


中国科学院大学
University of Chinese Academy of Sciences

UCAS XGS001CD SPRING 2022 Seminar

Thick Black Theory of Win the interview at the bottom for CS NPEE

计算机考研复试抄底厚黑学

Lecture 3 : Basic of Artificial Intelligence Security

Jing Li 李敬

2022年3月27日

壬寅二月廿五

北京 · 海淀

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

课程介绍 Seminar

- 课程代码：UCAS XGS001CD SPRING 2022 Seminar
- 课程名称：Thick Black Theory of Win the interview at the bottom for CS NPEE，计算机考研复试抄底厚黑学

- As we all know, participating in the NPEE (National Post-graduate Entrance Examination) and getting a good score is only a ticket to the reexamine or interview. There are many cases in which high scorers are eliminated in the retest. It's not that they didn't work hard or are not excellent, but that they didn't find then use the right method. Hence, it is very important to carefully prepare for the interview.
- In this course, we will talk about how to prepare the retest, especially interview. We will explore how to make an effective introduction letter and brand yourself. In particular, we will discuss make an excellent résumé or CV (Curriculum Vitae). We will also discuss how computer scientists and engineers are using machine learning to design state-of-the-art hardware and software security platforms. In particular, we will cover topics such as Artificial Intelligence, Computer System, Network, Open Source and Security. After completing the course, you should be able to appreciate the new trends of using thick black-driven techniques and skills in both interview and reexamine and should be prepared to start your own graduate career.

课程介绍（Cont.）

- 授课团队：
  - Jing LI 李敬 <lixeon.lij@gmail.com>，中科院信工所20级硕士研究生
  - 助教TA：Zhilu WANG 汪芷璐 <wangzhilu20@mails.ucas.edu.cn>，中科院信工所20级直博生
- 课程网站：https://lixeon.com/courses/ucas-xgs001cd-spring2022.html
- 授课方式：线上，腾讯会议
- 授课安排：

Lecture 1 : Introduction

复试面试的重要性, 抄底厚黑学思维, 润色攻心

Lecture 2 : Résumé and Brand Yourself

如何优雅制作有效简历, 社交牛逼症, 套磁的艺术, 恰到好处的恭维, 打造自己的品牌

Lecture 3 : Basic of Artificial Intelligence Security

人工智能基础、对抗机器学习、最佳实践串讲

Lecture 4 : Techniques II

计算机科学&网络空间安全&人工智能基础知识、最佳实践串讲 II

- 课程评估：https://www.wjx.cn/vj/PWcQRxR.aspx

Reference

1. 081203M04003H 高级人工智能, Fall2020, 沈华伟, 中国科学院大学
2. 081201M05005H 智能计算系统, Spring2021, 陈云霁, 中国科学院大学
3. 081201M07002H 人工智能时代的系统芯片设计, Summer2021, 陈春章, 中国科学院大学
4. CS498 Trustworthy Machine Learning, Spring2020, Bo Li, UIUC
5. A6人工智能系统, 微软人工智能教育与共建社区
6. CS404/504 Special Topics Adversarial Machine Learning, Fall2021, Alex Vakanski, UIdaho
7. 0839X2M07006H 大数据与人工智能技术, Summer2021, 岳银亮, 中国科学院大学
8. Applied Deep Learning, Spring2020, YUN-NUNG (VIVIAN) CHEN (陳縕儂), 台湾大学
9. Machine Learning, Spring 2021, HUNG-YI LEE (李宏毅), 台湾大学
10. CY7790 Machine Learning Security and Privacy, Fall2021, Alina Oprea, NEU

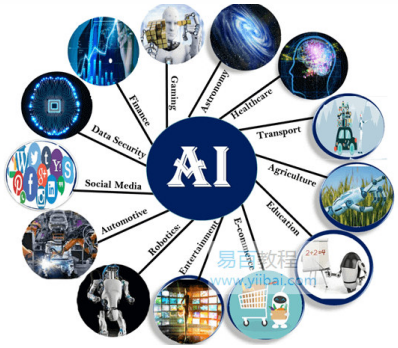
Contents

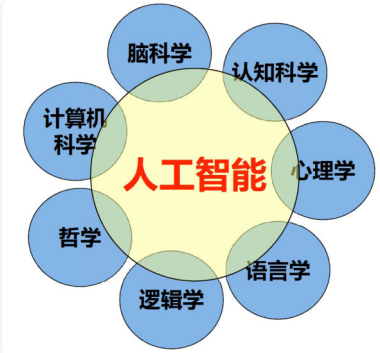
Napoleon Crossing the Alps

https://en.wikipedia.org/wiki/Napoleon_Crossing_the_Alps

Definition/Classification of Artificial Intelligence

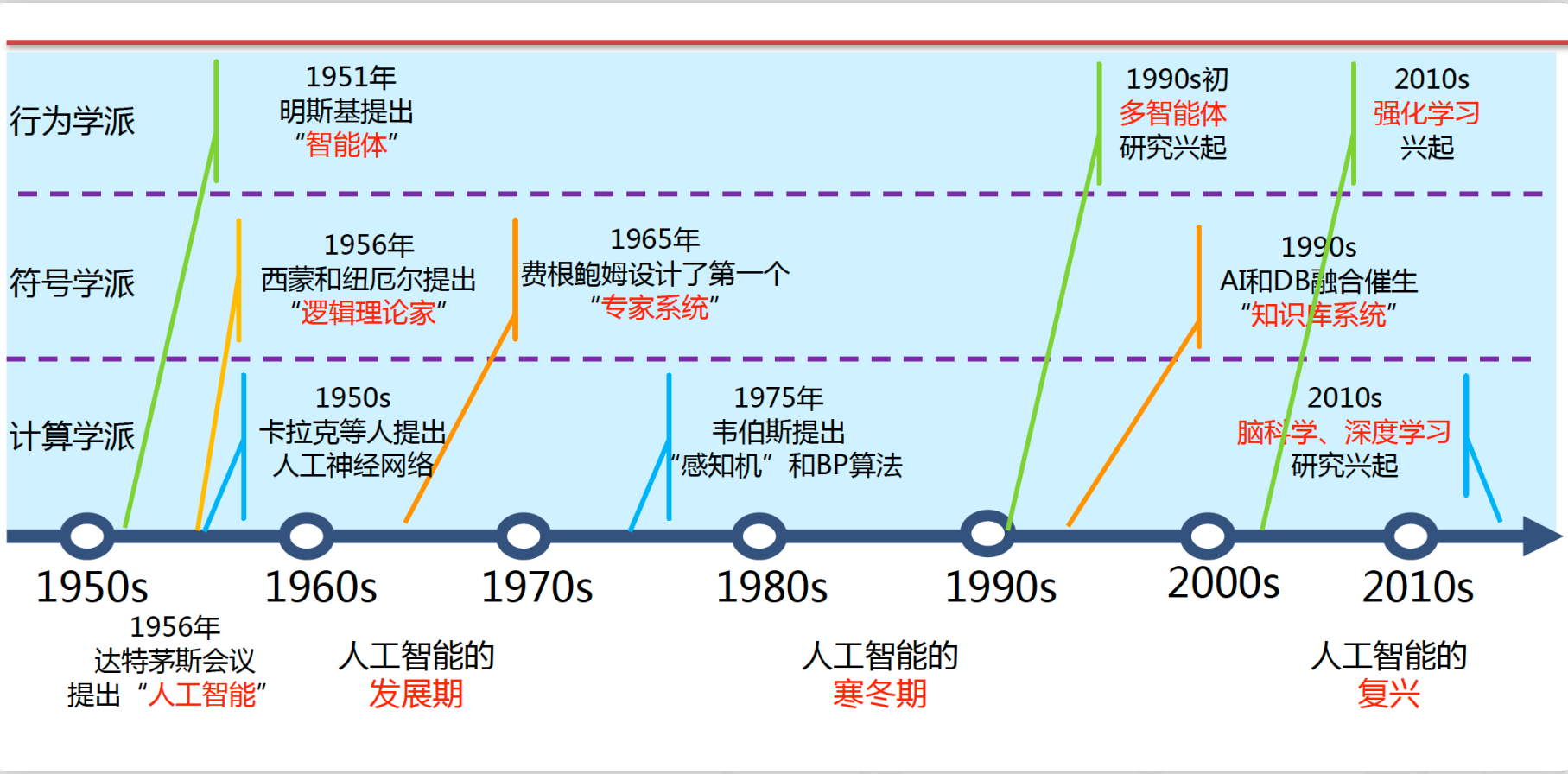
- What is AI? Merriam-Webster Dictionary:
  - "An area of computer science that deals with giving machines the ability to seem like they have human intelligence"
  - "The power of a machine to copy intelligent human behavior"
- How AI is classified?
  - Artificial Weak/Narrow Intelligence (**ANI**)
    - Focuses on improvement of individual ability, *e.g.* Siri
  - Artificial General Intelligence (**AGI**)
    - On humankind, human's brains, *e.g.* TrueNorth
  - Artificial Superintelligence (**ASI**)
    - Smarter than human brains, including innovation, recognition and social





- 机器智能：使机器具备计算和"判别"的行为能力
  - Artificial intelligence (AI) is the branch of computer science concerned with making computers intelligent, just like people.
- 类脑智能：仿生智能，让机器像人或生物一样思考
  - AI is the multidisciplinary study of human intelligence through attempts to artificially model it.
- 群体智能：社会智能的机器重现与利用、涌现智能
  - Intelligence emerged from collective behaviors of lots of agents without or with little intelligence

人工智能三大学派路线图



行为主义学派（进化、控制论）：利用机器对环境的迭代学习进化实现智能

-> 控制论、多智能体、强化学习……

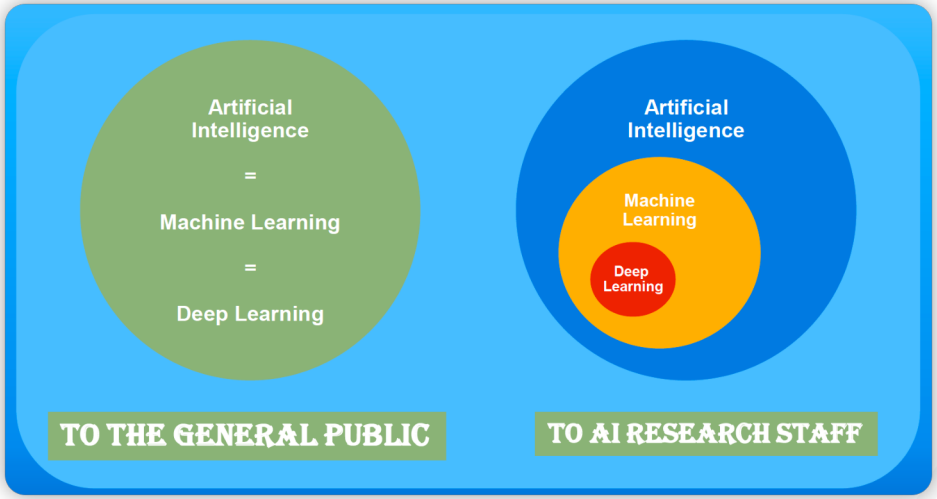
符号主义学派（逻辑）：认为人工智能的核心是知识表示、知识推理和知识运用

-> 逻辑、专家系统、知识库……

联结主义学派（仿生计算）：人工神经网络及神经网络间的连接机制和学习算法

-> 人工神经网络、认知科学、类脑计算

人工智能、机器学习、深度学习



- 机器学习：
  - 监督学习：SVM、线性判别
  - 无监督学习：k-聚类、主成分分析
  - 半监督学习：自训练法、基于图的半监督算法、半监督支持向量机
  - 强化学习：使用表格学习q_learning、sarsa 以及使用神经网络学习的DQN、直接输出行为的 Policy Gradients及Actor Critic等
- 深度学习
  - 深度神经网络、图神经网络 ...

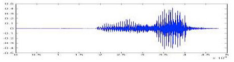
人工智能：为机器赋予人的智能

机器学习： 一种实现人工智能的方法

深度学习：一种实现机器学习的技术

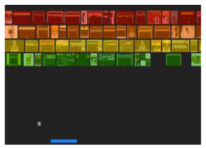
深度学习基于神经网络算法

Learning ≈ Looking for a Function

Speech Recognition $f($  $)=$ "你好"

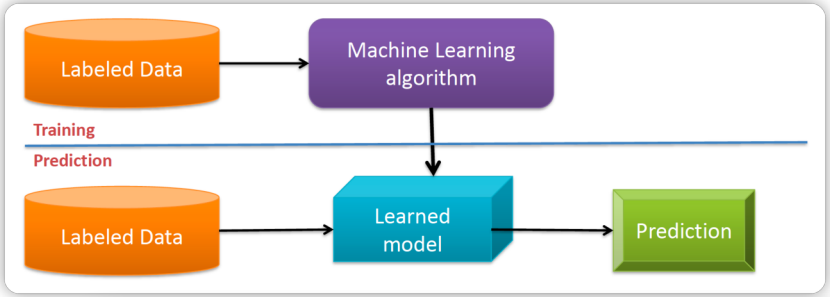
Handwritten Recognition $f($  $)=$ "2"

Weather forecast $f($ Thursday $)=$ "  Saturday"

Play video games $f($  $)=$ "move left"

机器学习数学基础

机器学习是机器从历史数据中学习规律来提升系统的某个性能度量。



- 机器学习：
  - 监督学习：SVM、线性判别
  - 无监督学习：k-聚类、主成分分析
  - 半监督学习：自训练法、基于图的半监督算法、半监督支持向量机
  - 强化学习：使用表格学习q_learning、sarsa 以及使用神经网络学习的DQN、直接输出行为的 Policy Gradients及Actor Critic等
- 深度学习
  - 深度神经网络、图神经网络 ...

