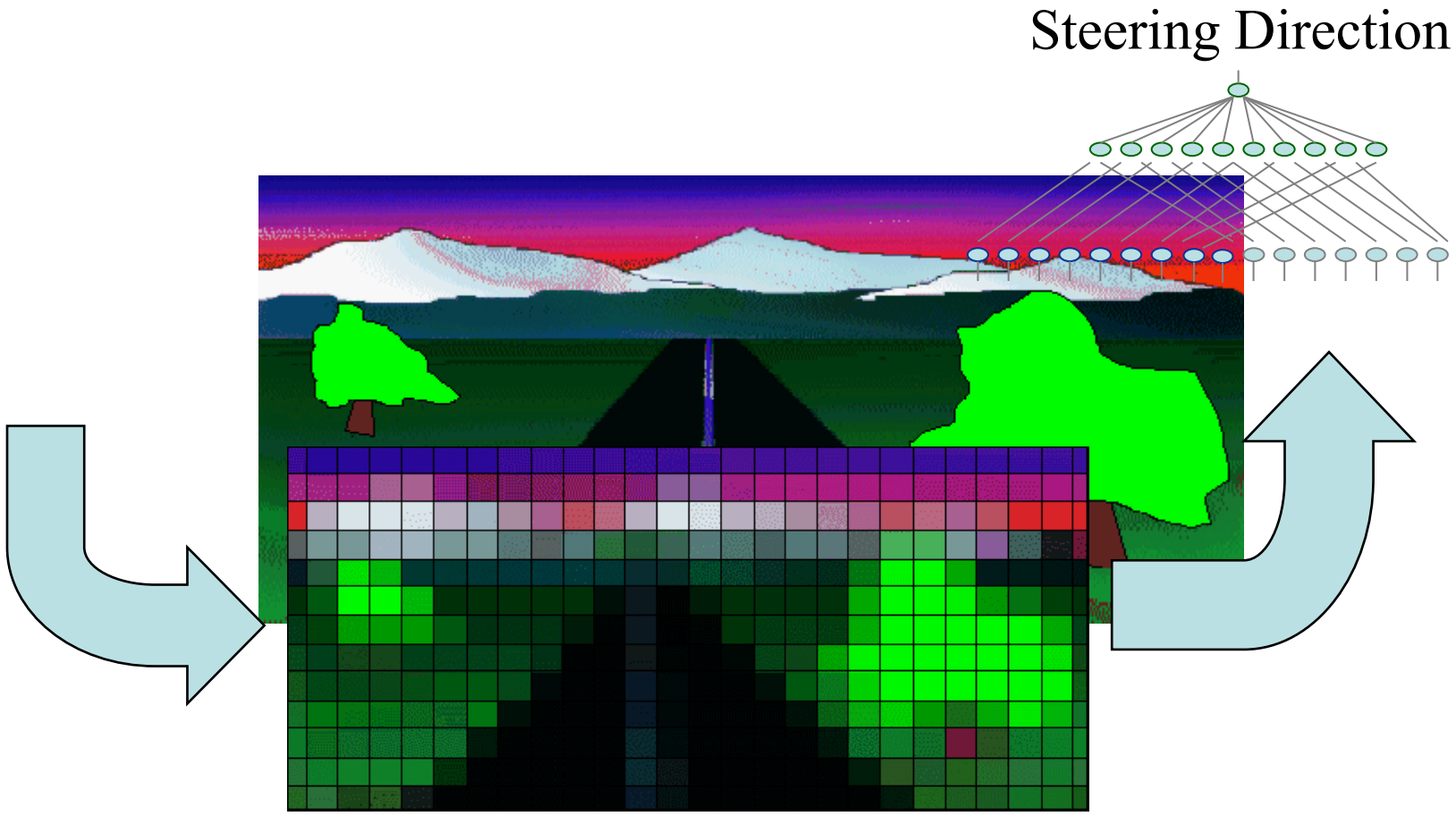
Applications in Statistical Learning

- Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack.
- Identify the risk factors for prostate cancer, based on clinical and demographic variables.
- Predict the price of a stock in 6 months from now, on the basis of company performance measures and economic data.
- Object recognition such as Human Face / ZIP code / plate numbers, from a digitized image.
- Latest applications of AI, such as Automatic driving, COVID19 diagnosis and treatment
- Image generation: Product Advertising / Virtual Clothes Try-on