Training data: $\{(\boldsymbol{x}^1, \hat{y}^1), (\boldsymbol{x}^2, \hat{y}^2), \ldots, (\boldsymbol{x}^N, \hat{y}^N)\}$

Training:

Step 1: function with unknown

$y = f_{\boldsymbol{\theta}}(\boldsymbol{x})$

Step 2: define loss from training data

$L(\boldsymbol{\theta})$

Step 3: optimization

$\boldsymbol{\theta}^* = arg \min_{\boldsymbol{\theta}} L$

Testing data: $\{\boldsymbol{x}^{N+1}, \boldsymbol{x}^{N+2}, \ldots, \boldsymbol{x}^{N+M}\}$

Use $y = f_{\boldsymbol{\theta}^*}(\boldsymbol{x})$ to label the testing data

The function we want to find …

$y = f($ no. of views on 2/26 $)$

頻道

Hung-yi Lee

篩選器

影片 | 流量來源 | 地理位置 | 觀眾年齡 | 觀眾性別 | 日期 | 訂閱狀態 | 訂閱來源 | 播放清單

日期	喜歡的人數		訂閱人數		觀看次數	
2021年1月26日	54	4.9%	69	5.5%	6,788	5.2%
2021年1月27日	60	5.4%	71	5.6%	6,242	4.7%
2021年1月28日	36	3.2%	63	5.0%	5,868	4.5%
2021年1月29日	27	2.4%	40	3.2%	4,413	3.4%
2021年1月30日	40	3.6%	40	3.2%	4,372	3.3%
2021年1月31日	47	4.2%	51	4.0%	5,135	3.9%
2021年2月1日	61	5.5%	29	2.3%	5,527	4.2%
2021年2月2日	49	4.4%	43	3.4%	5,911	4.5%
2021年2月3日	26	2.3%	44	3.5%	5,248	4.0%
2021年2月4日	43	3.9%	33	2.6%	4,771	3.6%
2021年2月5日	45	4.0%	49	3.9%	3,850	2.9%
2021年2月6日	29	2.6%	42	3.3%	3,828	2.9%
2021年2月7日	26	2.3%	46	3.6%	4,559	3.5%
2021年2月8日	38	3.4%	26	2.1%	4,772	3.6%
2021年2月9日	29	2.6%	25	2.0%	3,847	2.9%
2021年2月10日	31	2.8%	35	2.8%	3,382	2.6%

1. Function with Unknown Parameters

$y = f($  $)$

y: no. of views on 2/26, x_1: no. of views on 2/25

w and b are unknown parameters (learned from data)

Model $y = b + wx_1$ based on domain knowledge

feature

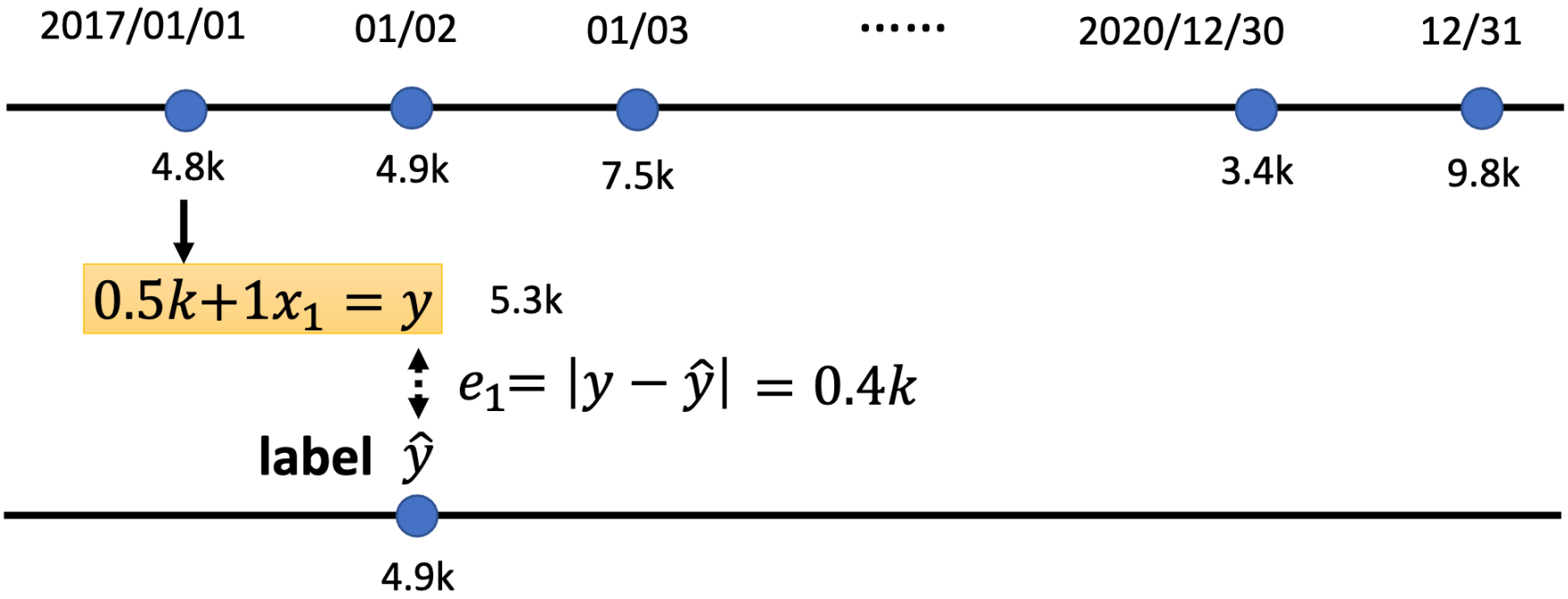
weight **bias**

2. Define Loss from Training Data

- Loss is a function of parameters $L(b, w)$
- Loss: how good a set of values is.

$L(0.5k, 1)$ $y = b + wx_1 \longrightarrow y = 0.5k + 1x_1$ How good it is?

Data from 2017/01/01 – 2020/12/31

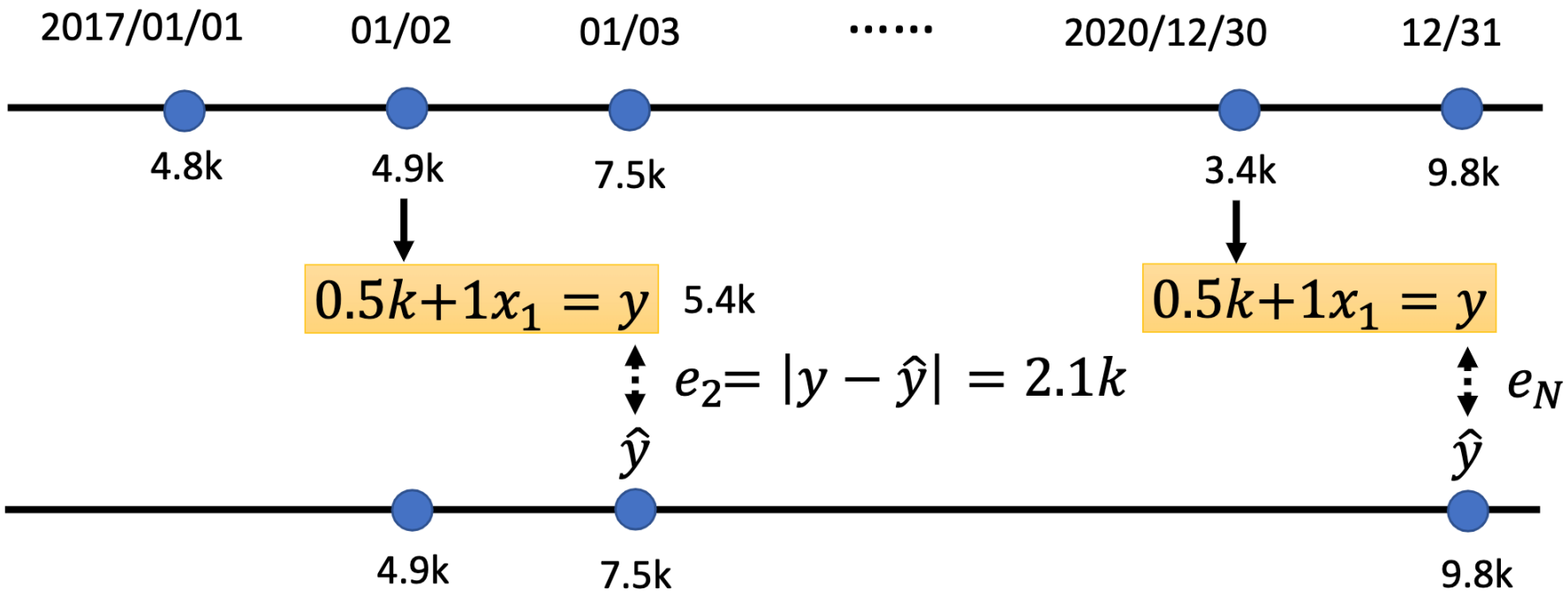


2. Define Loss from Training Data

- Loss is a function of parameters $L(b, w)$
- Loss: how good a set of values is.

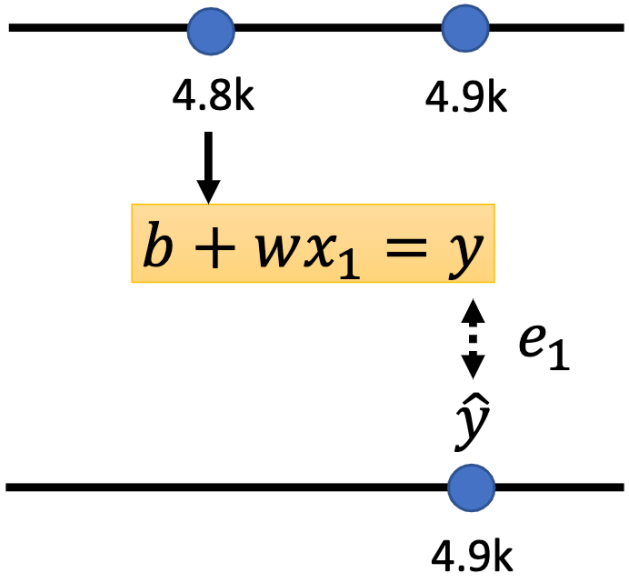
$L(0.5k, 1)$ $y = b + wx_1 \longrightarrow y = 0.5k + 1x_1$ How good it is?

Data from 2017/01/01 – 2020/12/31



2. Define Loss from Training Data

- Loss is a function of parameters $L(b, w)$
- Loss: how good a set of values is.



Loss: $L = \frac{1}{N}\sum_n e_n$

$e = |y - \hat{y}|$ L is mean absolute error (MAE)

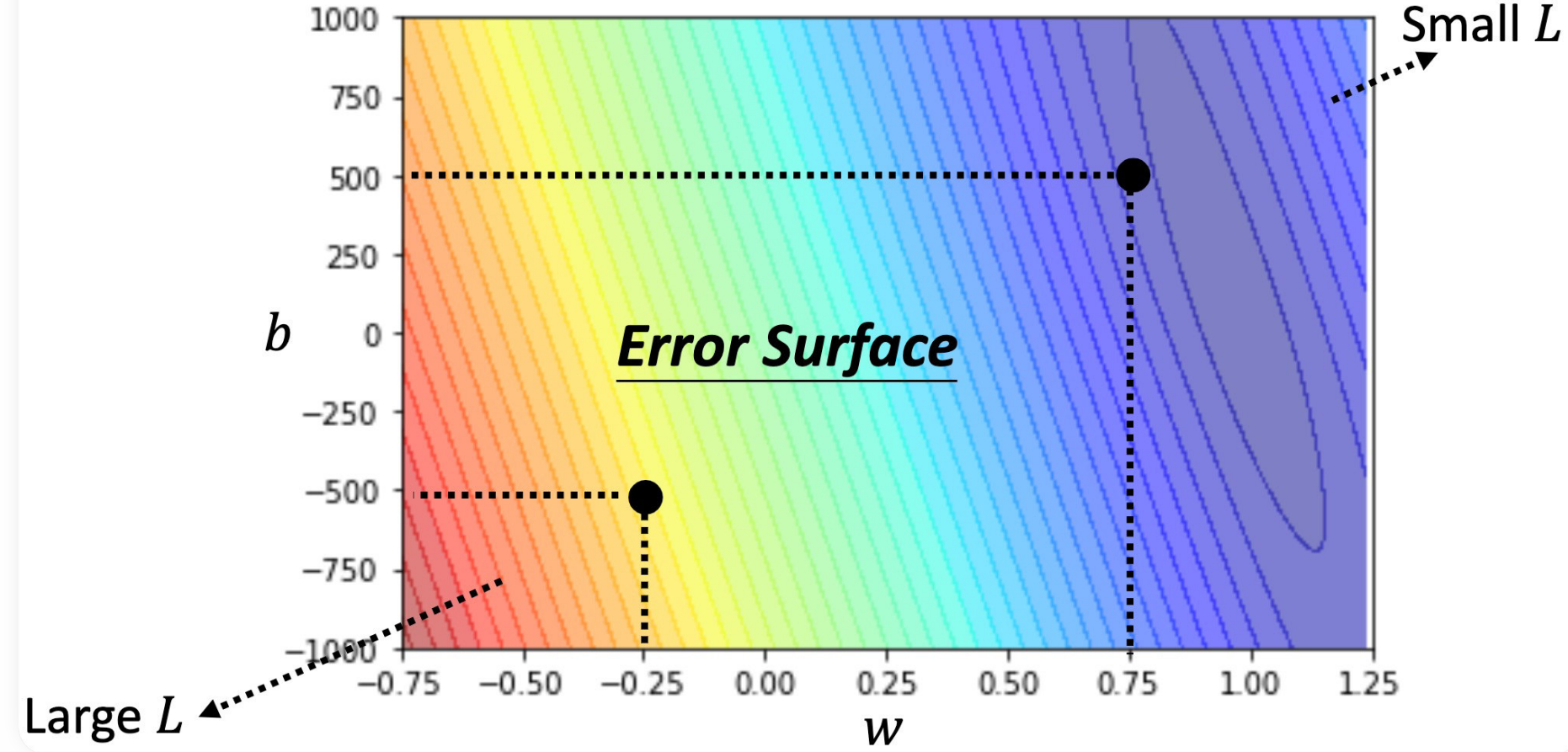
$e = (y - \hat{y})^2$ L is mean square error (MSE)

If y and $\hat{y}$ are both probability distributions ➡ Cross-entropy

2. Define Loss from Training Data

- Loss is a function of parameters $L(b, w)$
- Loss: how good a set of values is.

Model $y = b + wx_1$

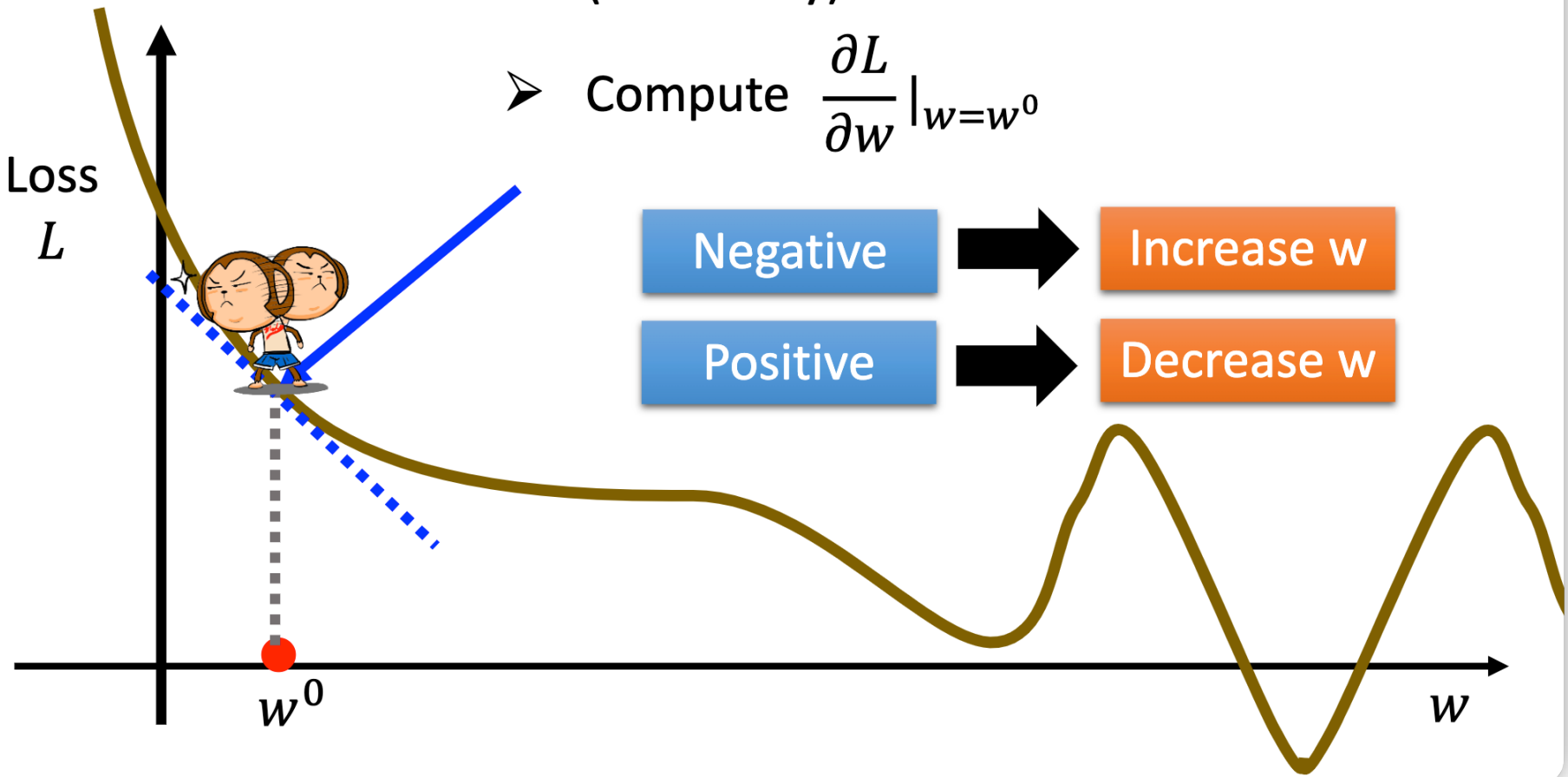


3. Optimization

$$w^* = arg \min_w L$$

Gradient Descent

- (Randomly) Pick an initial value w^0
- Compute $\frac{\partial L}{\partial w}|_{w=w^0}$



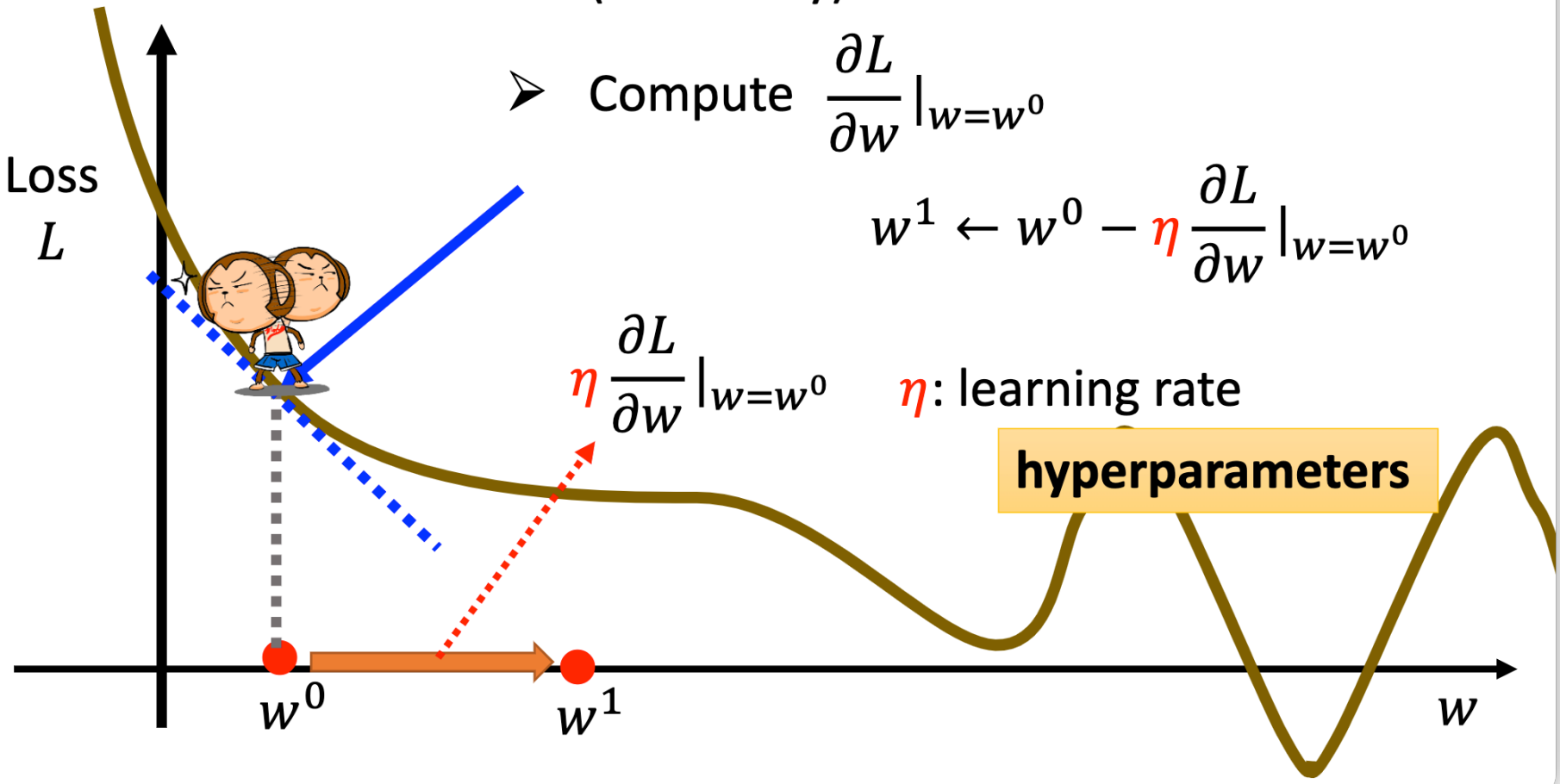
3. Optimization

$$w^* = arg \min_w L$$

Gradient Descent

- (Randomly) Pick an initial value w^0
- Compute $\frac{\partial L}{\partial w}|_{w=w^0}$

$$w^1 \leftarrow w^0 - \eta \frac{\partial L}{\partial w}|_{w=w^0}$$



3. Optimization

$$w^* = arg \min_w L$$

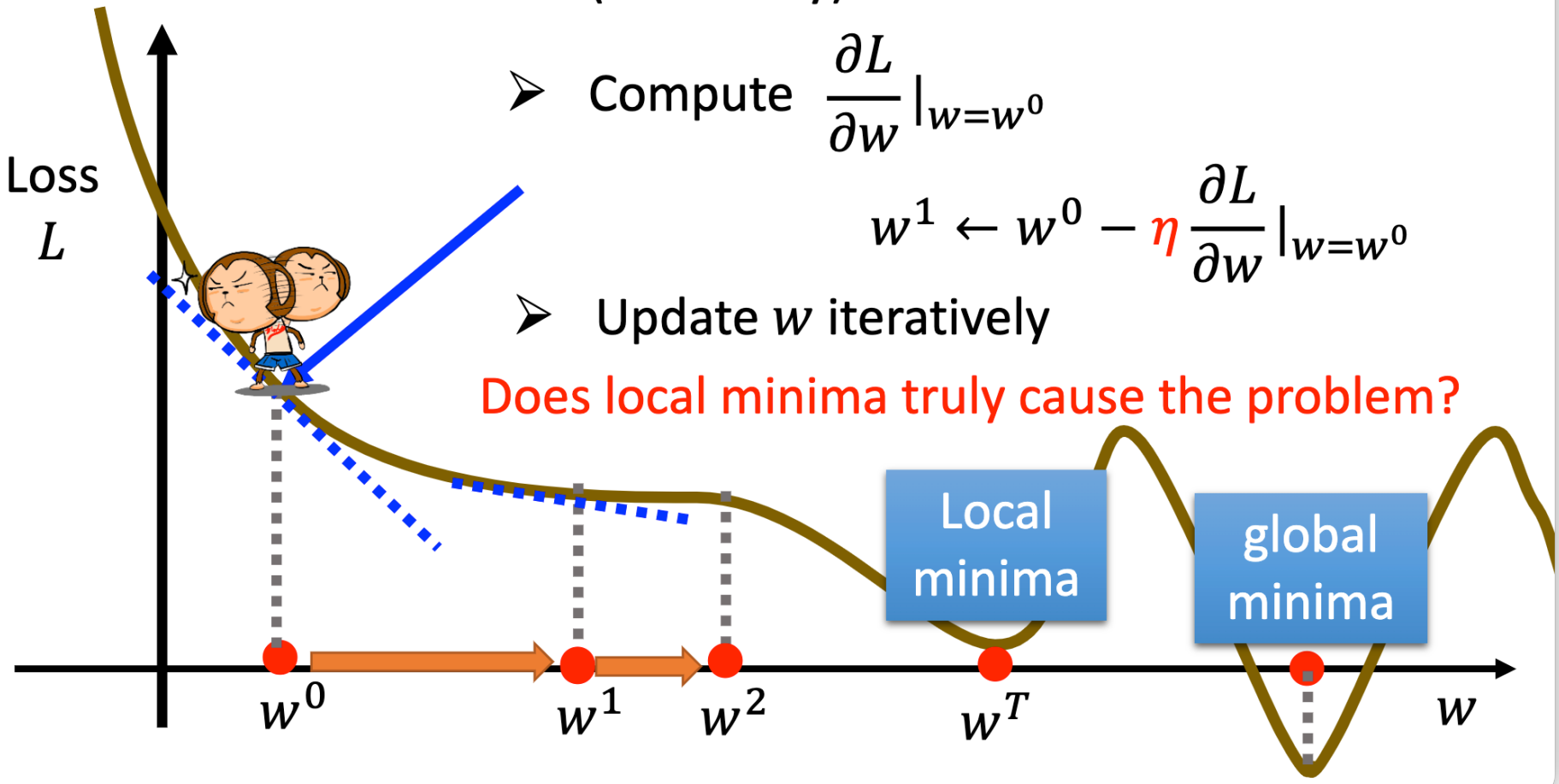
Gradient Descent

- (Randomly) Pick an initial value w^0
- Compute $\frac{\partial L}{\partial w}|_{w=w^0}$

$$w^1 \leftarrow w^0 - \eta \frac{\partial L}{\partial w}|_{w=w^0}$$

- Update w iteratively

Does local minima truly cause the problem?



3. Optimization

$$w^*, b^* = arg \min_{w,b} L$$

- (Randomly) Pick initial values w^0, b^0
- Compute

$$\frac{\partial L}{\partial w}\Big|_{w=w^0,b=b^0}$$

$$\frac{\partial L}{\partial b}\Big|_{w=w^0,b=b^0}$$

$$w^1 \leftarrow w^0 - \eta \frac{\partial L}{\partial w}\Big|_{w=w^0,b=b^0}$$

$$b^1 \leftarrow b^0 - \eta \frac{\partial L}{\partial b}\Big|_{w=w^0,b=b^0}$$

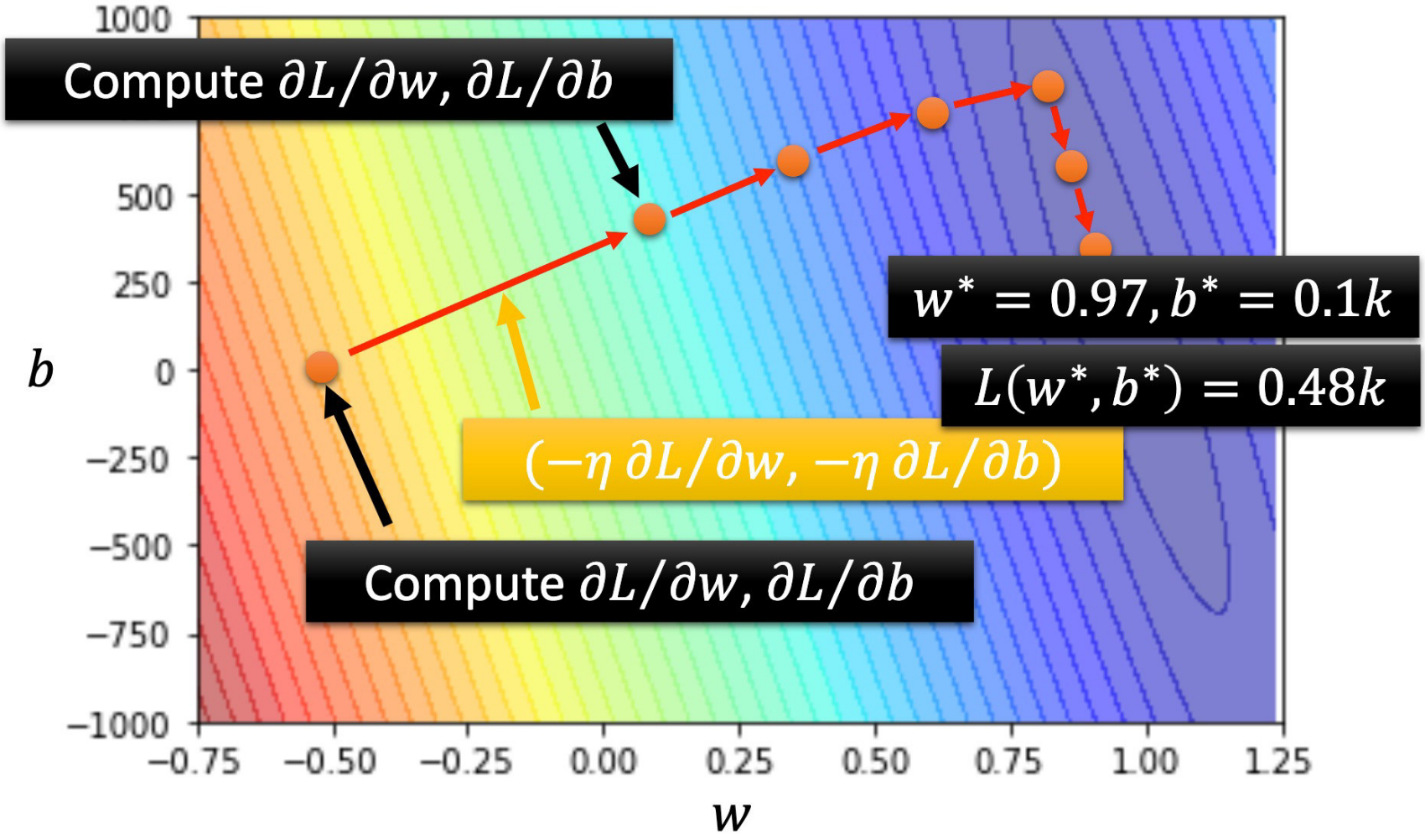
Can be done in one line in most deep learning frameworks