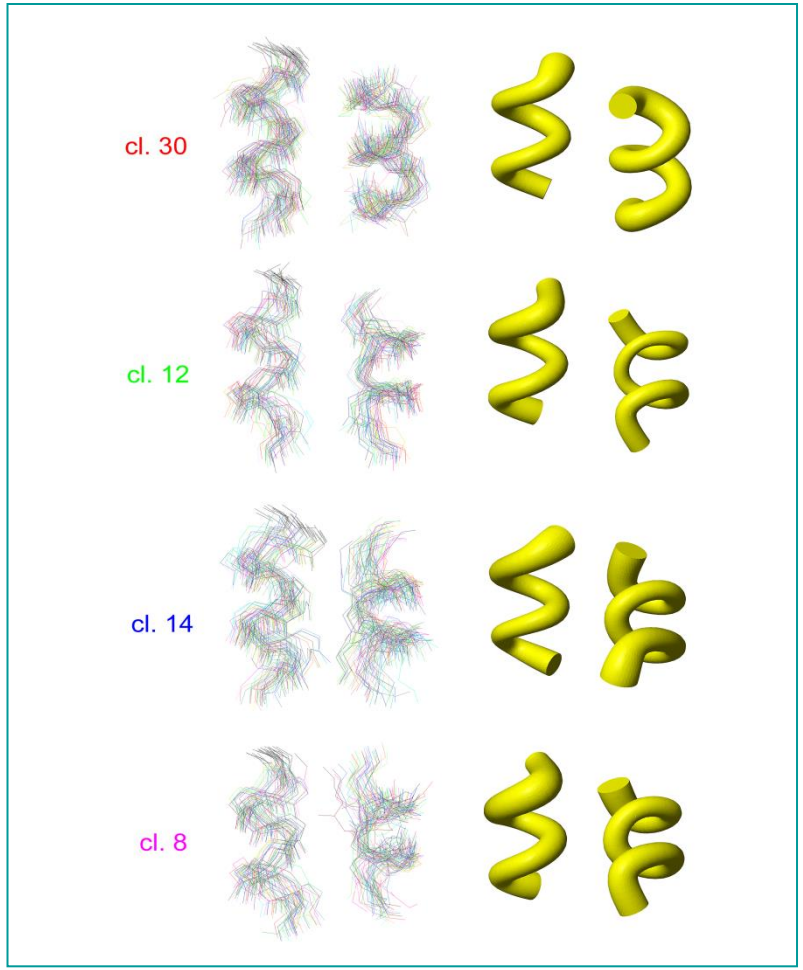
ML: COVID19 Risk Prediction

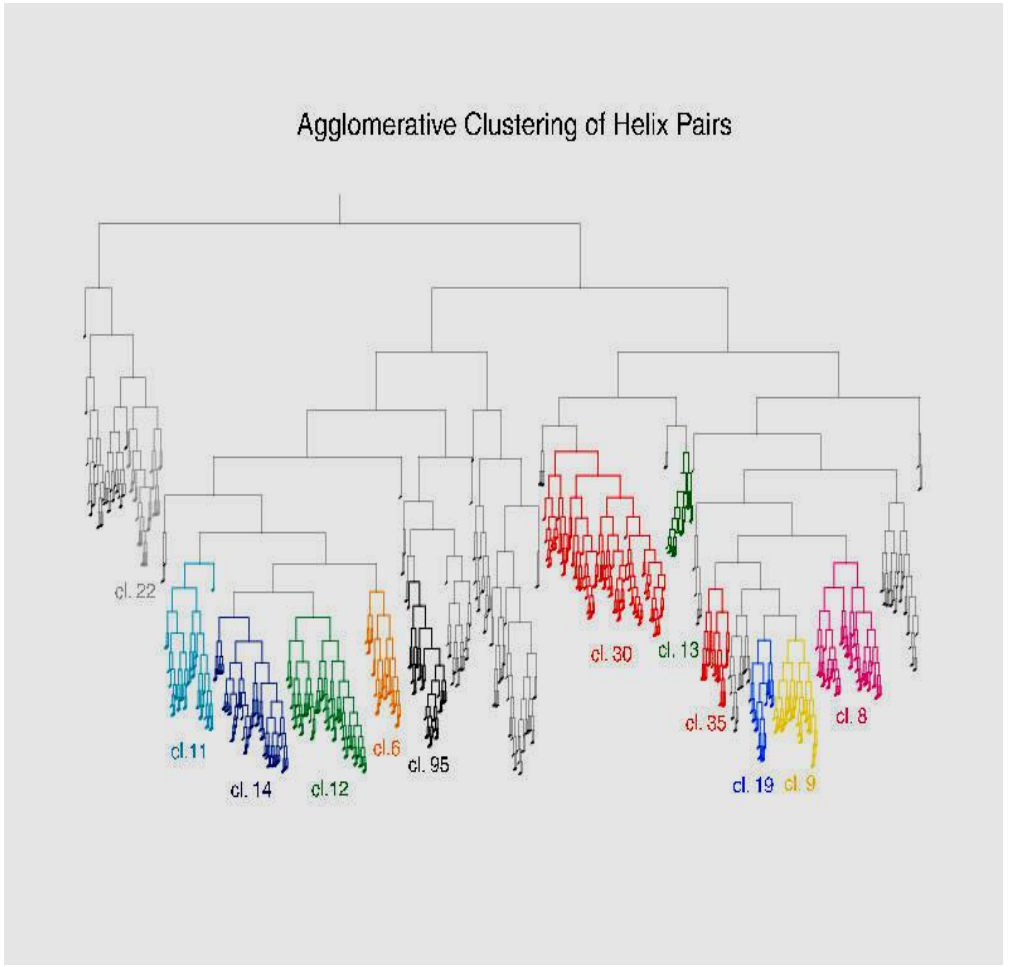


ML: Auto Vehicle Navigation



Protein Folding



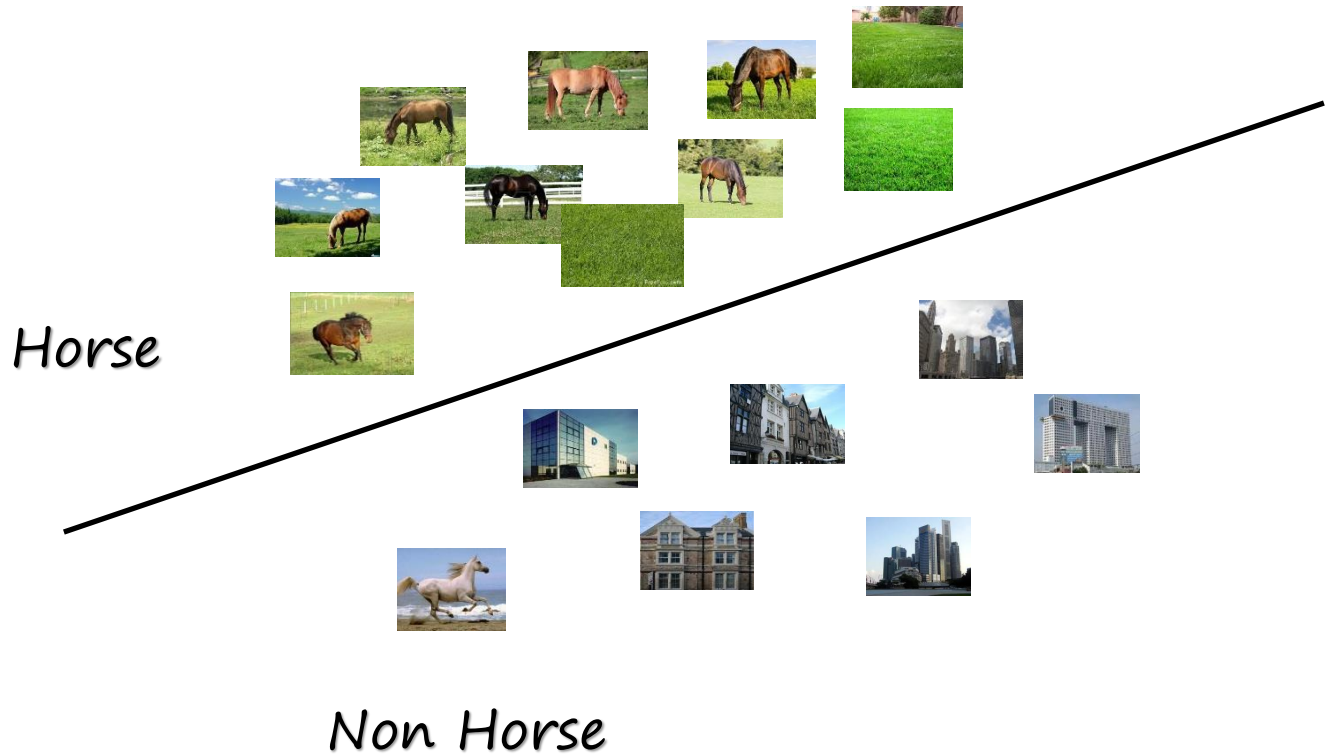


EX. Pattern Classification

- **Objective: To recognize