- Update w and b interatively

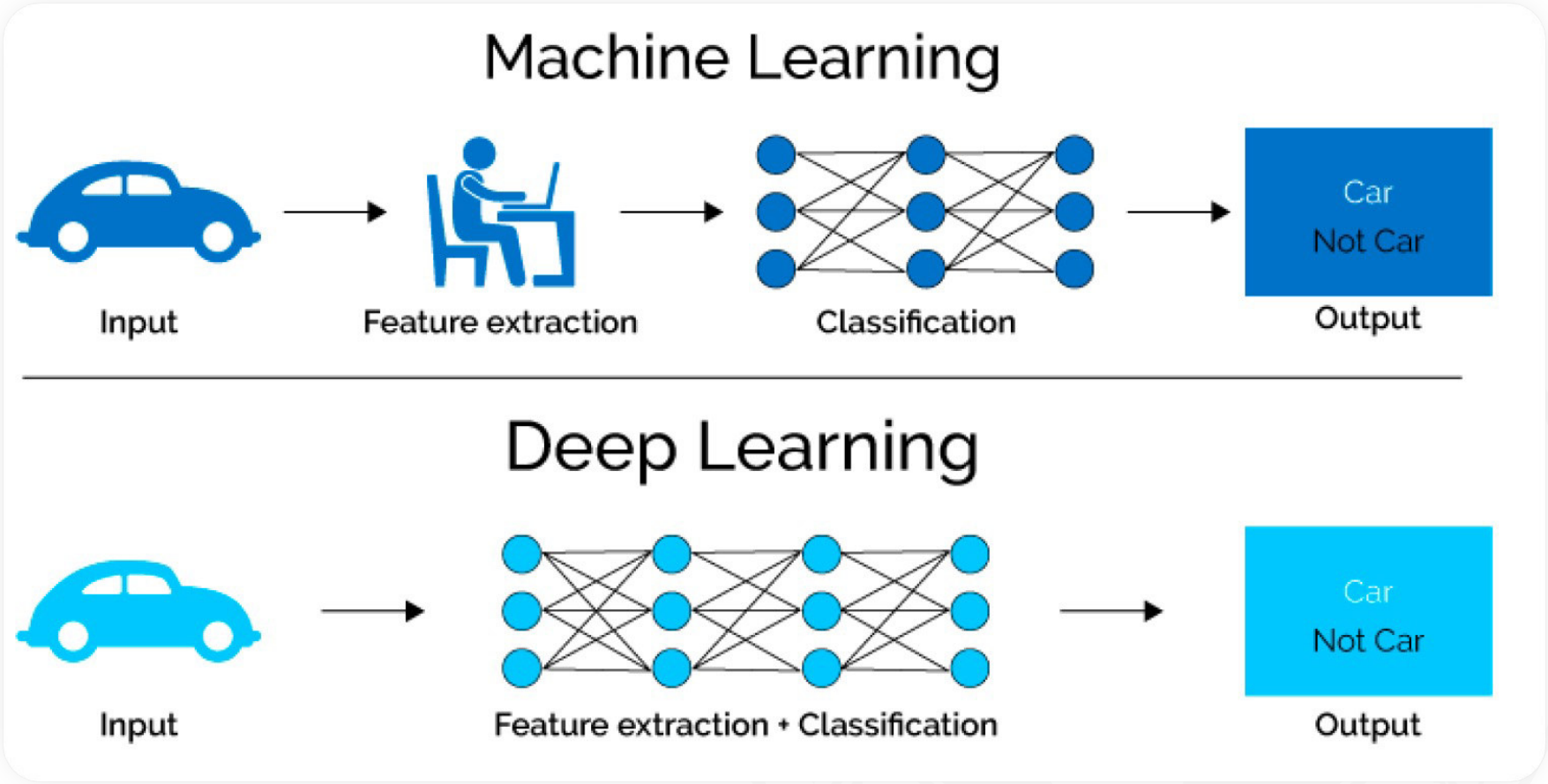
3. Optimization

Model $y = b + wx_1$

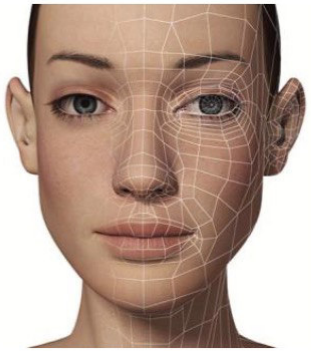
$$w^*, b^* = arg \min_{w,b} L$$



机器学习、深度学习


Machine Learning
Input
Feature extraction
Classification
Car
Not Car
Output
Deep Learning
Input
Feature extraction + Classification
Car
Not Car
Output

- 例子2：人脸识别
  - 传统的方法：基于手动特征



  - 鼻子多大
  - 双眼距离
  - 眼睛多大
  - 肤色白不白?
  - 嘴巴对鼻子距离
  - 嘴巴多大
  - 脸多宽
  - 。。。

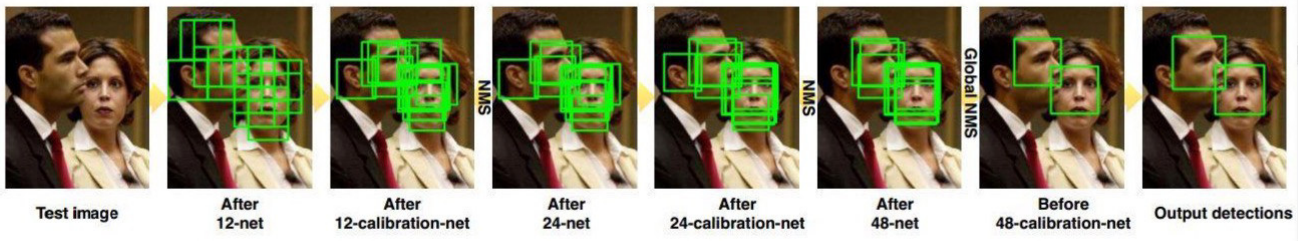
扎克伯格 ✓

扎克伯格 ✗

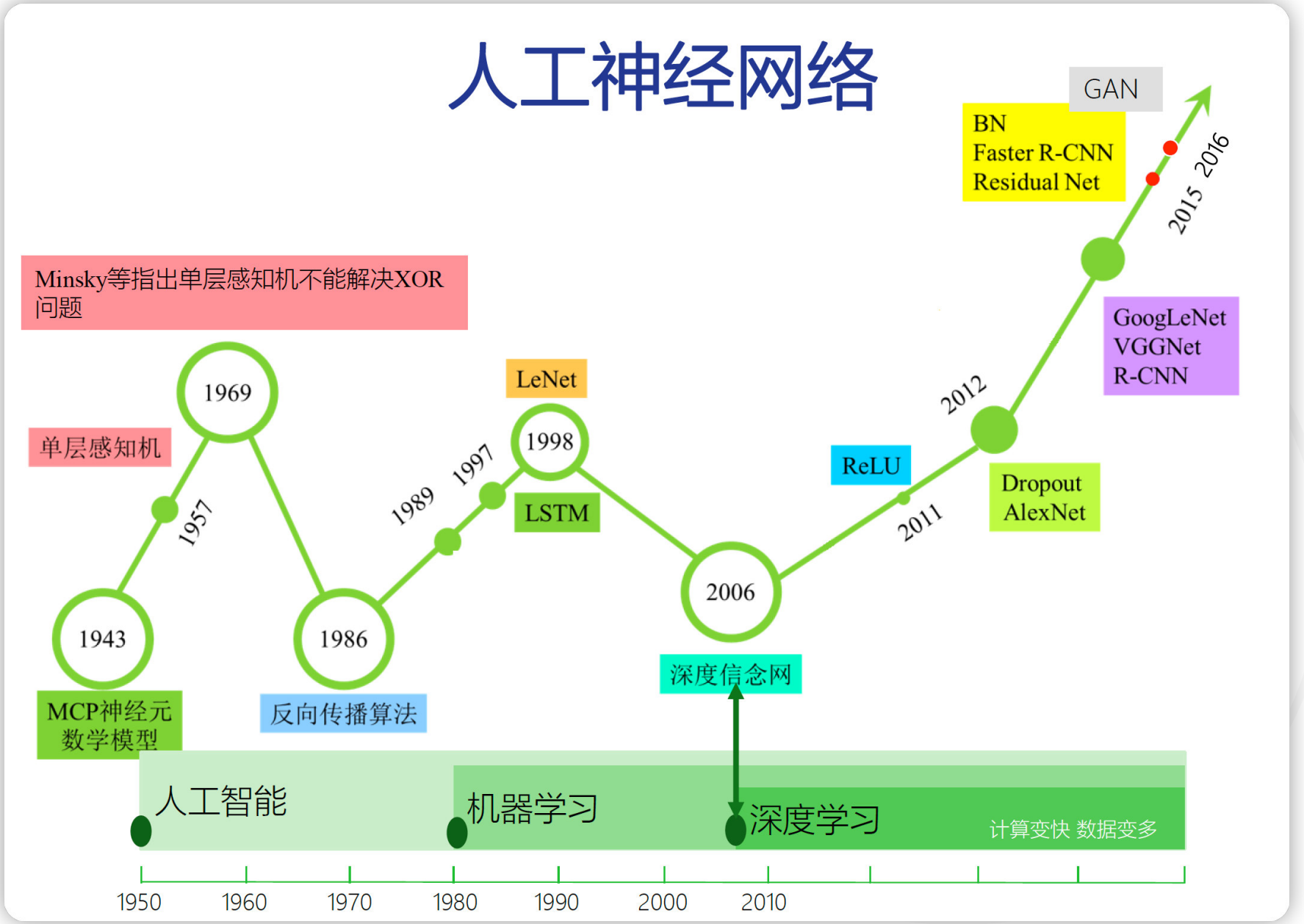
- 有几张脸能完全一样?
- 我们白着呢!
- 这么多人怎么量距离?

“前世的五百次回眸换得今生的一次擦肩而过”

  - 现在的方法：机器自动提特征



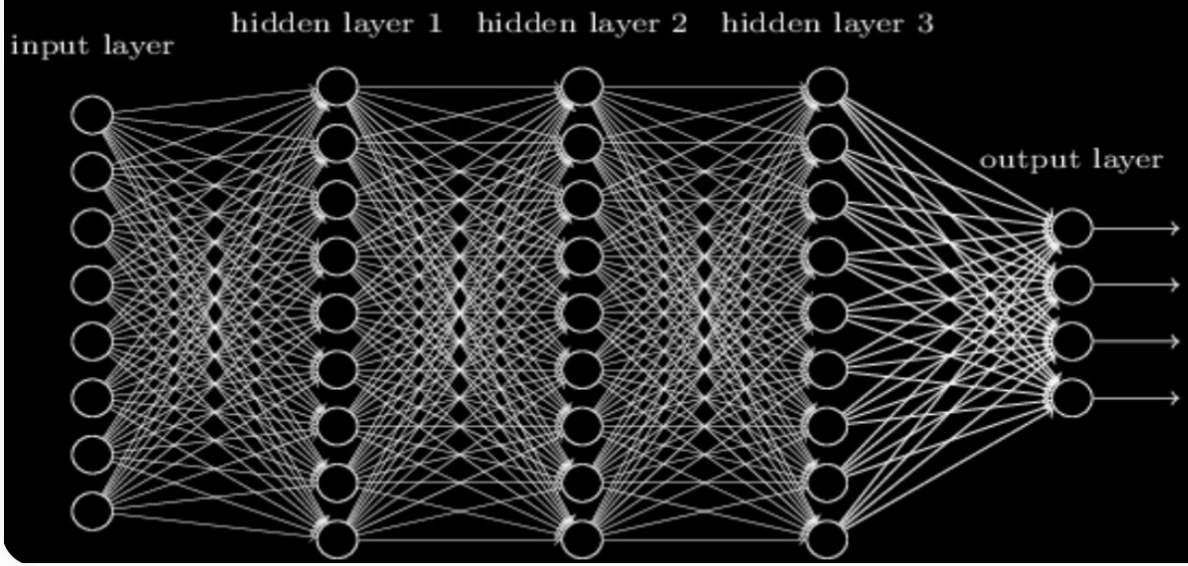
人工神经网络发展


人工神经网络
GAN
BN
Faster R-CNN
Residual Net
2015
2016
Minsky等指出单层感知机不能解决XOR问题
GoogLeNet
VGGNet
R-CNN
1969
单层感知机
LeNet
1998
1989
1997
LSTM
1957
ReLU
2012
Dropout
AlexNet
2011
1943
1986
2006
MCP神经元
数学模型
反向传播算法
深度信念网
人工智能
机器学习
深度学习
计算变快 数据变多
1950
1960
1970
1980
1990
2000
2010

神经网络最近的成功

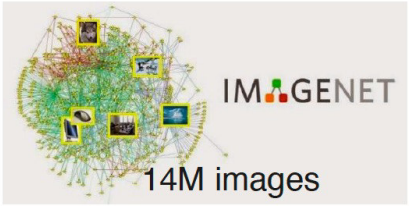
Neural networks return and excel at image recognition, speech recognition, …

The 2018 Turing award was given to Yoshua Bengio, Geoff Hinton, and Yann LeCun.