horses in images**

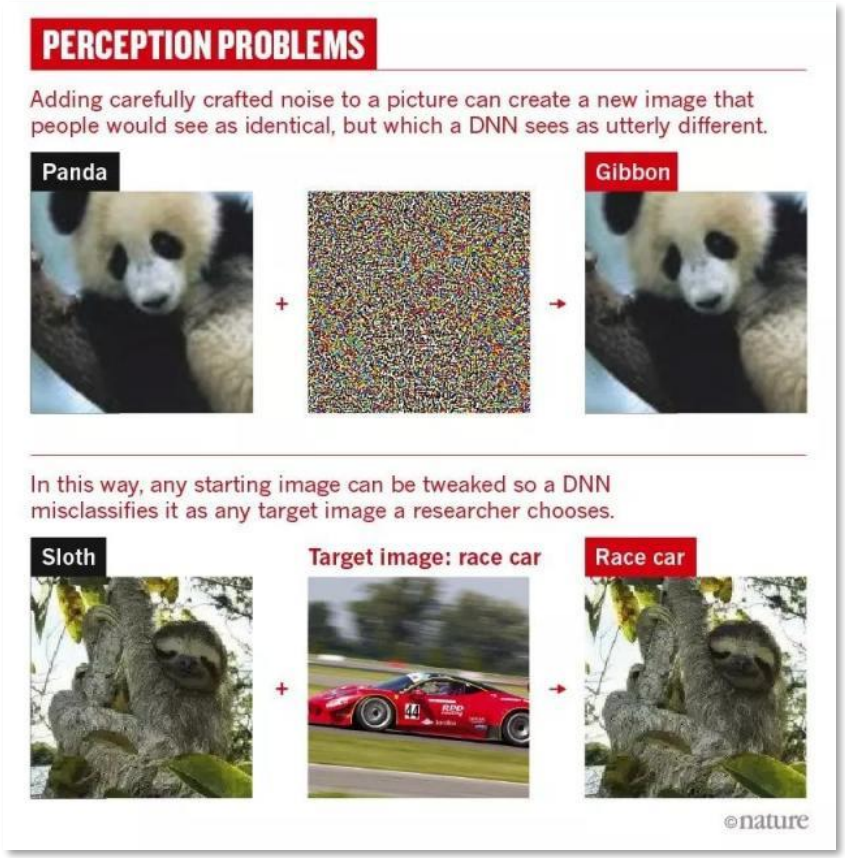
- **Procedure: Feature ➔ Classifier Training ➔ Cross Validation**

Failure Case: Wrong features

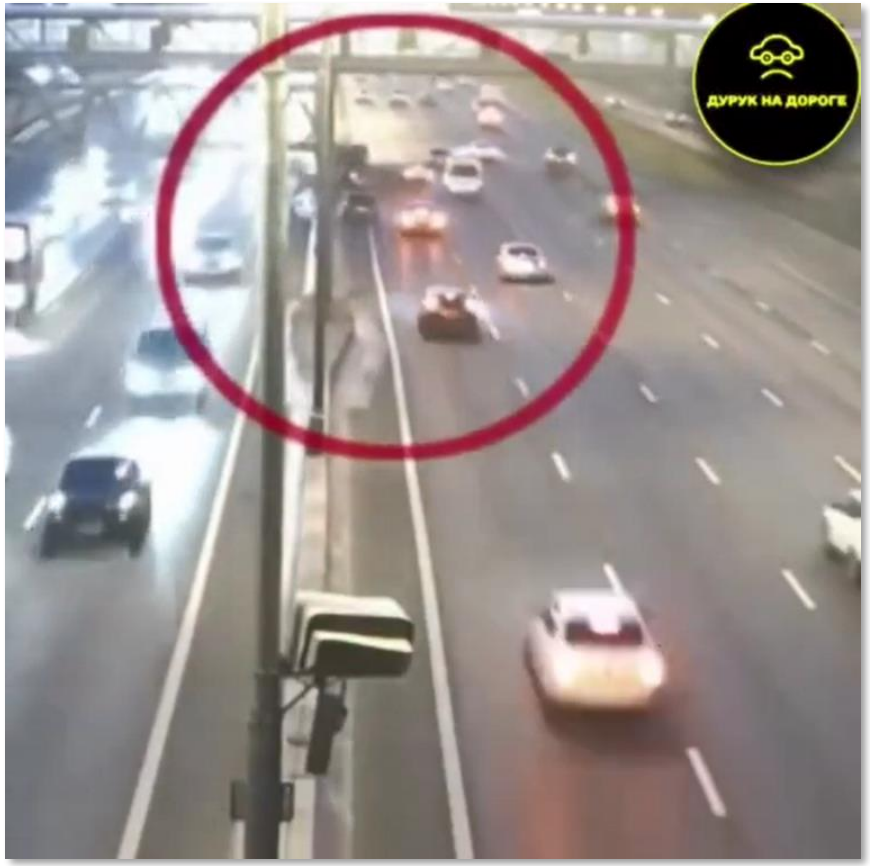


为什么深度学习模型泛化能力差?

为什么深度学习模型容易被欺骗?



为什么自动驾驶老是出事?