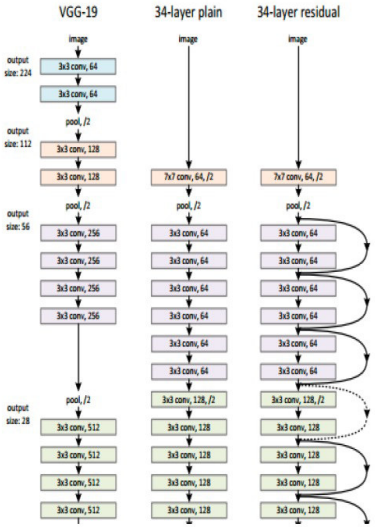


深度学习最近十年成功的原因

海量的（标识）数据

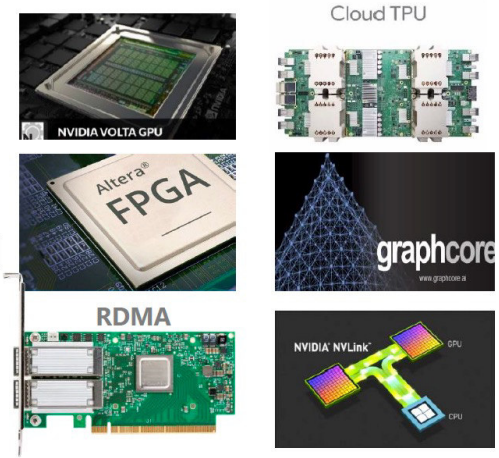


深度学习算法的进步



语言、框架

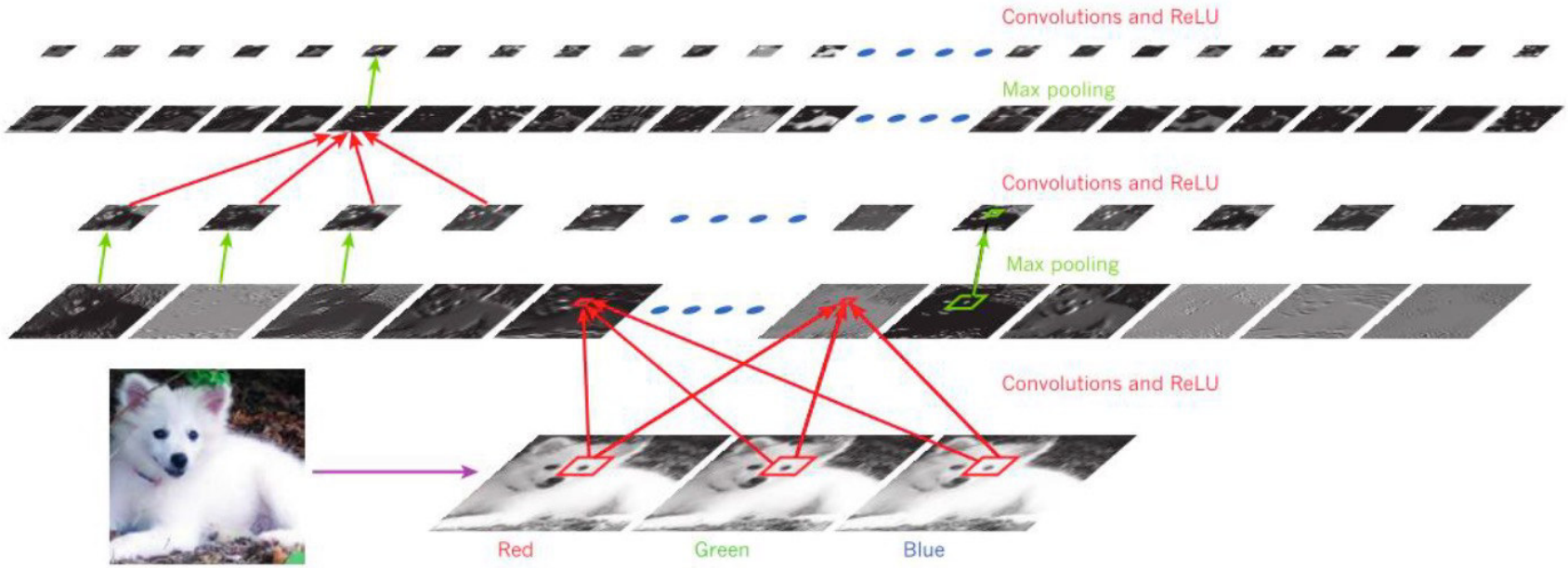
计算能力



深度学习+系统的进步: 编程语言、优化、计算机体系结构、并行计算以及分布式系统

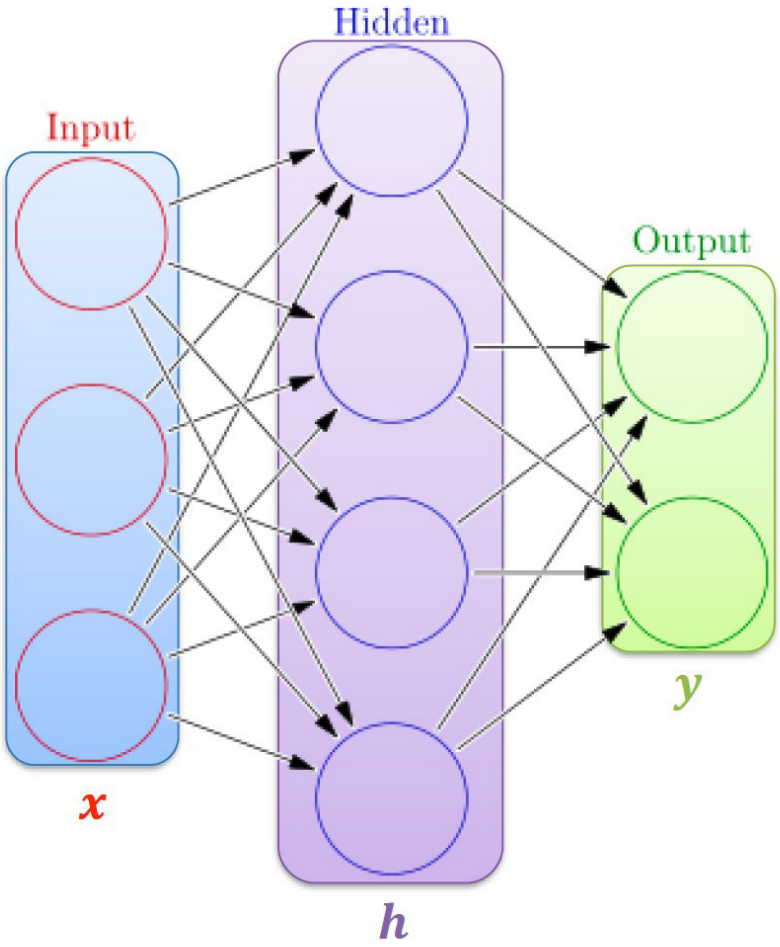
神经网络的本质

本质：是一个**多维拟合**的过程，允许机器获得原始数据，通过组合简单但非线性的模块获得所需的表示，每个模块能够**自动调整神经网络的权值**



一个简单的神经网络例子

- A NN with one hidden layer and one output layer



Weights Biases

$$hidden\ layer\ h = \sigma(W_1 x + b_1)$$

$$output\ layer\ y = \sigma(W_2 h + b_2)$$

Activation functions

4 + 2 = 6 neurons (not counting inputs)

[3 × 4] + [4 × 2] = 20 weights

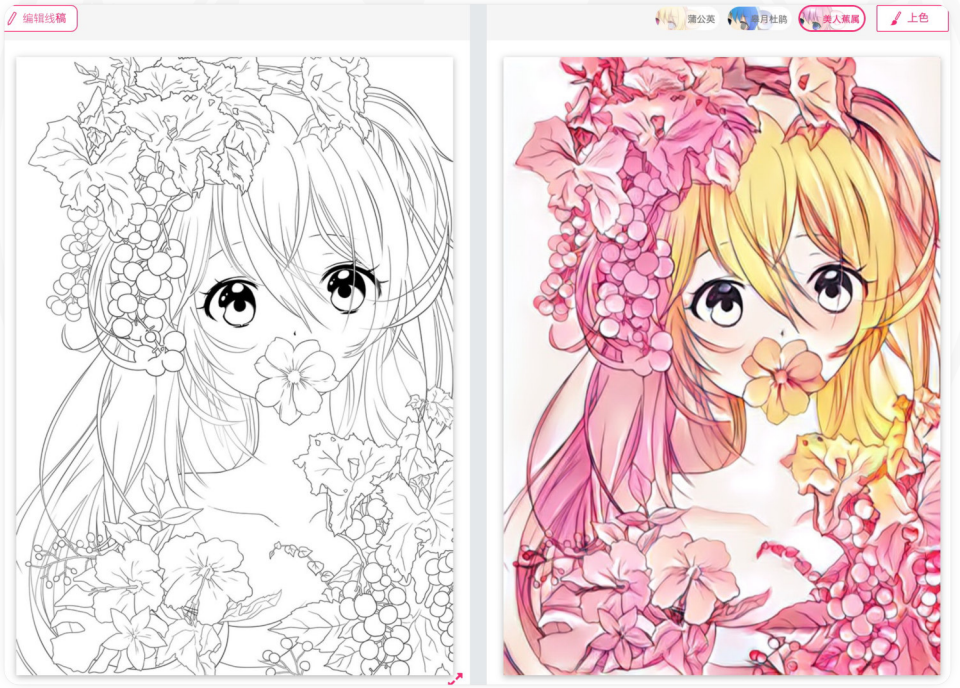
4 + 2 = 6 biases

26 learnable parameters

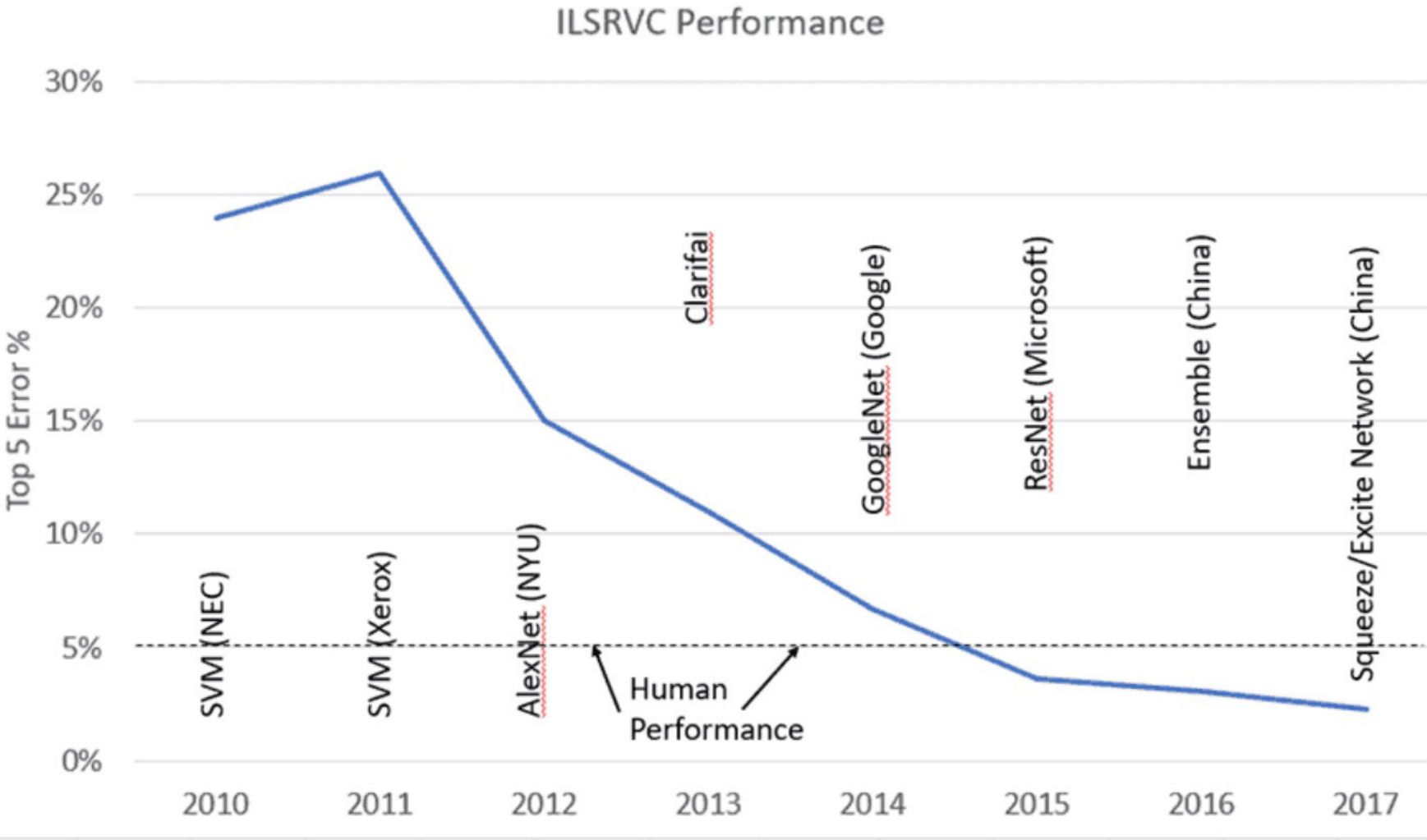
机器学习应用

- Healthcare
- Vision
- NLP
- Speech recognition
- Self-driving cars
- Stock market analysis
- Recommendations
- Sentiment analysis
- Human behavior
- Quality of life
- Business
- Sports
- Bots / chatbots
- Science / engineering
- Bioinformatics
- Precision medicine
- Unsupervised learning
- Reinforcement learning

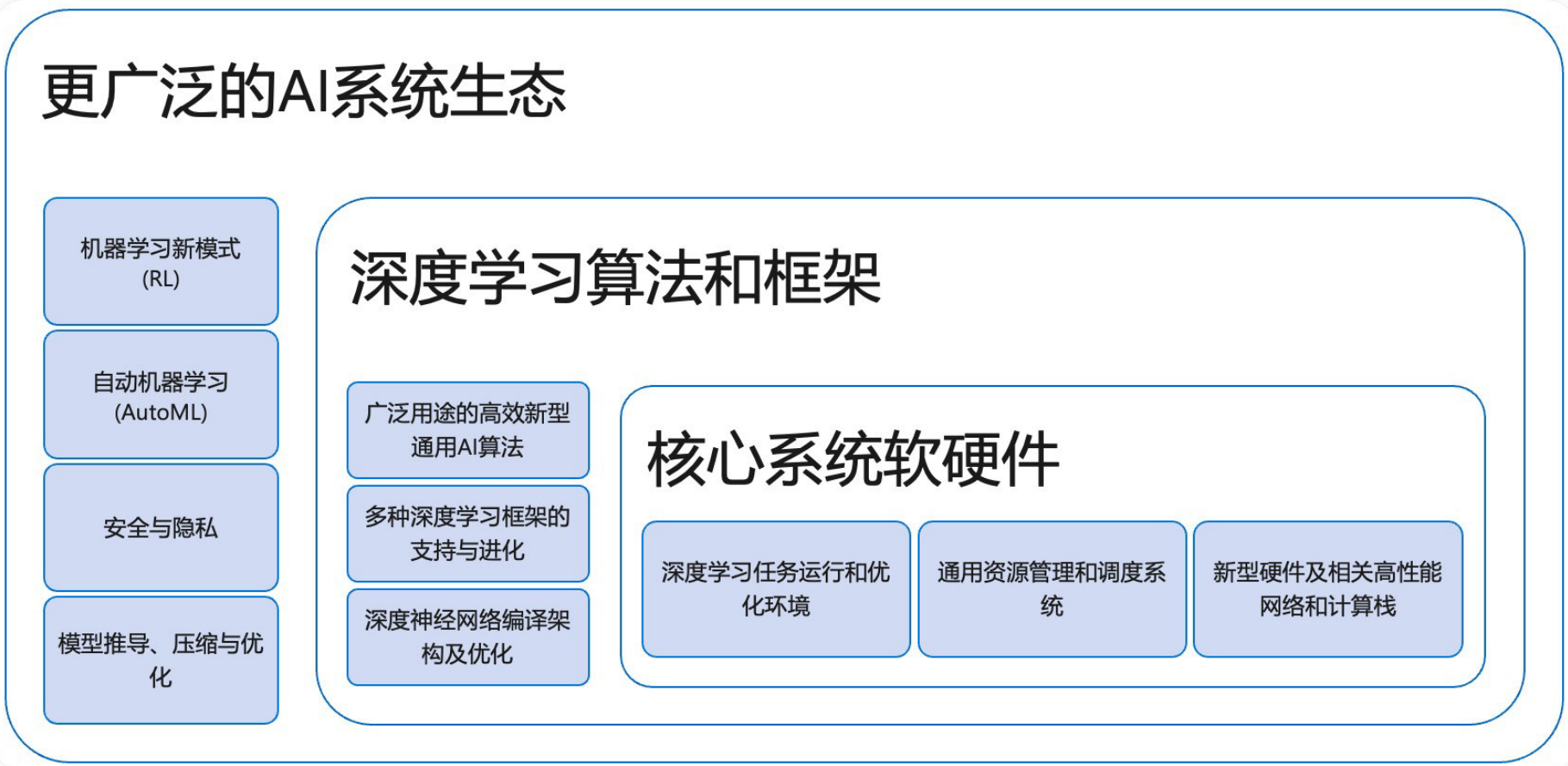
- 自然语言处理：
  - 句法语义分析、信息抽取、文本挖掘、信息检索、机器翻译、问答系统、对话系统、...
- 计算机视觉：
  - 图像分类、目标跟踪、语义分割、...
  - 人脸识别、场景文字识别、旗帜识别、台标识别、...
- 语音识别：
  - 说话人识别、声纹识别、...
  - 语音识别、语音转文字、...