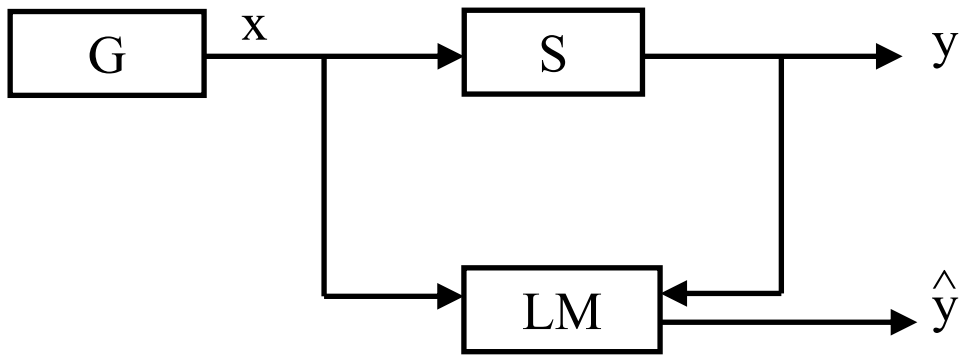
Heaven D. Why deep-learning AIs are so easy to fool[J]. Nature, 2019

Automatic Driving

- Baidu Automatic Driving Demo（2017）
- Tesla Car Accident（2019）

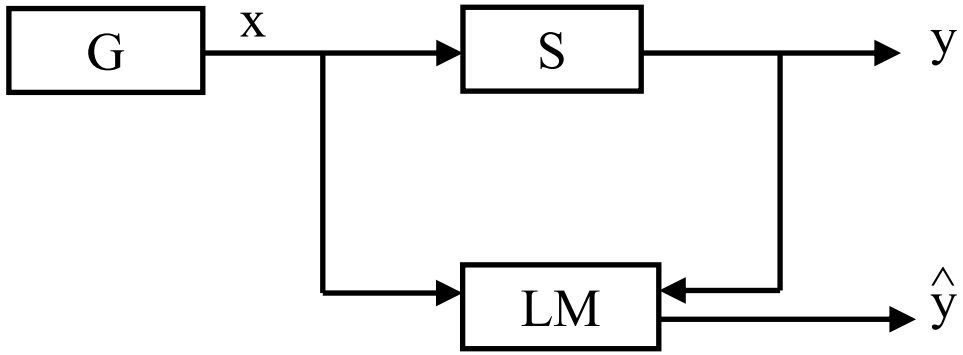
Function Estimation Model

- **Generator (G)** generates observations x (typically in R^n), independently drawn from some fixed distribution $F(x)$
- **Supervisor (S)** labels each input x with an output value y according to some fixed distribution $F(y|x)$
- **Learning Machine (LM)** "learns" from an i.i.d. sample of (x,y)-pairs output from G and S, by choosing a function that best approximates S from a parameterised function class $f(x,\alpha)$, where α is in Λ the parameter set



Function Estimation Model

- **Key concepts:** $F(x,y)$, an i.i.d. k-sample on F, functions $f(x,\alpha)$ and the equivalent representation of each f using its index α



The Problem of Risk Minimization

- **The loss functional (*L*, *Q*)**
  - the error of a given function on a given example

$$L:(x,y,f_\alpha) \mapsto L(y,f(x,\alpha))$$
$$Q:(z,\alpha) \mapsto L(z_y,f(z_x,\alpha))$$

- **The risk functional (*R*)**
  - the expected loss of a given function on an example drawn from $F(x,y)$
  - the (usual concept of) generalisation error of a given function

$$R(\alpha)=\int Q(z,\alpha)dF(z)$$

The Problem of Risk Minimization

- **Three Main Learning Problems**
  - **Pattern Recognition:**

$$y \in \{0,1\} \text{ and } L(y, f(x,\alpha)) = \mathbf{1}[y = f(x,\alpha)]$$

  - **Regression Estimation:**

$$y \in \Re \text{ and } L(y, f(x,\alpha)) = (y - f(x,\alpha))^2$$

  - **Density Estimation:**

$$y \in [0,1] \text{ and } L(p(x,\alpha)) = -\log p(x,\alpha)$$

General Formulation

- **The Goal of Learning**
  - Given an i.i.d. k-sample $z_1,\ldots, z_k$ drawn from a fixed distribution $F(z)$
  - For a function class' loss functionals $Q(z, \alpha)$, with α in Λ
  - We wish to minimise the risk, finding a value α^* such that
  - where