ILSRVC Top5 Error

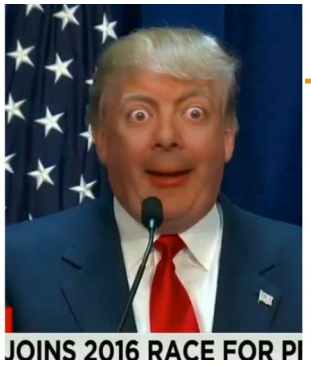


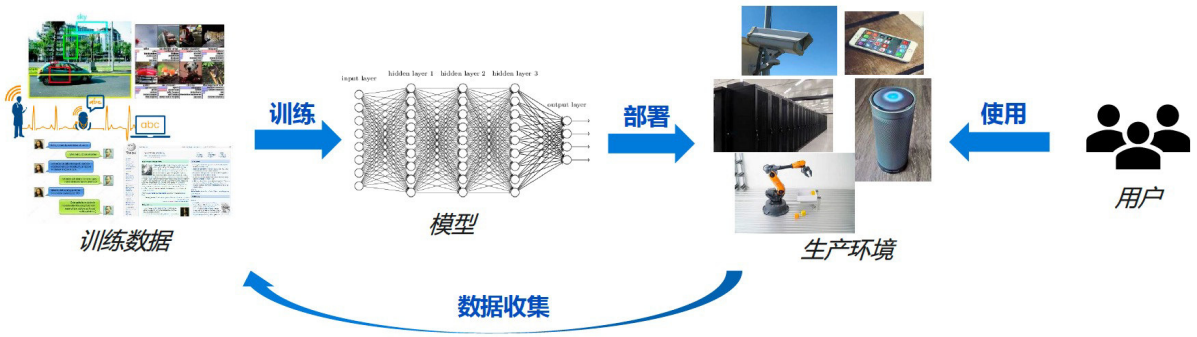
深度学习系统生态


更广泛的AI系统生态
机器学习新模式 (RL)
自动机器学习 (AutoML)
安全与隐私
模型推导、压缩与优化
深度学习算法和框架
广泛用途的高效新型通用AI算法
多种深度学习框架的支持与进化
深度神经网络编译架构及优化
核心系统软硬件
深度学习任务运行和优化环境
通用资源管理和调度系统
新型硬件及相关高性能网络和计算栈

深度学习安全

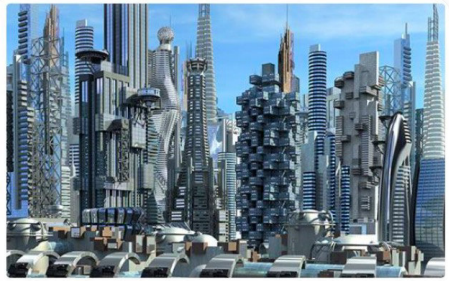
- **完整性** (Integrity): 模型输出正确的、符合预期的结果
  - 预测结果的鲁棒性 (Robustness)
  - 预测过程的可信性 (Trustworthy)
- **保密性** (Confidentiality): 训练数据、模型算法不被泄露
  - 数据保密性 (Data Confidentiality)
  - 模型保密性 (Model Confidentiality)
- **伦理** (Ethics): 应用和结果符合法律、道德、伦理
  - AI算法的公平性 (Fairness)
  - AI算法的滥用 (Misuse)





- **安全攻击**可能发生在训练数据、模型、生产环境中
- **隐私泄露**可能发生在训练、部署、使用、数据收集阶段
- 正常的数据和模型也可能产生**不当的后果**

WE NEED RESPONSIBLE AI!



VERIFIED

A Global Pandemic Will Force These Industry Sectors To Strike Smart

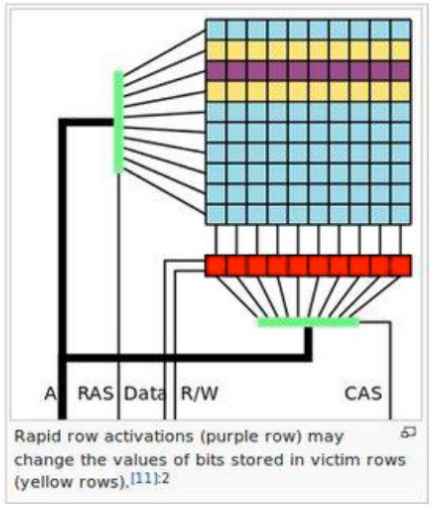
As the world is gearing up for the coronavirus pandemic, cities are vowing to not let the virus move in to prevent the

notrealnews.net

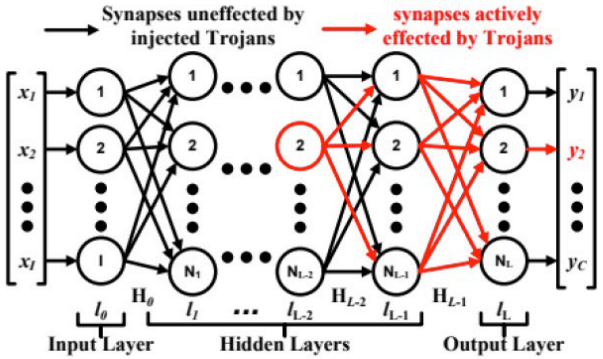
AI生成假新闻

深度学习系统攻击

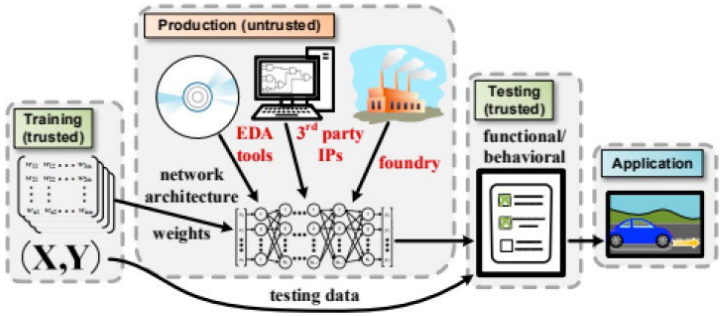
- 位翻转（Bit-flipping）攻击
  - [Rakin ICCV'2019] ResNet18模型中，翻转13个字节即可使模型准确率从70%降到0.1%。
  - 位翻转操作可以使用Row-Hammer Attacker完成
- 硬件后门植入
  - [Clements 2018] 通过硬件电路设计，在检测到trigger时改变神经网络中的某个neuron激活值



Row-Hammer Attack (RHA)

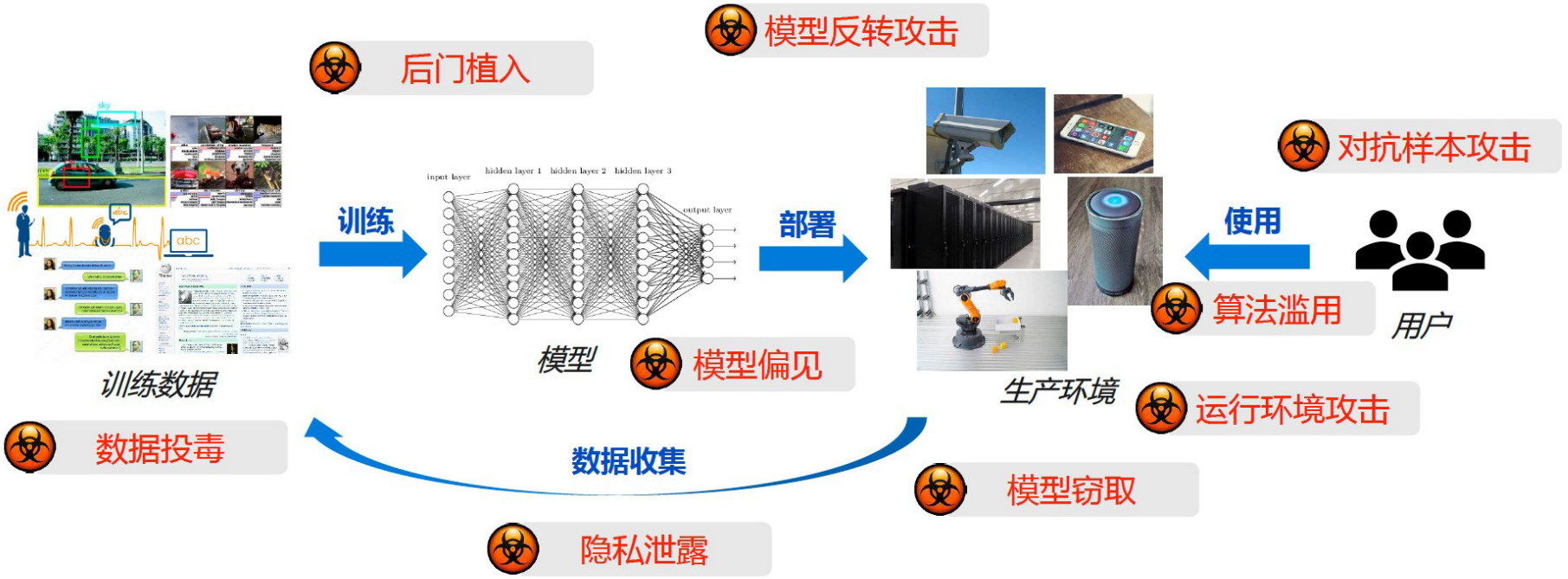


[Rakin ICCV2019]



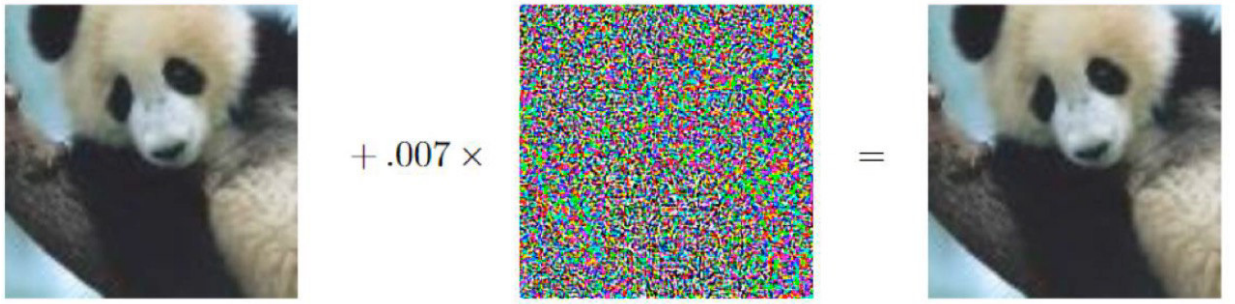
[Clements 2018]

机器学习安全攻击方式汇总

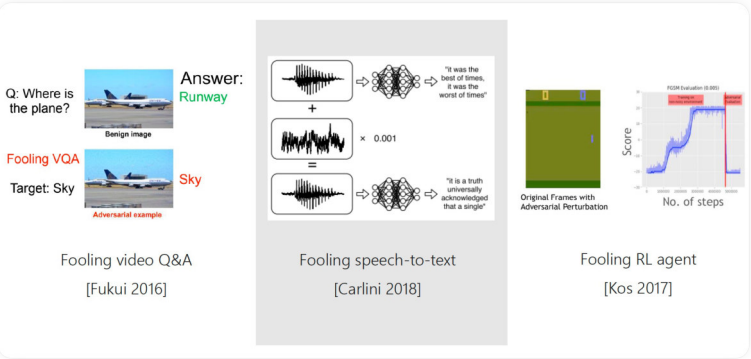


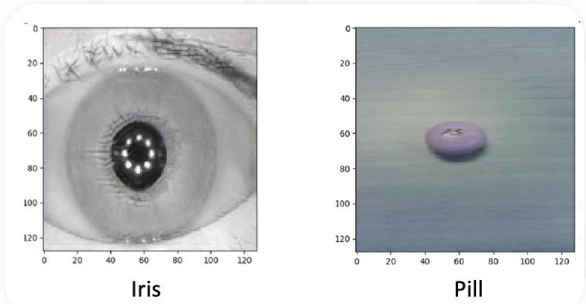
对抗样本攻击

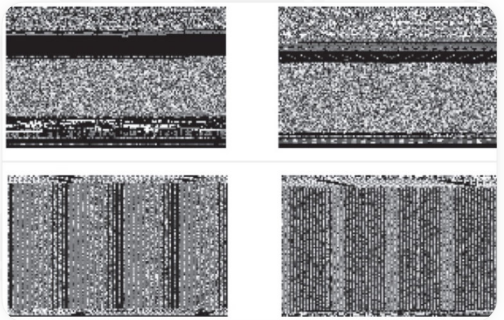
对正常样本增加一个微弱的（肉眼无法识别的）扰动，就能导致模型预测出错 [Szegedy ICLR2014],[Goodfellow ICLR2015]。



Fast gradient sign method (FGSM): $\boldsymbol{\eta} = \epsilon\text{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right)$







图像、文本、音频、视频、物理域（安防、自动驾驶）、网络空间（恶意流量、恶意软件）、生信影像

寻找对抗样本的一种传统方法的原理

One common approach to finding adversarial examples is as follows

- Take an image x, which is labeled by the classifier C (e.g., Logistic Regression, SVM, or NN) as class y, i.e., $C(x) = y$
- Create an adversarial image x_{adv} by adding small perturbations δ to the original image x, i.e., $x_{adv} = x + \delta$, such that the distance $D(x, x_{adv}) = D(x, x + \delta)$ is minimal
- The aim for a non-targeted attack is that the classifier assigns a label to the adversarial image that is different than y, i.e., $C(x_{adv}) = C(x + \delta) \neq y$
  - Or, for a targeted attack, the aim is that $C(x_{adv}) = C(x + \delta) = t \neq y$, where t is the target class

$$\begin{aligned} \text{minimize} \quad & \mathcal{D}(x, x+\delta) \\ \text{such that} \quad & C(x+\delta) = t \\ & x + \delta \in [0,1]^n \end{aligned}$$

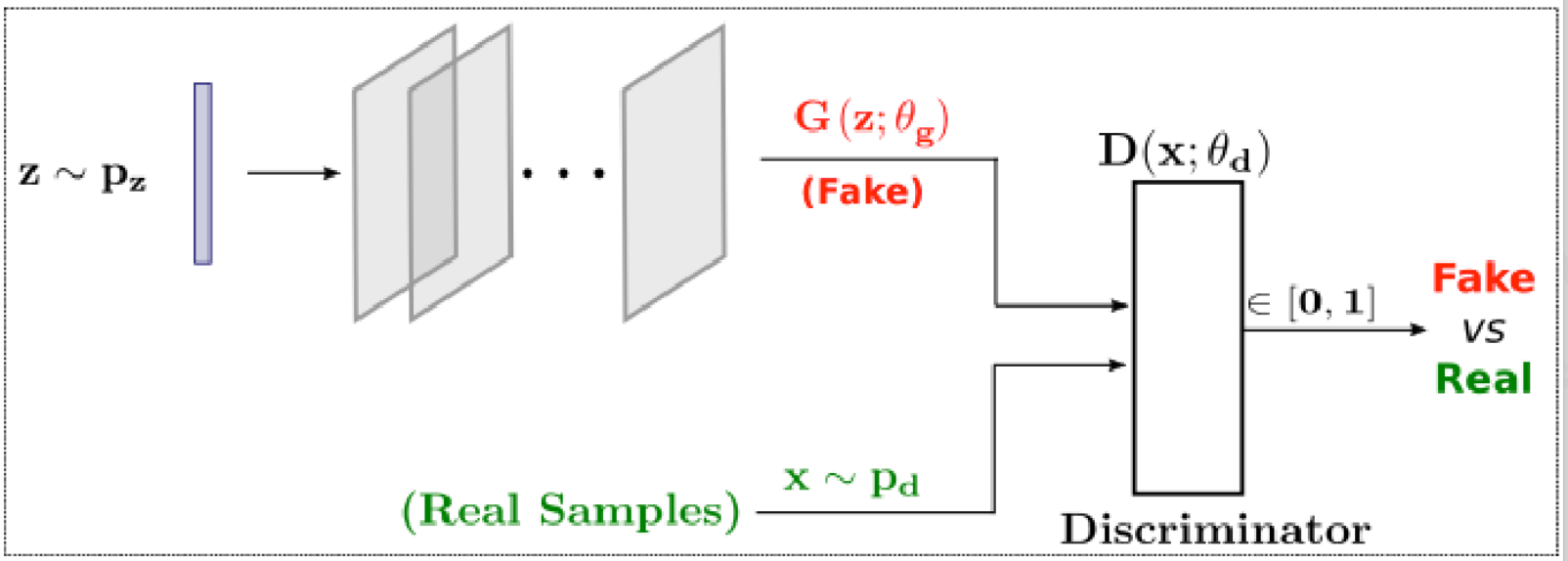
distance between x and $x+\delta = x_{adv}$

$x+\delta$ is classified as target class t

each element of $x+\delta$ is in [0,1] (to be a valid image)

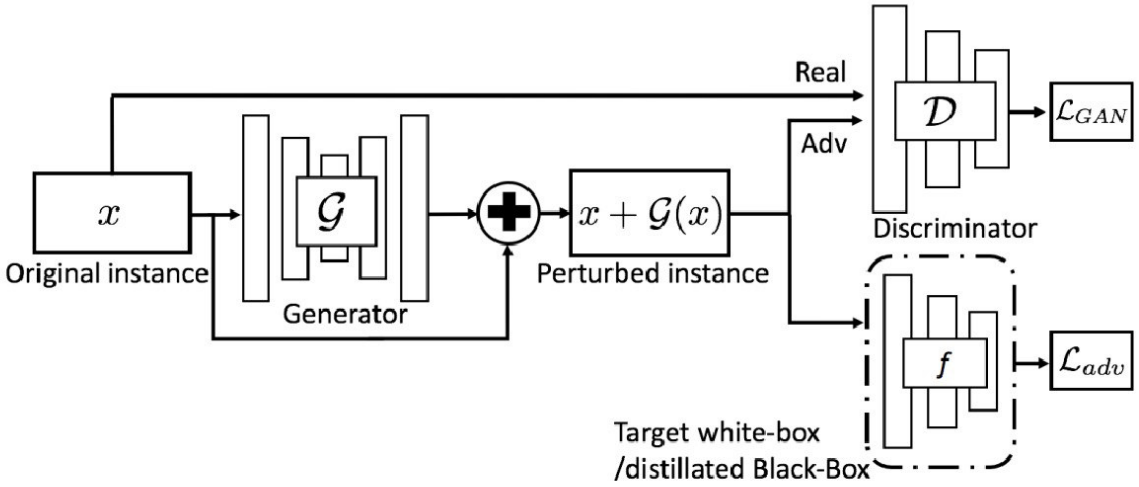
寻找对抗样本的GAN方法的原理

Generative adversarial networks (GANs)



- Generate more realistic instances
- Approximate certain distribution
- Efficient once the generator is trained

寻找对抗样本的GAN方法的原理 (Cont.)



Black-box can be performed here via distillation

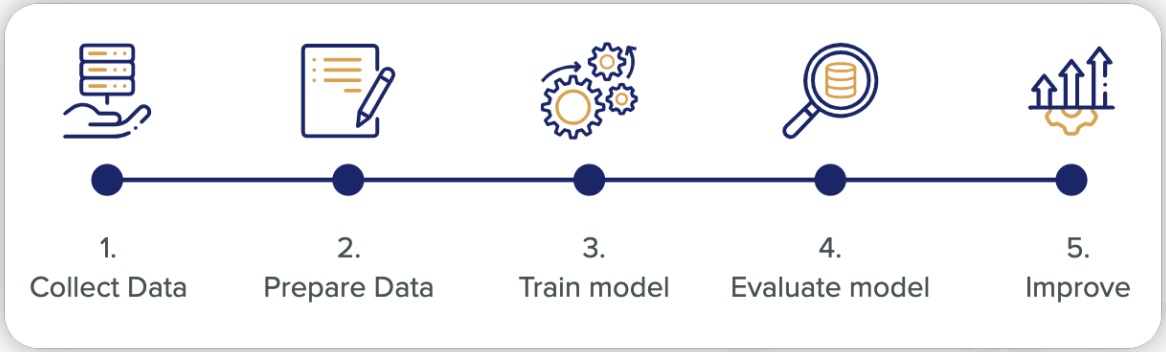
$$\mathcal{L} = \mathcal{L}_{adv}^{f} + \alpha \mathcal{L}_{GAN} + \beta \mathcal{L}_{hinge}$$

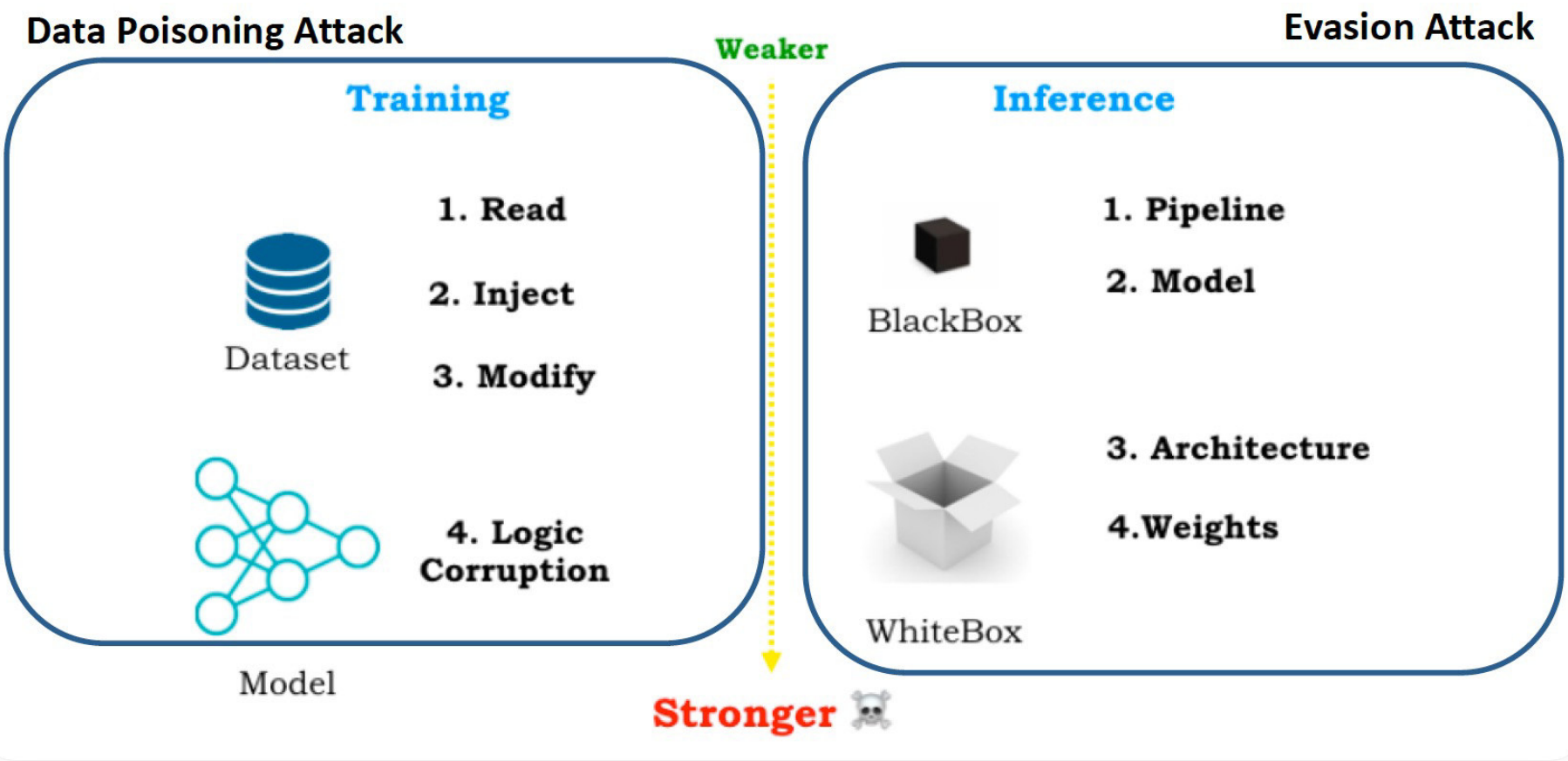
$$\mathcal{L}_{GAN} = \mathbb{E}_{x \sim \mathcal{P}_{\text{data}}(x)} \log \mathcal{D}(x) + \mathbb{E}_{x \sim \mathcal{P}_{\text{data}}(x)} \log(1 - \mathcal{D}(x + \mathcal{G}(x)))$$

$$\mathcal{L}_{\text{adv}}^{f} = \mathbb{E}_{x} \ell_{f}(x + \mathcal{G}(x), t)$$

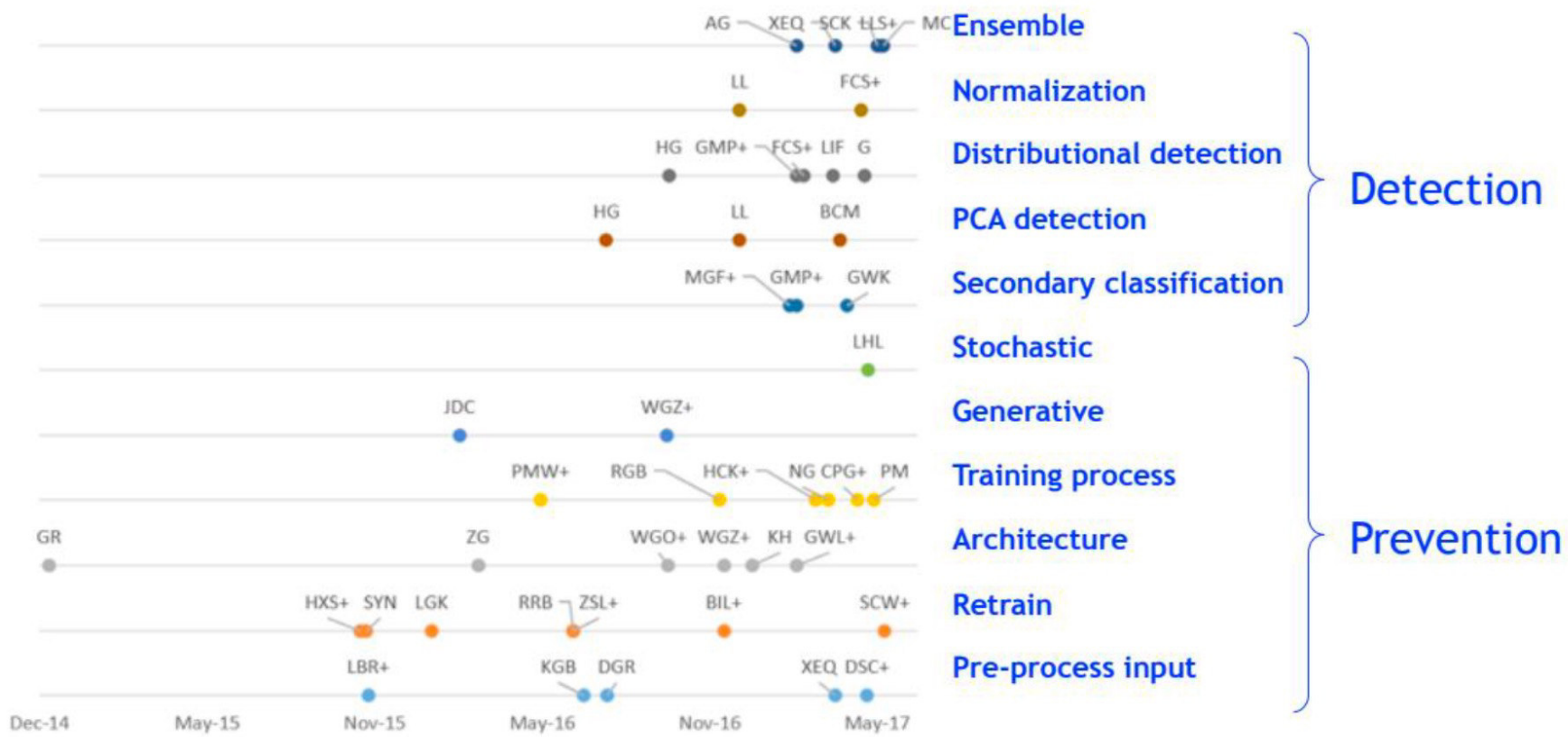
$$\mathcal{L}_{\text{hinge}} = \mathbb{E}_{x} \max(0, \|\mathcal{G}(x)\|_2 - c)$$