$$\alpha^* = \arg\min_{\alpha \in \Lambda} R(\alpha)$$

$$R(\alpha) = \int Q(z, \alpha) dF(z)$$

General Formulation

- **The Empirical Risk Minimization (ERM)**
  - **Define the empirical risk (sample/training error):**

$$R_{\text{emp}}(\alpha) = \frac{1}{k}\sum_{i=1}^{k} Q(z_i, \alpha)$$

  - **Define the empirical risk minimiser:**

$$\alpha_k = \arg\min_{\alpha \in \Lambda} R_{\text{emp}}(\alpha)$$

  - ERM approximates $Q(z, \alpha^*)$ with $Q(z, \alpha_k)$, the R_{emp} minimiser…that is ERM approximates α^* with α_k
  - Least-squares and Maximum-likelihood are realisations of ERM

Four Issues of Learning Theory

1. **Theory of consistency** of learning processes
   - What are (necessary and sufficient) conditions for consistency (convergence of R_{emp} to R) of a learning process based on the ERM Principle?
2. **The rate of convergence** of learning processes
   - How fast is the rate of convergence of a learning process?
3. **Generalization ability** of learning processes
   - How can one control the rate of convergence (the generalization ability) of a learning process?
4. **Constructing learning algorithms** (i.e. the SVM)
   - How can one construct algorithms that can control the generalization ability?

Types of learning

- **Supervised learning**
  - Given a category label for each pattern in a training set.
  - Such as Face Recognition, Text Classification, ……
- **Unsupervised learning**
  - Data Clustering, Data Quantization, Dimensional Reduction, ……
- **Reinforcement learning**
  - Multi-Agents, Robots, Automatic Driving ……
- **Semi- / Weakly /self- supervised learning**

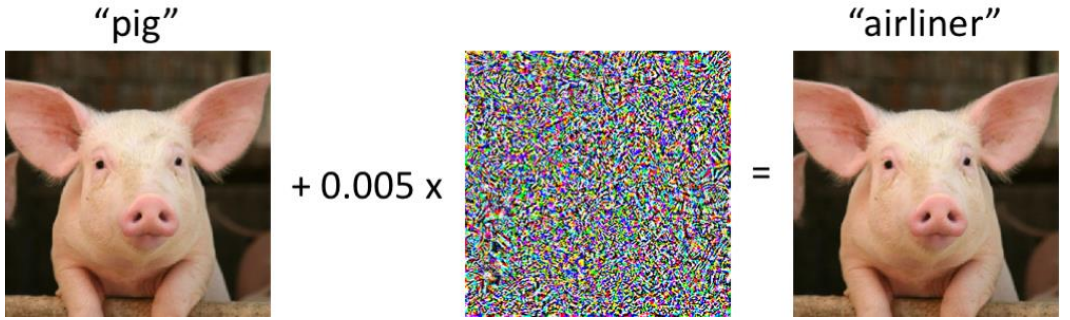
Course Web

- https://oc.sjtu.edu.cn/courses/ 58141

- **Teaching Assistant:**
  - **王剑挺　glory1229@sjtu.edu.cn**

- **成绩构成**
  - **作业20%；课程设计 30%；期末考试 50%**

讨论题



- **为什么图像加噪后容易被误分为其他物体？ 可能有哪些技术路线解决该问题？**

- 在交通拥挤期间，萝卜快跑为什么会出现停着不走，且不听交警指挥？需要我们解决关键技术是什么？

第一次作业

- 请通过测试大模型，指出大模型在回答问题/图像生成中存在的问题，写一份简易的调研报告，指出可能是什么原因，其科学问题是什么？
- 提交报告要求：字数要求>1000字，包括：测试系统、测试问题，出现不合理问题是什么？针对该问题，凝练进一步研究的科学问题。