对抗样本攻击按供应链分类





对抗样本防御

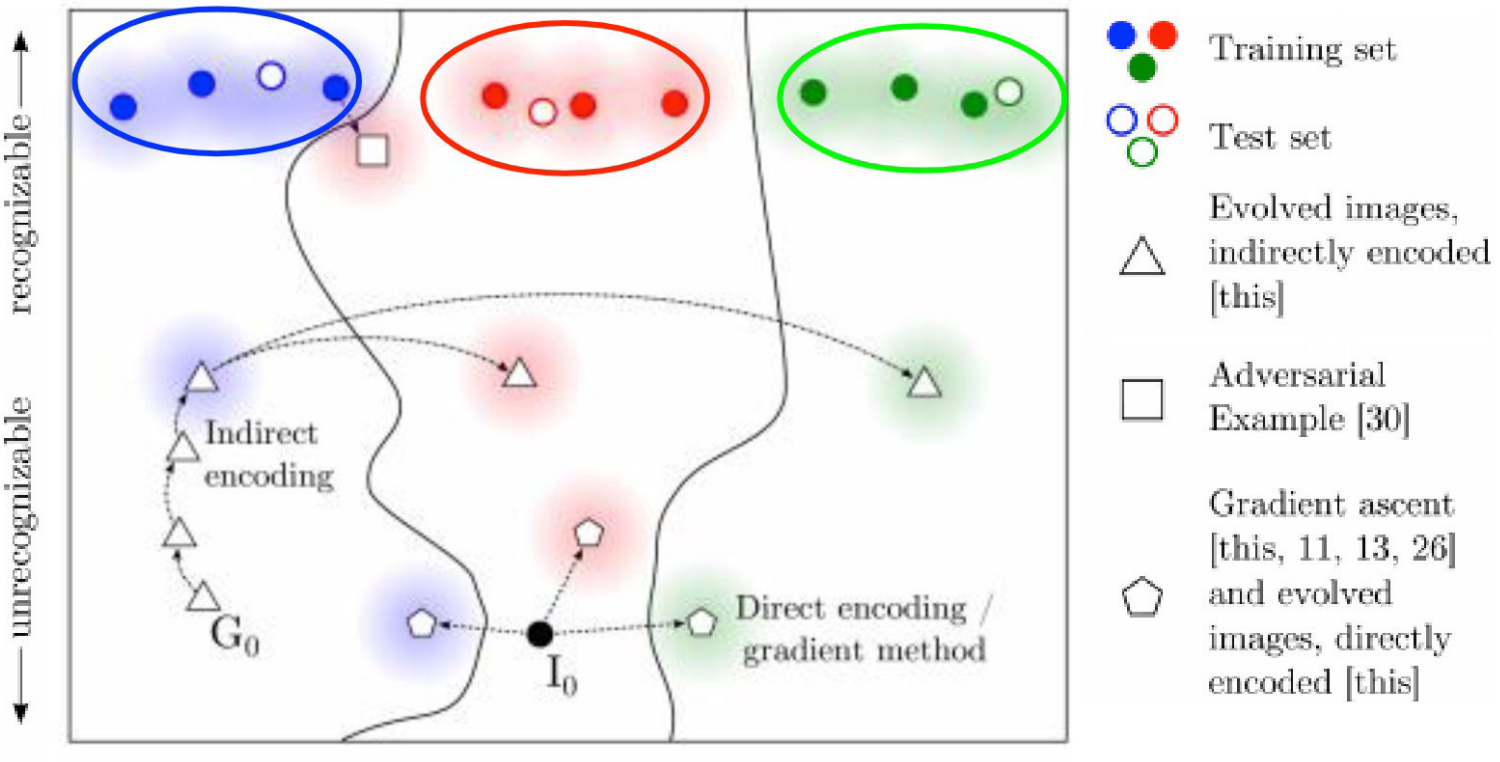


事前防御：异常输入检测

事中防御：模型鲁棒增强（蒸馏梯度、对抗训练）

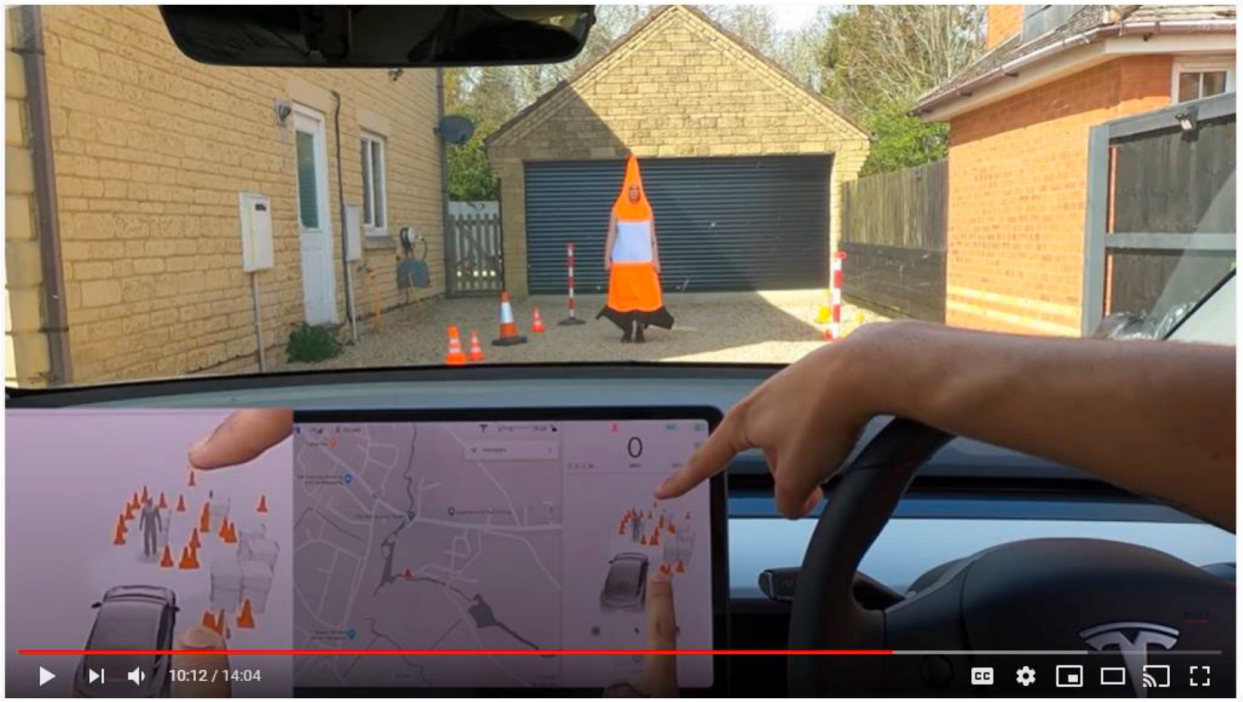
事后防御：审计、集成、级联、拟态防御

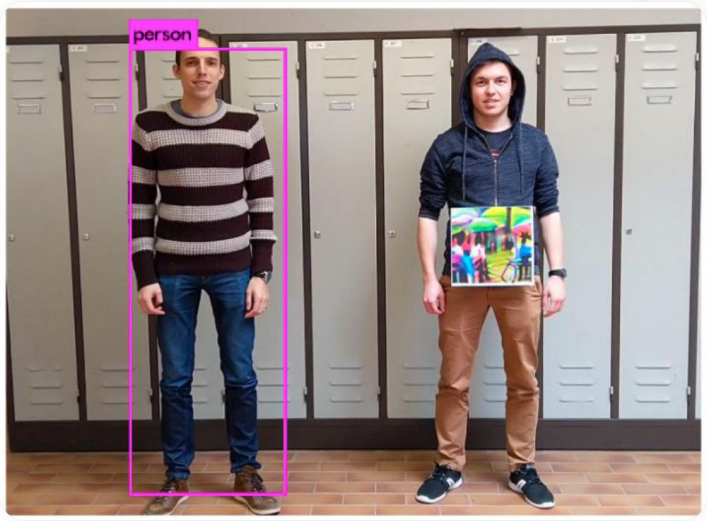
对抗样本存在的原因

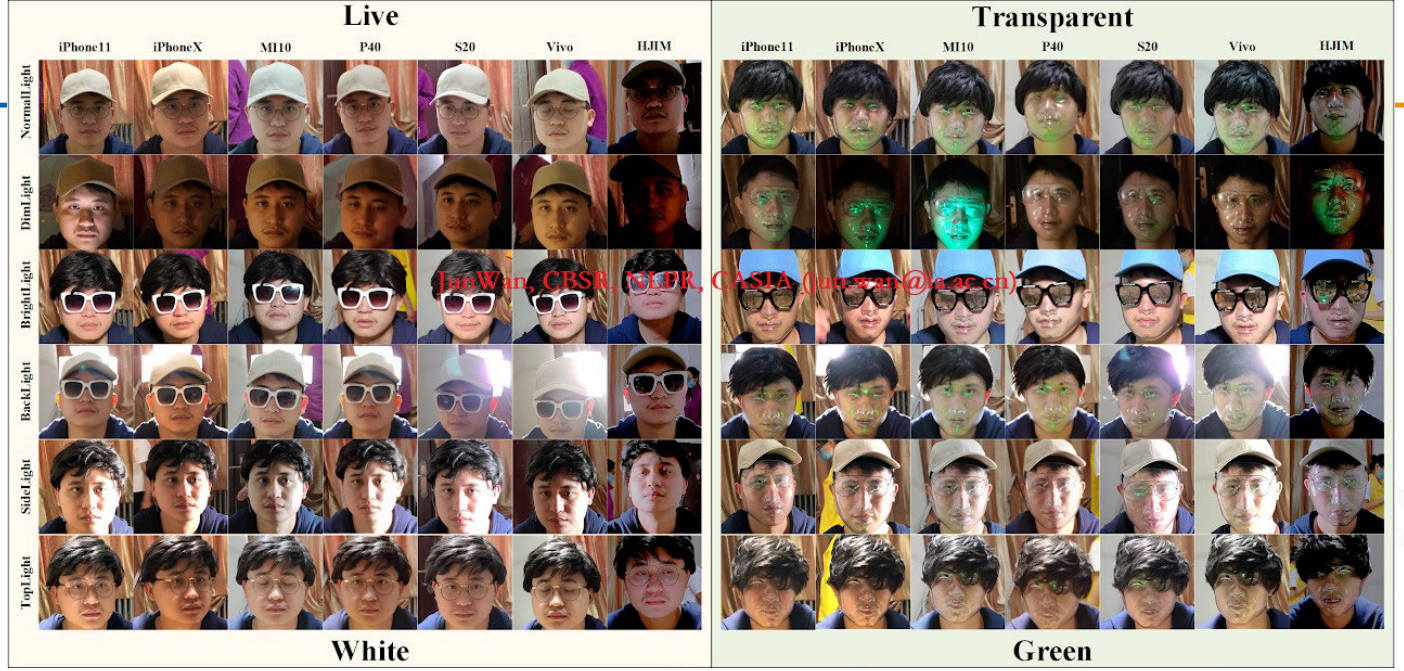


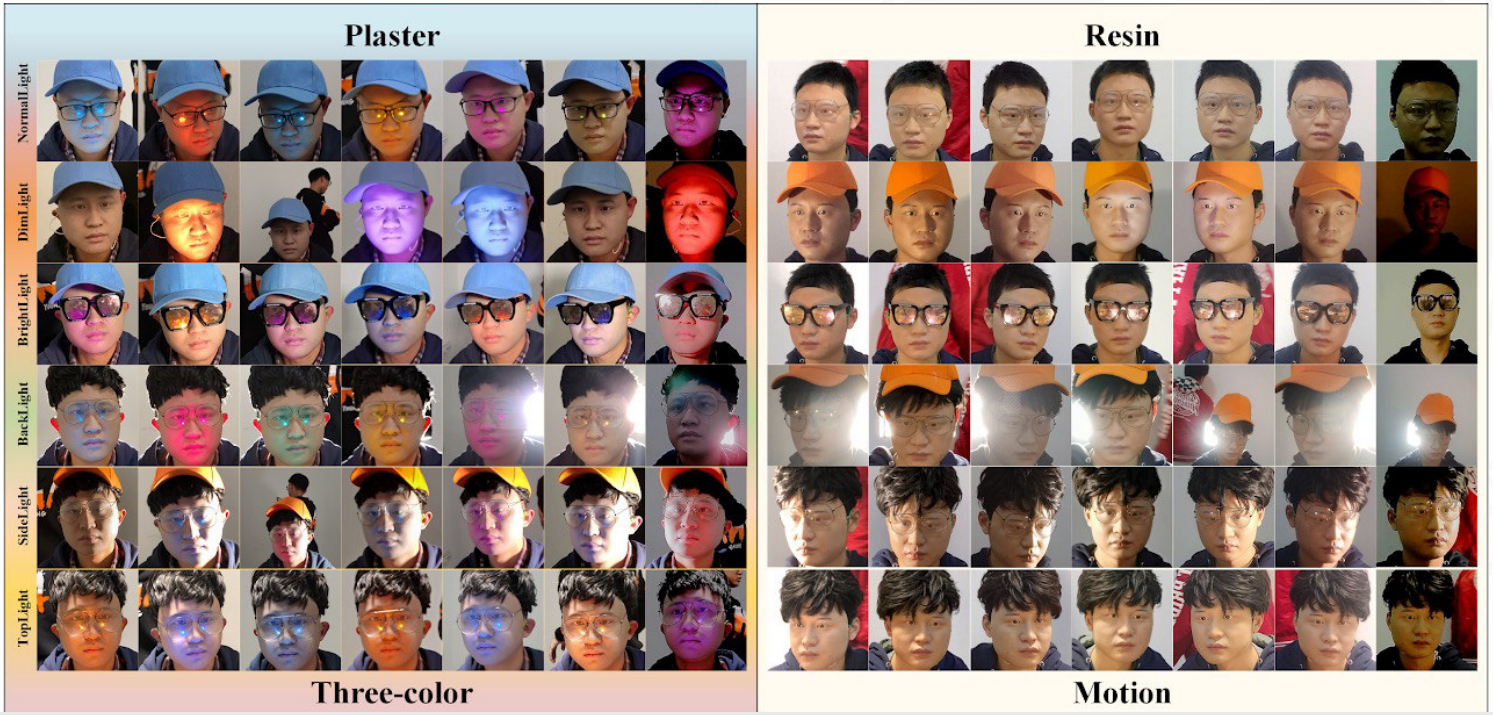
In discriminative models, decision boundary is loose. Data points occupy much less space than what is assigned to them. Generative models would not be easily fooled.

计算机视觉物理域对抗样本攻击案例









神经网络基础最佳实践

1. 动手学深度学习 李沐 https://courses.d2l.ai/zh-v2/
2. AI-EDU Microsoft https://microsoft.github.io/ai-edu/index.html
3. 深度学习面试宝典 Amusi https://github.com/amusi/Deep-Learning-Interview-Book
4. 零基础实践深度学习 百度 https://github.com/PaddlePaddle/awesome-DeepLearning
5. AI千集 https://github.com/weslynn/AlphaTree-graphic-deep-neural-network
6. MindSpore教程 华为 https://www.mindspore.cn/tutorials/zh-CN/r1.6/index.html

人工智能基础：

- 三大学派：行为主义、符号主义、联结主义
- 深度学习（联结主义）：神经网络为代表学习找到一个函数

深度学习安全：

- 完整性、保密性、伦理
  - ▶后门、对抗样本、隐私泄漏、模型偏见、框架系统供应链攻击

对抗样本攻防：

- 电子域攻防
- 计算机视觉物理域对抗样本攻击

感谢批评指正

THANKS

lixeon.lij@gmail.com

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS